



Competition on Spatial Statistics for Large Datasets

Huang HUANG, Sameh ABDULAH, Ying SUN, Hatem LTAIEF, David E. KEYES, and Marc G. GENTON

As spatial datasets are becoming increasingly large and unwieldy, exact inference on spatial models becomes computationally prohibitive. Various approximation methods have been proposed to reduce the computational burden. Although comprehensive reviews on these approximation methods exist, comparisons of their performances are limited to small and medium sizes of datasets for a few selected methods. To achieve a comprehensive comparison comprising as many methods as possible, we organized the *Competition on Spatial Statistics for Large Datasets*. This competition had the following novel features: (1) we generated synthetic datasets with the *ExaGeoStat* software so that the number of generated realizations ranged from 100 thousand to 1 million; (2) we systematically designed the data-generating models to represent spatial processes with a wide range of statistical properties for both Gaussian and non-Gaussian cases; (3) the competition tasks included both estimation and prediction, and the results were assessed by multiple criteria; and (4) we have made all the datasets and competition results publicly available to serve as a benchmark for other approximation methods. In this paper, we disclose all the competition details and results along with some analysis of the competition outcomes.

Key Words: Gaussian processes; Matérn covariance function; Parameter estimation; Prediction; Tukey g -and- h random fields.

1. INTRODUCTION

With the development of better observing techniques and advanced computing devices, it has become easier and more common to obtain large spatial datasets. Therefore, statistical inference in spatial statistics has become computationally challenging. For decades, various approximation methods have been proposed to model and analyze large-scale spatial data when the exact computation is infeasible. However, in the literature, the performance of the statistical inference using those proposed approximation methods has generally been assessed with small and medium datasets only, for which the exact solution can be obtained.

H. Huang · S. Abdullah · Y. Sun · H. Ltaief · D. E. Keyes · M. G. Genton (✉) King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. (E-mail: marc.genton@kaust.edu.sa).

© 2021 International Biometric Society
Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 4, Pages 580–595
<https://doi.org/10.1007/s13253-021-00457-z>

However, for large real-world datasets, the exact computation is no longer feasible. The inference with approximation methods is typically validated empirically or via prediction accuracy with the fitted model.

Motivated by the challenge to compare the statistical and computational efficiency of different approximation methods, several pioneering works were triggered. Englund (1990) performed a very early research investigating the inference performance from different spatial models. The study used the Walker Lake dataset (Srivastava 1987) in two areas with tens of thousands of data points, and they observed considerable variability in the interpolation results from the different spatial models. Bradley et al. (2016) reviewed various spatial predictors, including both deterministic and stochastic approaches, and applied them to satellite measurements of CO₂. Datasets of three different sizes were studied (the largest one comprised tens of thousands of observations), and the assessment of the different methods relied on the prediction error. Heaton et al. (2019) mainly focused on the Gaussian Process (GP) and proposed a competition in which research groups used their selected approximation methods of GP or other model-free algorithmic approaches to make predictions for both a simulated and real-world dataset. Both datasets consisted of 150,000 realizations. The covariance of the simulated data from GP was known and disclosed to the competition participants; the real-world data consisted of land surface temperatures measured by satellites for which the true underlying covariance was unknown. The performance was examined based on the prediction error or the predictive distribution. Wikle et al. (2017) discussed the design of a common task framework to compare different methods. In addition, they developed a website so that researchers could upload their prediction scripts to the website server for the NASA OCO-2 data, and then the associated prediction performance would be published on the website leaderboard.

Inspired by these works, we organized a competition, the *Competition on Spatial Statistics for Large Datasets* (<https://cemse.kaust.edu.sa/stsds/2021-kaust-competition-spatial-statistics-large-datasets>), to involve more recent methodologies and overcome weaknesses existing in previous studies. Compared to previous competitions or comparison works, our competition has the following key features:

- We were able to generate synthetic spatial datasets where we know the true process with size on a much larger scale, ranging from 100 thousand to 1 million realizations, thanks to the *ExaGeoStat* (Abdulah et al. 2018a) software (<https://github.com/ecrc/exageostat>). With the simulated large datasets, we could better understand the statistical efficiency of different methods.
- The datasets were simulated from various spatial models, including both Gaussian and non-Gaussian process models. For GP models, the datasets were simulated with Matérn covariance functions for a selected set of parameters, representing a wide range of statistical properties of the spatial random field. The non-Gaussian spatial datasets were generated by Tukey *g*-and-*h* random fields (Xu and Genton 2017), which generalize GP to account for skewness and heavy tails.
- The competition tasks included both estimation and prediction. It is insightful to examine the extent of departure of a model inferred by GP approximation methods from the

truth. One sub-competition focused on assessing the model misspecification, where we used as criteria the Mean Loss of Efficiency (MLOE) and Mean Misspecification of the Mean Square Error (MMOM), both proposed by [Hong et al. \(2021\)](#). For sub-competitions focusing on predictions, the Root Mean Square Error (RMSE) was used to evaluate the prediction accuracy.

- We shared all the prepared datasets and competition results in a public repository. Future approximation methods for large spatial datasets can use these datasets as benchmark data and compare their performance with existing methods. In addition, we also posted the model parameter estimate results and predictions with exact computation by *ExaGeoStat*, which can be referenced as exact inference results.

The competition was launched on November 23, 2020, and attracted 29 research teams worldwide to register. These registered teams included active researchers in the spatial statistics community. The competition ended on February 1, 2021, and 21 teams successfully submitted their results; see Table S1 in the Supplementary Material for details about the teams that submitted their results.

The rest of the paper is organized as follows: In Sect. 2, we provide an overview of the competition. In Sect. 3, we briefly discuss the methods used in the competition by the participating teams. In Sect. 4, we show details of the competition results with some analysis. In Sect. 5, we conclude and give final remarks.

