2023 KAUST Competition on Spatial Statistics for Large Datasets

January 17, 2023

1 Introduction

The rapid increase in the volume of geospatial data over recent years has added more challenges to processing these data using traditional methods. Thus, geospatial applications have brought High-Performance Computing (HPC) into the mainstream and further increased its use in the spatial statistics field. *ExaGeoStat*[1] is one example of an HPC software that enables large-scale parallel generation, modeling, and prediction of large geospatial data via covariance matrices. Unlike other existing tools which typically rely on approximations to deal with the vast data volume on day-use machines, *ExaGeoStat* allows the processing of big geospatial data using modern HPC hardware in the exact mode without approximation. Therefore, *ExaGeoStat* makes it possible to fairly compare various approximation methods on synthetic datasets via large-scale simulations, not only to the true models but also to the exact solutions provided by *ExaGeoStat*. In 2021 and 2022, we hosted two competitions to assess the performance of existing methods/tools in the estimation and prediction of the given spatial and spatio-temporal datasets. Those datasets were synthetic datasets generated from specified statistical models, in particular, covariance models, using *ExaGeoStat*. All the datasets for the two previous competitions are available for download from Huang et al. (2021b) and Abdulah et al. (2022a). The associated descriptions of 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).

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

This year, we are hosting a third competition with different objectives and datasets. Instead of point estimation and prediction, the 2023 competition focuses on the construction of confidence and prediction intervals. It includes four sub-competitions, i.e., 1a, 1b, 2a, and 2b. The datasets were generated from stationary Gaussian random fields with an isotropic Matérn covariance function. There are several datasets with various designs of irregularly spaced locations. We provide two training dataset sizes, 90K and 900K, and two testing dataset sizes, 10K and 100K, to accommodate participants with different levels of computing resources.

2 Datasets

In this competition, spatial data $Z(\boldsymbol{s}_i), i \in \{1, \ldots, n\}$ were generated from a zero-mean stationary Gaussian random fields with an isotropic Matérn covariance function:

$$\mathrm{Cov}\big\{Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\big\} = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta}\Big)^{\nu} K_\nu\Big(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\beta}\Big) + \tau^2 \mathbb{1}_{\{\boldsymbol{s}_i = \boldsymbol{s}_j\}}, \tag{1}$$

where $\mathrm{Cov}\big\{Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\big\}$ is the covariance between $Z(\cdot)$ at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order $\nu > 0$, $\Gamma(\cdot)$ is the Gamma function, $\mathbb{1}_{\{\cdot\}}$ is the indicator function, σ^2 is the variance parameter, τ^2 is the nugget effect, and $\beta > 0$ is the range parameter.

The competition provides datasets generated using the *ExaGeoStat* software. The settings are described as follows:

1. We have generated 5 training datasets at 90K locations with various designs of the spatial locations and 2 corresponding testing datasets at 10K locations for the first four training datasets and 1 testing dataset at 10K locations for the last case. Each dataset was generated from a zero-mean Gaussian process with a set of specified parameters for the Matérn covariance function (1).

2. We have generated 5 training datasets at 900K locations with various designs of the spatial locations and 2 corresponding testing datasets at 100K locations for the first four training datasets and 1 testing dataset at 100K locations for the last case. Each dataset was generated from a zero-mean Gaussian process with a set of specified parameters for the Matérn covariance function (1).

The competition focuses on two main objectives. Sub-competitions 1a and 1b aim at constructing the confidence intervals of the unknown parameters for the Matérn covariance function (1) for the given datasets. Sub-competitions 2a and 2b ask for providing the prediction intervals for the given datasets at a set of new locations in the testing datasets. Table 1 provides a summary of the datasets' settings.

Table 1: 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	90K	–
1b	Gaussian Matérn	Estimation (95% confidence interval)	5	900K	–
2a	Gaussian Matérn	Prediction (95% prediction interval)	5	90K	10K
2b	Gaussian Matérn	Prediction (95% prediction interval)	5	900K	100K

In Sub-competitions 1a and 1b, the participants should provide the 95% independent confidence intervals for each of the σ^2, β, ν, and τ^2 parameters. We do not consider simultaneous confidence intervals for these parameters. In Sub-competitions 2a, we provide 5 training datasets with 90K locations and 9 testing datasets with 10K locations, while in 2b, we provide 5 training datasets with 900K locations and 9 testing datasets with 100K locations. The participants should provide the corresponding independent 95% prediction intervals for each testing point. The confidence and prediction intervals are mandatory for our competition, while the parameters estimation and prediction values for testing points are not required.

