# Discussion on Competition for Spatial Statistics for Large Datasets

Lewis R. BLAKE, Olga KHALIUKOVA, Alexander PINARD, Douglas NYCHKA, Dorit HAMMERLING, and Soutir BANDYOPADHYAY

Firstly, we'd like to thank the organizers of the competition for facilitating this compelling investigation. The paper provides a clear overview of modern approximation methods for large spatial data sets. The competition was well structured and executed, and the organizers were professional, timely, and direct in communication. We'd also like to congratulate many of the teams for an impressive performance.

Our team was led by Lewis R. Blake, primarily consisted of graduate students at Colorado School of Mines who had just completed a first course in Spatial Statistics, and was supported by a team of Postdoctoral and Faculty advisors. We saw the competition as an excellent learning opportunity to explore more advanced topics in Spatial Statistics and we were successful in this objective. Competing in this competition was an illuminating firsthand experience in dealing with large spatial problems, and the results will guide in further exploration of the state-of-the-art spatial approximation methods.

This competition differs from similar previous competitions in a few keys ways. Expanding the size of usable data in competitions is a valuable endeavor as is considering both Gaussian and non-Gaussian fields. The decision to allow teams to make multiple submissions made the competition more informative. Comparing less successful approaches to a team's best submission provides insight into a method's performance. For example, an important finding was that the winning model closest to the true relationship did not result in the best out-of-sample predictions. This phenomena illustrates the importance in the choice and implementation of the approximation methods for spatial prediction. In addition, providing results for several similar submissions (such as the $k$-nearest neighbor submissions) helps show how much variance there may be in the implementation of similar approaches.

This competition was valuable by comparing approximation methods for large spatial data knowing the true process and spatial model. There are some shortcomings of a competition

Lewis R. Blake (✉), Olga Khaliukova, Alexander Pinard, Douglas Nychka, Dorit Hammerling,Soutir Bandyopadhyay Colorado School of Mines, Golden, Colorado, USA. (E-mail: *lblake@mines.edu*)
(E-mail: *okhaliukova@mines.edu*), (E-mail: *apinard@mines.edu*), (E-mail: *nychka@mines.edu*),
(E-mail: *hammerling@mines.edu*), (E-mail: *sbandyopadhyay@mines.edu*) .


*Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 4, Pages 596–598*
*https://doi.org/10.1007/s13253-021-00460-4*

design and spatial model based on analyzing synthetic data. For example, the data have characteristics which are not typical in the analysis of large real-world spatial data. The generating process behind these data was based on tools which some of the organizers had developed. Thus, the generative model may have been inferred. If the competition used real data to test predictive performance instead, such a potential issue would not arise.

With regards to modeling, one challenge faced in this competition, which differs from analysis of real-world data, is that a priori incorporation of information about reasonable parameters was not available (e.g., no scientific knowledge could inform reasonable choices for the Matérn smoothness parameter $\nu$). In addition, this competition was based on a very specific pattern of the training and testing locations. The locations of the testing sets were uniformly spaced across the domain, and so each testing point was located near a training point. In such cases, focusing on local structure can yield good predictive performance because dependence on longer-range correlation becomes less important, due to the well-known screening effect. Future competitions may attempt to compare predictive performance when testing sets are swaths of the spatial domain not covered by the training locations.

Our experience in geophysical applications suggests that the larger the spatial data set, the more likely the process will be nonstationary. This competition motivates studying how successful approaches generalize to highly nonstationary data, often encountered in climate research and other geophysical processes. Besides being nonstationary, spatial data sets can be three-dimensional and have a time dimension. It is also worth investigating how the results from this competition could be generalized to these higher dimensional problems.

The organizers have provided the competition data sets to the public as a benchmark for future research. While standardized data sets are common and often required for comparative purposes in the machine learning literature, we believe there may be disadvantages to relying on this type of comparison. For example, we are concerned about "training to the testing data" as publishers and researchers alike are biased to publish positive results on the testing data. Benchmarking against the provided data sets is potentially useful, however, multiple realizations of fields from a given parameter set would be more useful for such a comparison to avoid some of the pitfalls of single training and testing examples.

We found it interesting that many submissions were close in their accuracy, and small differences might be discounted based on other practical issues. Looking at the scales on many of the presented figures, we see that many teams performed very well, and within a small range of each other. It is unclear how much of a practical difference these methods would manifest in suitability for various applications. Moreover, while inconsistent access to, and experience with, state-of-the-art computing facilities was present between teams (and even within our own team), we think it could be informative to have access to an order-of-magnitude comparison between runtimes and predictive performance. This might provide insight into the performance gained versus computational resources required. For example, the compute times for our second submission in sub-competition 2b were about 50 s. This is significantly faster compared to the median runtime across submissions of 2700 s reported by the authors. However, we are unsure of how much improvement is gained by methods requiring runtimes on the order of 2700 s, both because of the lack of comparison and use of rankings to compare submissions. Large differences in runtimes (e.g., wall clock time) are significant because they could be the distinction between interactive data analysis

and a more batch style approach. If possible in future competitions, providing standardize computing facilities (perhaps in the form of access to high-performance computing facilities or Google Colaboratory Notebooks) could aid to provide additional metrics for comparison when computational requirements are taken into account.

*[Received June 2021. Revised June 2021. Accepted July 2021. Published Online August 2021.]*