2. OVERVIEW OF THE COMPETITION

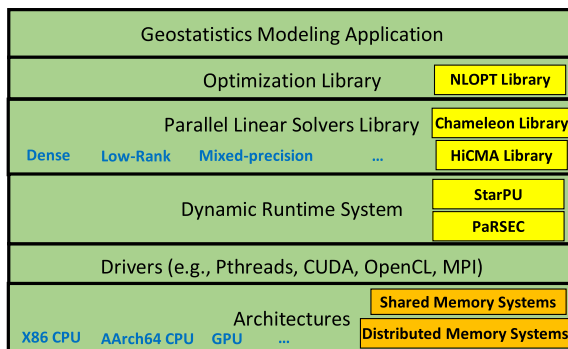
In this section, we give a brief overview of the *ExaGeoStat* software framework and how it was used to generate the datasets in this competition as well as a detailed description of the four sub-competitions, which focused on either the model inference of a zero-mean GP or the spatial prediction.

2.1. *ExaGeoStat* SOFTWARE IN THE COMPETITION PREPARATION

The heart of this competition was *ExaGeoStat*, a C-based high-performance software for geospatial statistics in climate and environment modeling ([Abdulah et al. 2018a](#)). *ExaGeoStat* provides a High-Performance Computing (HPC)-tailored framework that is able to maximize the utilization of cutting-edge parallel hardware architectures with the aid of state-of-the-art high-performance dense linear algebra libraries. Thus, this software is able to tackle the scaling limitations of the Maximum Likelihood Estimation (MLE) and prediction operations, i.e., $O(n^2)$ memory complexity and $O(n^3)$ computation complexity, where n represents the number of spatial locations.

Figure 1 illustrates the set of software libraries that *ExaGeoStat* relies upon to enable its HPC capabilities. The MLE optimization is performed using the *NLOPT* optimization library ([Johnson 2014](#)), which aims to maximize the likelihood function by using different sets of statistical model parameters based on the given covariance function. Furthermore, to perform the underlying linear matrix operations, *ExaGeoStat* relies on state-of-the-art

Figure 1. *ExaGeoStat* software layers for geostatistics applications: *NLOPT* optimization library, parallel linear solvers libraries, dynamic runtime systems, programming drivers on shared and distributed memory architectures.



high-performance linear algebra libraries, specifically, *Chameleon* (CHAMELEON 2021) and *HiCMA* (HICMA 2021). These are tile-based high-performance linear algebra libraries that rely on task-parallel programming models instead of the less efficient block-based algorithms. The *HiCMA* library is the hierarchical approximation version of the *Chameleon* dense library.

Generally, the tile-based algorithms split a given matrix into a set of tiles to perform the required matrix operations and maximize utilization of the underlying hardware resources. The numerical algorithm is translated into a Directed Acyclic Graph (DAG), where the nodes represent tasks and the edges represent data dependencies (e.g., read, write, and read-write) where runtime systems, e.g., *StarPU* and *PaRSEC*, can be used to schedule the DAG tasks across different hardware resources, ensuring that the data dependencies rules predefined by the user are not violated. More detail about task-based MLE operations and *ExaGeoStat* can be found in Abdulah et al. (2018a), Abdulah et al. (2018b) and Abdulah et al. (2019).

The *ExaGeoStat* software has three main components: a synthetic data generator, a modeling tool, and a predictor. The synthetic data generator provides a reference set of synthetic measurements and locations, which generates test cases of prescribed size for standardized comparisons with other methods. This tool facilitates the quality assessment of any proposed approximation method across a wide range of datasets with different features. The modeling tool can compute the MLE function through computation methods that vary from exact to approximate (see Fig. 1). The *ExaGeoStat* predictor aims to predict a set of unknown measurements at new spatial locations. In Abdulah et al. (2019), the capabilities of *ExaGeoStat* were exported to the R environment (R Core Team 2019) through the *ExaGeoStatR* package. With *ExaGeoStatR*, large-scale Gaussian calculations in R are now possible by mitigating its memory and computing limitations. The *ExaGeoStatR* package provides the same functionality of the *ExaGeoStat* software through a set of R functions that abstract the underlying hardware architecture to a set of input-parameters.

In the following sections, we elaborate on the set of synthetic datasets that were used in this competition and generated using the *ExaGeoStat* data generator. The datasets were generated with different true parameters and dataset sizes. We also used the *ExaGeoStat* modeling and predictor tools to provide exact estimations and predictions of the given Gaussian datasets as benchmarks for the solutions submitted by the participants.

2.2. DATASETS USED IN THE COMPETITION

First, we generated 16 datasets, denoted by Dataset G1 – G16, from different zero-mean stationary isotropic GP $Z(\mathbf{s})$ at one million locations in the unit square $[0, 1] \times [0, 1]$ with *ExaGeoStat*. The Matérn covariance was used:

$$\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\} = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\beta} \right)^\nu K_\nu \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\beta} \right) + \tau^2 \mathbb{1}_{\{i=j\}}, \quad (1)$$

where $\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\}$ is the Matérn covariance between realizations of $Z(\cdot)$ at locations \mathbf{s}_i and \mathbf{s}_j , $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν , $\Gamma(\cdot)$ is the Gamma function, and $\mathbb{1}$ is the indicator function. The four parameters determining the covariance structure are: the partial sill σ^2 , the range $\beta > 0$, the smoothness $\nu > 0$, and the nugget τ^2 .