3 Assessment and Ranking Strategy

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, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^\top = (\sigma^2, \beta, \nu, \tau^2)^\top$ be the unknown parameters and $[lc_i, uc_i]$, $i = 1, 2, 3, 4$, be the corresponding 95% confidence intervals for θ_i. The confidence interval scoring rule is defined by

$$\text{IS}_{\text{est}} = \frac{1}{4} \sum_{i=1}^{4} S_\alpha^{\text{int}}(lc_i, uc_i; \theta_i).$$

Assuming that $K^{(1a)}$ teams participated in Sub-competition 1a, let $\text{IS}^{(1a)}_{\text{est},k,j}$, for $k = 1, \ldots, K^{(1a)}$, $j = 1, \ldots, 5$, denote the confidence interval score from team k for dataset j. For each dataset j, we sort $\text{IS}^{(1a)}_{\text{est},k,j}$, $k = 1, \ldots, K^{(1a)}$ in ascending order and assign rank $Rank^{(1a)}_{k,j}$ to each team. The final score for team k in Sub-competition 1a is calculated as $S^{(1a)}_k = \sum_{j=1}^{5} Rank^{(1a)}_{k,j}$, and the final rank is assigned by sorting $S^{(1a)}_k$ in ascending order. A similar ranking strategy is used for Sub-competition 1b.

In the second part, Sub-competitions 2a and 2b, the prediction interval scoring rule is 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 $\text{IS}^{(2a)}_{\text{pred},k,t}$, $k = 1, \ldots, K^{(2a)}, t = 1, \ldots, 9$, denote the prediction interval score from team k for testing dataset t. For each testing dataset t, we sort $\text{IS}^{(2a)}_{\text{pred},k,t}$, $k = 1, \ldots, K^{(2a)}$ in ascending order and assign rank $Rank^{(2a)}_{k,t}$ to each team. The final score for team k in Sub-competition 2a is calculated as $S^{(2a)}_k = \sum_{t=1}^{9} Rank^{(2a)}_{k,t}$, and the final rank is assigned by sorting $S^{(2a)}_k$ in ascending order. A similar ranking strategy is used for Sub-competition 2b.

4 Rules

- Teams from any background are welcome to participate. Participant teams can choose to participate in one or more sub-competitions.

- In each sub-competition, all the required results need to be submitted before the deadline to secure a rank.
- Each team is allowed and encouraged to have more than one submission if different methods are used to solve a given problem.

5 Results and Prizes

The ranks of participant teams will be announced on our KAUST web page. The prize for the top-ranked team in each sub-competition will be one Apple iPad. Ties will be broken as deemed suitable by the organizers.

6 Getting Started

Participants should register for the competition by filling in this Registration Google Form: https://forms.gle/D2jT11t6ae3rweLp7 by February 1, 2023. More details and the competition datasets links will be sent to all the registered teams by February 1, 2023.

7 Timeline

Registration will remain open until April 1, 2023, but we will send the datasets to early-registered teams by February 1. All submissions should be received by 11:59pm (UTC±00:00) on April 1, 2023, by filling in this Google Form https://forms.gle/E2XsN6qiie3n8kAJ6.

8 Deliverables

Participants can find an example of the submission files at https://shorturl.at/itDQ5. All required files should be uploaded to Google drive via the Results Submission Google Form, which can be found here: https://forms.gle/E2XsN6qiie3n8kAJ6. The required files are:

- A text description file "TeamName-Description.txt", showing the participating sub-competitions and briefly describing the adopted methods. If the adopted methods

include priors, hyper-parameters, tuning parameters, etc., an explicit description of their choices should be provided. If data preprocessing is used, it should also be explained.

- Other files for each sub-competition described in the following subsections.

8.1 Sub-competitions 1a and 1b

- One "TeamName-1a.csv" or "TeamName-1b.csv" file (delimited by commas) is needed with 5 rows and 8 columns, and an additional header row. The header row is fixed as "lc_sigma_squared, uc_sigma_squared, lc_beta, uc_beta, lc_nu, uc_nu, lc_tau_squared, uc_tau_squared". Then, row i reports the 95% confidence intervals represented as two values for each parameter for dataset i. Each value should be reported with six digits after the decimal point.

8.2 Sub-competition 2a

- One "TeamName-2a-j-t.csv" file (delimited by commas) is needed with 10K rows and 4 columns, and an additional header row, for each training dataset j and testing dataset t. The header row is fixed as "x, y, lp, up." Then, row i reports the 95% prediction interval represented as two values for each testing location i defined by coordinates x, y. Each value should be reported with six digits after the decimal point.

8.3 Sub-competition 2b

- One "TeamName-2b-j-t.csv" file (delimited by commas) is needed with 100K rows and 4 columns, and an additional header row, for each training dataset j and testing dataset t. The header row is fixed as "x, y, lp, up." Then, row i reports the 95% prediction interval represented as two values for each testing location i defined by coordinates x, y. Each value should be reported with six digits after the decimal point.

9 Contact

If you have any questions about this competition, you can contact us at kaustcompspat@gmail.com

References

Abdulah, S., Alamri, F., Ltaief, H., Sun, Y., Keyes, D. E., and Genton, M. G. (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, D. E., and Genton, M. G. (2022b), "The second competition on spatial statistics for large datasets," *Journal of Data Science*, 20, 439–460.

Gneiting, T. and Raftery, A. E. (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, 102, 359–378.

Huang, H., Abdulah, S., Sun, Y., Ltaief, H., Keyes, D. E., and Genton, M. G. (2021a), "Competition on spatial statistics for large datasets," *Journal of Agricultural, Biological and Environmental Statistics*, 26, 580–595.

— (2021b), "Data for competition on spatial statistics for large datasets," http://hdl.handle.net/10754/669153.