The covariance setup of Datasets G1–G8 is given in Table 1, which also shows the effective range (the distance beyond which the covariance drops below 5%) for each dataset covariance. The partial sill parameter controls the common variability and can be easily scaled. Therefore, we did not vary it but fixed it as $\sigma^2 = 1.5$. We chose three values, 2.3, 1.5, and 0.6, to consider different smoothness scenarios from smooth to rough. With the selected smoothness parameter, we also varied the range parameter such that the resulting effective range matched 0.1, 0.3, and 0.8 for weak, medium, and strong dependence. We did not include the long-range and smoothest case because we found that it leads to numerical instability issues due to the covariance matrix near-singularity problem. Datasets G9–G16 have the same covariance structures as Dataset G1–G8, respectively, except that all the nugget parameters are 0.27, indicating a 18% noise-to-signal ratio (see Table 1). Thus, Datasets G1–G16 cover a broad variety of covariance properties.

To account for non-Gaussian datasets, we used the Tukey g -and- h random fields (Xu and Genton 2017), which generalize GP to account for skewness and heavy tails. More precisely, for the generated GP $Z(\mathbf{s})$, the Tukey g -and- h random process $T(\mathbf{s})$ was defined by marginal transformation at each location \mathbf{s} as follows:

$$T(\mathbf{s}) = \xi + \omega \times \frac{\exp(gZ(\mathbf{s})) - 1}{g} \times \exp\left(\frac{hZ^2(\mathbf{s})}{2}\right),$$

Table 1. Covariance setup for Datasets G1–G16

Dataset	σ^2	β	ν	τ^2	Effective range
G1/G9	1.5	0.017526	2.3	0 (for G1), 0.27 (for G9)	0.1
G2/G10	1.5	0.021080	1.5	0 (for G2), 0.27 (for G10)	0.1
G3/G11	1.5	0.030933	0.6	0 (for G3), 0.27 (for G11)	0.1
G4/G12	1.5	0.052579	2.3	0 (for G4), 0.27 (for G12)	0.3
G5/G13	1.5	0.063240	1.5	0 (for G5), 0.27 (for G13)	0.3
G6/G14	1.5	0.092798	0.6	0 (for G6), 0.27 (for G14)	0.3
G7/G15	1.5	0.168639	1.5	0 (for G7), 0.27 (for G15)	0.8
G8/G16	1.5	0.247462	0.6	0 (for G8), 0.27 (for G16)	0.8

Table 2. Parameters used to generate datasets NG1 and NG2

Dataset	ξ	ω	g	h	σ^2	β	ν	τ^2
NG1	1	2	0.2	0.2	1	0.1	1	0
NG2	1	2	0.5	0.3	1	0.1	1	0

where ξ and ω are the location and scale parameters, respectively, g controls the skewness, and $h \geq 0$ determines the tail-heaviness. We chose two sets of values g and h to consider random processes with medium and strong departure from GP. The parameters used to generate the two non-Gaussian datasets, denoted by Datasets NG1 and NG2, are summarized in Table 2.

2.3. DETAILS OF SUB-COMPETITIONS

The first sub-competition (Sub-competition 1a) was about examining the ability of different methods to infer the correct GP model on a moderately large dataset. We chose 90,000 realizations in each of Datasets G1–G16 and asked the participating teams to estimate the four parameters σ^2 , $\beta > 0$, $\nu > 0$, and τ^2 . The metrics used to evaluate the performances are MLOE and MMOM (Hong et al. 2021) across different datasets. MLOE characterizes the average loss of prediction efficiency when the approximated model is used for predictions instead of the true model. MMOM characterizes the average misspecification of the mean square error when calculated under the approximated model. Details of MLOE and MMOM are given in Section S3 in the Supplementary Material.

The second sub-competition (Sub-competition 1b) was about assessing spatial prediction performance on a moderately large dataset generated from a GP model. For each of Datasets G1–G16, we gave 10,000 new locations to participating teams and asked them to predict over these locations conditional on the 90,000 realizations provided in Sub-competition 1a. RMSE was used to evaluate the prediction accuracy.

The third sub-competition (Sub-competition 2a) focused on prediction for non-Gaussian data, where we asked participating teams to predict over 10,000 new locations conditional on 90,000 realizations for each of Datasets NG1 and NG2.

The fourth sub-competition (Sub-competition 2b) was about modeling much larger datasets. One Gaussian dataset (Dataset G5) and one non-Gaussian dataset (Dataset NG1) were chosen. We increased the conditional data size to a very large number, 900,000, and the participating teams needed to predict over 100,000 new locations.

A summary of the four sub-competitions is given in Table 3. The teams could choose to participate in one or more sub-competitions, and we used separate rankings for each of the four sub-competitions because each sub-competition had its own emphasis. We also allowed and encouraged participating teams to have more than one submission if they used different methods to solve the given problems. The participating teams were not informed whether the data were Gaussian or not in Sub-competitions 2a and 2b.

Table 3. Summary of the four sub-competitions

Sub-competition	Task	True data model	Metric	Data size
1a	GP estimation	GP	MLOE & MMOM	90,000
1b	prediction	GP	RMSE	predict 10,000 conditional on 90,000
2a	prediction	non-GP (Tukey <i>g</i> -and- <i>h</i>)	RMSE	predict 10,000 conditional on 90,000
2b	prediction	GP & non-GP (Tukey <i>g</i> -and- <i>h</i>)	RMSE	predict 100,000 conditional on 900,000

Ideally, we would have replicates of the datasets for each setting to better assess the methods. However, the problems proposed here in this competition were of very big size; therefore, replications of the model inference and prediction would require too many computational resources for the participating teams. For this practical reason, we used only one dataset for each setting.

It is also noteworthy that the final ranks were computed based on the determined rule in the competition, which was that we applied equal weights to the *rank* rather than the *assessing metric* in each dataset (discussed in detail in Sect. 2.4). This means that if one team had an extremely poor estimation or prediction for one dataset, a good performance in another dataset would still be able to compensate. However, had the submissions been assessed using a different rule, such as the mean of all metrics across different datasets, then the final ranking of submissions may be different.

2.4. ASSESSMENT

We assessed and assigned the rank for each team in each sub-competition as follows.

In Sub-competition 1a, we let $K^{(1a)}$ denote the total number of different submissions for Sub-competition 1a, and $P_{ki1}, P_{ki2}, k = 1, \dots, K^{(1a)}, i = 1, \dots, 16$, denote the absolute MLOE and absolute MMOM from submission k for dataset i , respectively. Then, for each dataset i and metric $j = 1, 2$, we sorted $P_{kij}, k = 1, \dots, K^{(1a)}$ in ascending order and assigned rank $R_{kij}^{(1a)}$ to each submission (the averaged rank was used for ties). The final score for submission k in Sub-competition 1a was calculated as $S_k^{(1a)} = \sum_{i=1}^{16} (R_{ki1}^{(1a)} + R_{ki2}^{(1a)})$, and the final rank was assigned by sorting $S_k^{(1a)}$ in ascending order (the averaged rank was used for ties).

For Sub-competitions 1b, 2a, 2b, we let $K^{(\text{cmp})}$ denote the total number of different submissions for Sub-competition *cmp* (i.e., *cmp* = 1b, 2a, or 2b) and let the RMSE from submission k for dataset i be denoted by $\text{RMSE}_{ki}^{(\text{cmp})}, k = 1, \dots, K^{(\text{cmp})}, i = 1, \dots, 16$ when *cmp* = 1b and $i = 1, 2$ when *cmp* = 2a or 2b. For each dataset i , we sorted $\text{RMSE}_{ki}^{(\text{cmp})}, k =$

$1, \dots, K^{(\text{cmp})}$ in ascending order and assigned rank $R_{ki}^{(\text{cmp})}$ to each submission (the averaged rank was used for ties). The final score for submission k in Sub-competition cmp was calculated as $S_k^{(\text{cmp})} = \sum_{i=1}^{16} R_{ki}^{(\text{cmp})}$ when $\text{cmp} = 1\text{b}$ and $S_k^{(\text{cmp})} = \sum_{i=1}^2 R_{ki}^{(\text{cmp})}$ when $\text{cmp} = 2\text{a}$ or 2b , and the final rank was assigned by sorting $S_k^{(\text{cmp})}$ in ascending order (the averaged rank was used for ties).

2.5. RESULTS

The full competition results for each submission are given in Table S2 in the Supplementary Material. Many approximation methods were used, and we provide a brief summary of them in Sect. 3. To highlight the top performers, the top three submissions in each sub-competition (four submissions in Sub-competition 2a due to a tie among three teams for the second place) are as follows, with the best scoring team listed first:

- Sub-competition 1a: [1] SpatStat-Fans, [2] GpGp, [3] RESSTE(CL/krig)
- Sub-competition 1b: [1] RESSTE(CL/krig), [2] HCHISS, [3] Chile-Team
- Sub-competition 2a: [1] RESSTE(Tukey-g-h-trans-GPGP), [3] GpGp(quick), [3] HMatrix, [3] RESSTE (nonpara-trans-GPGP)
- Sub-competition 2b: [2] RESSTE(nonpara-trans-GPGP), [2] RESSTE(Tukey-g-h-trans-GPGP), [2] Tohoku-University

Besides the competition submissions, we also used *ExaGeoStat* to see the rank of the exact computations in Sub-competitions 1a and 1b. The augmented top lists for Sub-competitions 1a and 1b are shown in Table 4, where “ExaGeoStat(estimated-model)” in Sub-competition 1a means that we used *ExaGeoStat* to estimate the Matérn covariance parameters by maximizing the full likelihood with exact computation; in Sub-competition 1b, it means that we used the associated estimated model to make predictions with exact computation. “ExaGeoStat(true-model)” in Sub-competition 1b means that the prediction was made using the true model with exact computation by *ExaGeoStat*.

Table 4. Results for Sub-competitions 1a and 1b with submissions from *ExaGeoStat*

Sub-competition	Submission	Score	Rank
1a	ExaGeoStat(estimated-model)	154	1
1a	SpatStat-Fans	156	2
1a	GpGp	186	3
1a	RESSTE(CL/krig)	229	4
1b	ExaGeoStat(true-model)	72	1
1b	RESSTE(CL/krig)	78	2
1b	ExaGeoStat(estimated-model)	79	3
1b	HCHISS	93	4
1b	Chile-Team	113	5

We observe that ExaGeoStat(estimated-model) managed to find the closest model to the truth in Sub-competition 1a, but the prediction performance is slightly worse than RESSTE(CL/krig) in Sub-competition 1b. This suggests that using a model closer to the truth does not guarantee a more accurate point-prediction performance for a given realization. However, ExaGeoStat(estimated-model) should yield the best predictions on average among other approximation methods over multiple realizations of a spatial process. When the true parameter values were used for prediction (ExaGeoStat(true-model)), the score reduced from 79 to 72, the best achieved in Sub-competition 1b.

3. OVERVIEW OF METHODS USED IN THE COMPETITION

In this section, we do not intend to provide an exhaustive literature review of existing approximation methods. Instead, we briefly discuss the methods used by participants in the competition. A detailed literature review can be found in [Sun et al. \(2012\)](#) and [Heaton et al. \(2019\)](#).

- **Composite likelihood methods** approximate the joint likelihood as a weighted product of a collection of component likelihoods ([Varin et al. 2011](#)). For example, Vecchia's approximation framework uses a series of conditional likelihoods where the conditioning sets are chosen sparsely ([Vecchia 1988](#)). Pairwise likelihood methods take the likelihoods of each pair of observations as the component likelihoods ([Varin 2008](#)). Therefore, each component in the composite likelihood can be obtained with fewer computations. Teams Among-Stats, Chile-Team, ExtStat, GpGp, HCHISS, RESSTE, etc., submitted results with composite likelihood approximation methods.
- **Low-rank approximation methods** generally project the entire random process to a certain low-dimensional space and use the low-rank representation as a surrogate to approximate the original process. For example, predictive processes ([Banerjee et al. 2008](#)) place knots in the spatial domain, and the expectation of the original process conditional on the realizations on the knots is used as the substitute. Fixed rank kriging ([Cressie and Johannesson 2008](#)) uses a small number of basis functions to represent the process so that the precision matrix can be obtained by inversion of a matrix with a much smaller dimension. Teams utilizing low-rank approximation methods in the competition are ExtStat, UOW, etc.
- Another direction is **approximating the covariance or the precision matrix with sparse structure** so that the computation becomes feasible. Covariance tapering ([Furrer et al. 2006](#); [Kaufman et al. 2008](#)) multiplies a correlation function with compact support to the original covariance function so that the correlation of distant locations is shrunk to zero, and sparsity is induced in the covariance matrix. For the precision matrix, the Gaussian Markov random fields naturally yield a sparse structure in the precision matrix ([Rue et al. 2009](#)). Team ExtStat submitted results using this technique.
- **Combinations or extensions of different approaches** are also possible. Hierarchical matrix methods ([Litvinenko et al. 2019](#)) apply the hierarchical matrix approximation

format to the covariance matrix. Then, the off-diagonal blocks of the covariance matrix are represented with low ranks so that the covariance matrix can be inverted with a lower computational cost. Team HMatrix submitted results by this approach. A full-scale approximation of covariance functions (Sang and Huang 2012) combines covariance tapering and predictive process models to account for small- and large-scale spatial dependence at the same time. Team SpatStat-Fans used this method. Multiresolution approximation (Katzfuss 2017) is an extension of predictive processes or the full-scale approximation, where basis functions with a hierarchical structure are used to capture spatial dependence at different scales. Teams Colorado-School-of-Mines and GPvecchia had submissions with this method. Nearest-Neighbor Gaussian Processes (Datta et al. 2016) extend the Vecchia approximation to a process-based model so that the parameters are estimated and predictions are made with a unifying framework. Teams ExtStat and NNGP applied this approach in their submissions.

Here, we provide brief descriptions and settings for the top teams in the competition.

For the GP model inference problem in Sub-competition 1a, SpatStat-Fans applied the smoothed full-scale approximation method where the entire domain is partitioned into 10×10 regular rectangular blocks, and the knot set is on a 20×20 grid. GpGp subsampled 30,000 observations and then used Vecchia's approximation conditional on 30 nearest neighbors by the R package *GpGp* (Guinness et al. 2021). RESSTE(CL/krig) used composite likelihoods to find the optimal covariance parameter estimates.

For the GP prediction problem in Sub-competition 1b, RESSTE(CL/krig) used plug-in kriging predictors with the inferred parameters by composite likelihoods. HCHISS used kriging conditional on 1,000 nearest neighbors with covariance parameters estimated by Vecchia's approximation. Chile-Team used kriging conditional on 800 nearest neighbors with covariance parameters estimated by Gaussian conditional pairwise likelihood.

For the non-Gaussian or very large prediction problem, RESSTE(Tukey-g-h-trans-GPGP) and RESSTE(nonpara-trans-GPGP) in Sub-competitions 2a and 2b applied the Tukey *g*-and-*h* transformation and a nonparametric transformation so that the transformed data are approximately Gaussian, respectively, and then used the R package *GpGp* for Gaussian predictions. HMatrix in Sub-competition 2a used hierarchical matrix approximation for the covariance matrix with accuracy 10^{-6} . GpGp(quick) in Sub-competition 2a used the "matern_nonstat_var" covariance function in the R package *GpGp*, where 50 basis functions were used to represent the spatially varying covariance function and the covariance parameters were estimated by 10,000 random samples with 20 conditional neighbors; then, the prediction was carried out by kriging with 30 conditional neighbors. Tohoku-University in Sub-competition 2b used covariance tapering in which the Matérn covariance function of the GP was applied with parameters estimated by cross-validation.

More details of the methods used by the top teams in each sub-competition will be provided by the discussants of the paper.

4. COMPETITION RESULT ANALYSIS

In this section, we provide more details about the competition results. Figure 2 illustrates the parameter estimates submitted by all the teams in Sub-competition 1a as well as *ExaGeoStat* with exact computation for comparison. We highlight the results of *ExaGeoStat* and the top three performers in Sub-competitions 1a (SpatStat-Fans, GpGp, RESSTE(CL/krig)) and 1b (RESSTE(CL/krig), HCHISS, Chile-Team). Note that the submission RESSTE(CL/krig) was among the top three in both sub-competitions. All submissions except HCHISS succeeded in estimating the nugget parameters very precisely. We observe that the parameter estimation was generally more difficult when the process was smoother (larger smoothness parameter) and had stronger dependence (larger effective range). In such cases, the partial sill and range parameter estimates from different submissions differed the most. For comparison purposes, we also show the model inference results of submissions HCHISS and Chile-Team, which ranked the second and third in Sub-competition 1b, respectively. However, we notice that their model estimates were not as good as SpatStat-Fans, GpGp, and RESSTE(CL/krig). Even though *ExaGeoStat* had the most accurate estimate overall, we note that for datasets G15 and G16, *ExaGeoStat* with exact computation tended to overestimate the partial sill and range. SpatStat-Fans and RESSTE(CL/krig) showed patterns similar to the exact computation results, but the estimates obtained by GpGp were more accurate and closer to the truth. Figure S1 in the Supplementary Material illustrates the absolute MLOE and MMOM, where we observe that GpGp indeed had smaller absolute MLOE and MMOM for datasets G15 and G16. The likelihood values at the estimated parameters can also be used for comparison. We used *ExaGeoStat* to calculate the exact loglikelihood when the parameter estimates from the submissions were plugged in. Figure S2 in the Supplementary Material depicts the loglikelihood from the submissions minus the loglikelihood with the true parameters. For those methods that had a smaller loglikelihood, such as Chile-Team and HCHISS, it means that they failed to find the maximizers of the likelihood due to approximation. Those with higher values, such as *ExaGeoStat* and SpatStat-Fans, may have obtained the optimal estimates for the given dataset.

Figure 3 shows the RMSE from different submissions for each dataset in Sub-competition 1b. We highlight the same submissions as we discussed before for Sub-competition 1a, including the top three submissions in Sub-competitions 1b and 1a. In addition, we used *ExaGeoStat* to make predictions with exact computation using the true parameters and the estimates by *ExaGeoStat* in Sub-competition 1a, and the corresponding RMSEs are also given and highlighted. In the top panel of Fig. 3, we use boxplots to summarize the overall prediction performance for different datasets. Because the RMSE from the top teams in Sub-competitions 1a and 1b cannot be differentiated well using the boxplot scale, we also show their RMSE with bar charts in the bottom panels of Fig. 3 for better comparisons among these top teams. We observe that the RMSE was generally larger when the nugget existed because the data had a higher level of noise. It is noteworthy that the top performers in Sub-competition 1a, SpatStat-Fans and GpGp, succeeded in finding excellent parameter estimates. However, their better-inferred models did not lead to better overall predictions compared to the other highlighted submissions. One possible reason is that their approximation was inadequate in kriging, even though the underlying model they used was

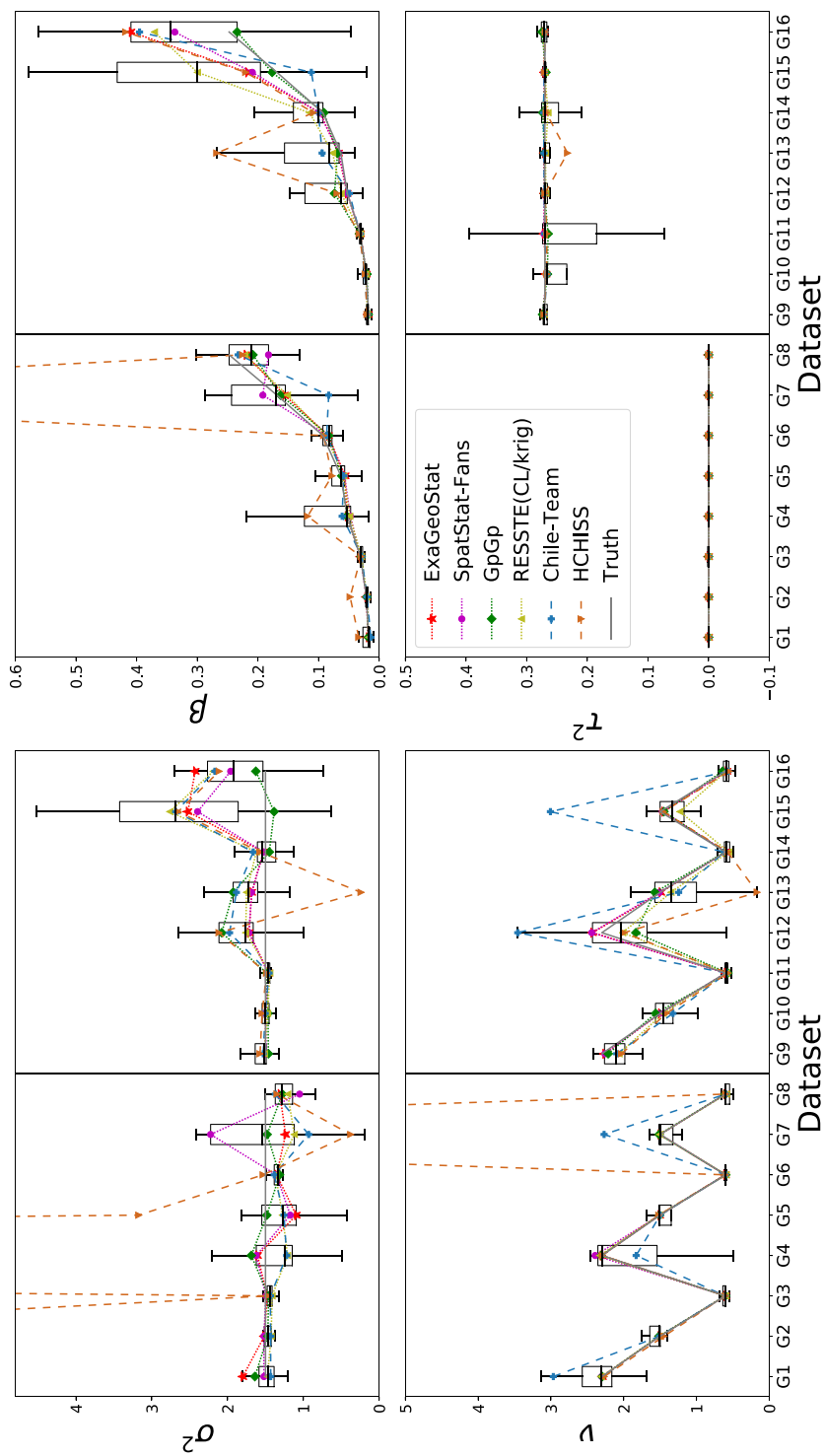


Figure 2. Boxplots of parameter estimates from all teams in Sub-competition 1a (outliers are not shown for clarity). The true values and estimates by *ExaGeoStat*, *SpatStat-Fans*, *GpGp*, *RESSTE(CL/krig)*, *Chile-Team*, and *HCHISS* are highlighted. The legend order of the highlighted submissions (except the truth) follows their ranks in Sub-competition 1a. Datasets G9–G16 share the same covariance structure as G1–G8, respectively, except with a nugget.

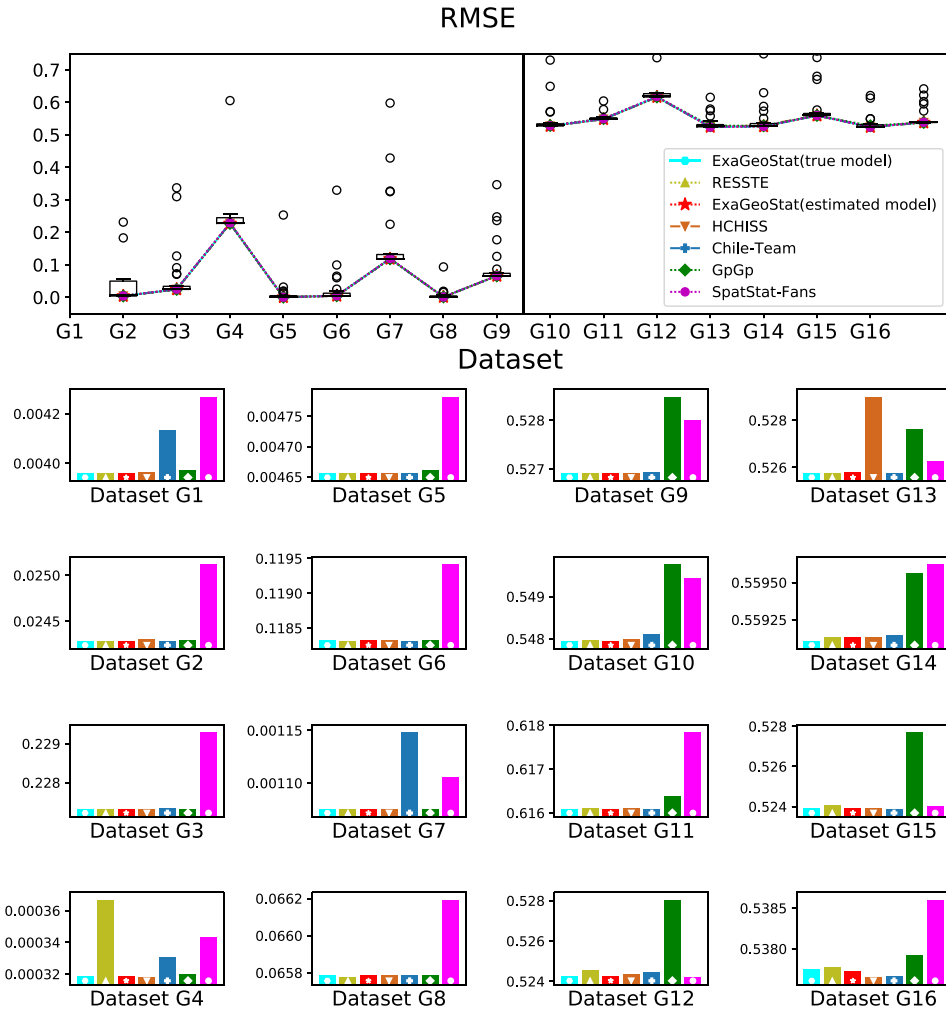


Figure 3. Boxplots of RMSE from all submissions in each dataset in Sub-competition 1b. *ExaGeoStat* predictions with the true parameters and estimated parameters by *ExaGeoStat* in Sub-competition 1a are also given. In the legend, the highlighted submissions are listed in order of their rank in Sub-competition 1b. RMSE from the top teams in Sub-competitions 1a and 1b is highlighted and shown in the bar charts. Datasets G9–G16 share the same covariance structure as G1–G8, respectively, except with a nugget.

more accurate. In fact, GpGp only used 50 nearest neighbors as the conditional set for each prediction, whereas HCHISS used 1000 nearest points. This demonstrates that both the model inference and the number of neighbors considered are important for local kriging predictions; it is difficult to tell to what extent the number of neighbors matters.

The RMSE summary in Sub-competitions 2a and 2b is given in Figures S3 and S4 in the Supplementary Material, respectively, where we highlight the top teams in both sub-competitions. The top performers include the application of the Tukey *g*-and-*h* transformation and nonparametric transformations to GPs as well as other local kriging predictions based on inferred (nonstationary) GP models.

5. DISCUSSION

In this competition, we created and released a set of benchmark data with different designs. We knew the true parameters used to generate the datasets as well as the exact maximum likelihood estimates by *ExaGeoStat*, which can be used to investigate future proposed methods. For practical reasons, we only selected and used subsets of the generated GP datasets in this competition. The full datasets with one million spatial locations are publicly available on <https://doi.org/10.25781/KAUST-8VP2V> for ease of use in future research. Future approximation methods can use this repository as a tool to assess their performance against the submissions from the different participating teams and the exact inference using *ExaGeoStat* in this competition (a detailed summary of the exact maximum likelihood estimates by *ExaGeoStat* in Sub-competition 1a is also given in Table S3 in the Supplementary Material).

We did not compare the computational time in this competition because the participating teams modeled the data on their own machines, and the execution time is not directly comparable. However, we summarize the execution time from all submissions for making predictions in Sub-competitions 1b, 2a, and 2b in Figure S5 in the Supplementary Material. The median time for making 10,000 predictions conditional on 90,000 observations was around 60 seconds for Gaussian data (in Sub-competition 1b) and 430 seconds for non-Gaussian data (in Sub-competition 2a). For a larger dataset in Sub-competition 2b, the median time for making 100,000 predictions conditional on 900,000 observations was around 2700 seconds.

We also note that replicates of the datasets with the same setting were ideally needed to better assess different methods from a statistical point of view. However, the datasets used in this competition were already quite large, making it infeasible for many teams to perform inference and prediction with many replicates. To make the competition workable for most participating teams, we only used one replicate in each setup. Nevertheless, the wide variety of covariance setups we considered provided a fair comparison for large spatial data modeling.

For decades, the big spatial data problem has been an active research area due to the challenges caused by the ubiquity of large spatial datasets, which often contain millions of observations, such as remote sensing climate data or numerical model outputs. The “big data” research field has been advanced by the size of the spatial data in real applications. In addition to developing efficient and accurate methods for larger spatial datasets, recent research has been focused on multivariate spatial and spatio-temporal data, where the data size can be magnified significantly. The prediction problem will then include both spatial interpolation and temporal forecasting for single or multiple variables. Providing a unified framework for understanding the performance of existing approximation methods is much more challenging in simulation and assessment but crucial for suggesting future research directions.

Funding Funding was provided by King Abdullah University of Science and Technology.

REFERENCES

- Abdulah S, Li Y, Cao J, Ltaief H, Keyes DE, Genton MG, Sun Y (2019) ExaGeoStatR: A package for large-scale geostatistics in R. arXiv preprint [arXiv:1908.06936](https://arxiv.org/abs/1908.06936)
- Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018a) ExaGeoStat: A high performance unified software for geostatistics on manycore systems. *IEEE Trans Parallel Distrib Syst* 29(12):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), pp. 98–108
- Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2019). Geostatistical modeling and prediction using mixed precision tile Cholesky factorization. In: 2019 IEEE 26th international conference on high performance computing, data, and analytics (HiPC), pp. 152–162
- Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. *J Royal Stat Soc: Ser B (Stat Methodol)* 70(4):825–848
- Bradley JR, Cressie N, Shi T (2016) A comparison of spatial predictors when datasets could be very large. *Stat Surv* 10:100–131
- CHAMELEON (2021, January). The Chameleon project. Available at <https://project.inria.fr/chameleon>
- Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. *J Royal Stat Soc: Ser B (Stat Methodol)* 70(1):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(514):800–812
- Englund EJ (1990) A variance of geostatisticians. *Math Geol* 22(4):417–455
- Furrer R, Genton MG, Nychka D (2006) Covariance tapering for interpolation of large spatial datasets. *J Comput Graph Stat* 15(3):502–523
- Guinness J, Katzfuss M, Fahmy Y (2021) GpGp: Fast Gaussian Process Computation Using Vecchia’s Approximation. R package version 0.3.2
- Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, Gerber F, Gramacy RB, Hammerling D, Katzfuss M, Lindgren F, Nychka DW, Sun F, Zammit-Mangion A (2019) A case study competition among methods for analyzing large spatial data. *J Agricult Biol Environ Stat* 24(3):398–425
- HICMA (2021, January). The HiCMA project. Available at <https://github.com/ecrc/hicma>
- Hong Y, Abdulah S, Genton MG, Sun Y (2021). Efficiency assessment of approximated spatial predictions for large datasets. *Spat Stat* 43:100517
- Johnson SG (2014) The NLOpt nonlinear-optimization package. Available at <https://github.com/stevengj/nlopt>
- Katzfuss M (2017) A multi-resolution approximation for massive spatial datasets. *J Am Stat Assoc* 112(517):201–214
- Kaufman CG, Schervish MJ, Nychka DW (2008) Covariance tapering for likelihood-based estimation in large spatial data sets. *J Am Stat Assoc* 103(484):1545–1555
- Litvinenko A, Sun Y, Genton MG, Keyes DE (2019) Likelihood approximation with hierarchical matrices for large spatial datasets. *Comput Stat Data Anal* 137:115–132
- R Core Team (2019) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Royal Stat Soc: Ser B (Stat Methodol)* 71(2):319–392
- Sang H, Huang JZ (2012) A full scale approximation of covariance functions for large spatial data sets. *J Royal Stat Soc: Ser B (Stat Methodol)* 74(1):111–132
- Srivastava RM (1987) A non-ergodic framework for variograms and covariance functions. Master’s thesis, Stanford University, Stanford, CA
- Sun Y, Li B, Genton MG (2012) Geostatistics for large datasets, Chapter 3. In: Porcu E, Montero J-M, Schlather M (eds) *Advances and challenges in space-time modelling of natural events*, vol 207. Springer, Berlin, pp 55–77

- Varin C (2008) On composite marginal likelihoods. *Adv Stat Anal* 92(1):1–28
- Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. *Statistica Sinica* 21:5–42
- Vecchia AV (1988) Estimation and model identification for continuous spatial processes. *J Roy Stat Soc: Ser B (Methodol)* 50(2):297–312
- Wikle CK, Cressie N, Zammit-Mangion A, Shumack C (2017). A common task framework (ctf) for objective comparison of spatial prediction methodologies. *Stats & data science views*. Available at <https://www.statisticsviews.com/article/a-common-task-framework-ctf-for-objective-comparison-of-spatial-prediction-methodologies>
- Xu G, Genton MG (2017) Tukey g -and- h random fields. *J Am Stat Assoc* 112(519):1236–1249

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.