Discussion on "Competition on Spatial Statistics for Large Datasets"

Quan VU, Yi CAO, Josh JACOBSON, Alan R. PEARSE, and Andrew ZAMMIT- MANGION

The *Competition on Spatial Statistics for Large Datasets* ran in late 2020 and early 2021 and attracted several researchers in spatial statistics, including some in our group at the University of Wollongong, Australia. In this discussion paper, we first summarize our submission to the competition. We then discuss some aspects of the competition and give suggestions for future competitions with regard to the datasets and the assessment methods used.

1. INTRODUCTION

We thank Huang et al. (2021) for organizing the *Competition on Spatial Statistics for Large Datasets* and the editor for the invitation to discuss the paper.

'Big-data' spatial statistics is a key research focus for many spatial statisticians and data scientists in this age of big data and data-driven decisions. The initiative by Huang et al. (2021) is a reflection of the worldwide activity in this research arena, and the number of groups participating in the competition is testament to the importance of this field of research. The competition itself aims to compare different methods for estimation and prediction with large datasets. The authors use the *ExaGeoStat* software to simulate several benchmark datasets and to also estimate and predict using these datasets.

Our discussion paper is divided into two main sections. In Sect. 2, we briefly detail our submission to the competition, while in Sect. 3, we discuss some aspects of the competition: the datasets, the *ExaGeoStat* software, and the methods used to assess the submissions.

2. OUR SUBMISSION TO THE COMPETITION

The competition had four sub-competitions. Each sub-competition itself consisted of a number of datasets, which we refer to as sub-competition datasets. Sub-competition 1a

Quan Vu (✉), Yi Cao, Josh Jacobson, Alan R. Pearse, Andrew Zammit-Mangion School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia
(E-mail: *quanv@uow.edu.au*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 4, Pages 614–618
https://doi.org/10.1007/s13253-021-00464-0

focused on parameter estimation, while sub-competitions 1b, 2a and 2b focused on spatial prediction with large spatial datasets. In each sub-competition, we experimented with several methods via their implementations in R (R Core Team 2020). These methods included local kriging using `gstat` (Pebesma 2004), multi-resolution kriging using `LatticeKrig` (Nychka et al. 2015), nearest-neighbor Gaussian processes (NNGP) using `spNNGP` (Finley et al. 2019), block composite likelihood (BCL; Eidsvik et al. 2014), multi-resolution approximations (MRAs; Huang et al. 2019), and fixed rank kriging using `FRK` (Zammit-Mangion and Cressie 2021). This pool of methods contains a range of approaches for spatial prediction with massive datasets, including low-rank approximations, a composite likelihood approach, and methods that use a small subset of data for each prediction. An internal competition was held to decide which method to use for estimation or prediction in our submission.

In our internal competition, 80% of the data in each sub-competition dataset were selected at random and designated as training data. The remaining 20% were then used as test data. For each of the sub-competition datasets, the prediction method that incurred the smallest root-mean square error (RMSE) between the out-of-sample predictions at the test locations and the test data was selected for our final submission. The parameter estimates submitted for sub-competition 1a were those associated with the prediction method based on the Matérn covariance model that achieved the lowest RMSE in sub-competition 1b. Note that although our internal competition only used 80% of the data within each dataset, our submitted predictions were based on the complete datasets. The majority of our submission was ultimately comprised of predictions using local kriging, with predictions using BCL, NNGP, MRA, and LatticeKrig for a few cases in sub-competitions 1a and 1b. Interestingly, we did use Tukey g-and-h transformations (Xu and Genton 2017) for dealing with non-Gaussian data in sub-competitions 2a and 2b; however, local kriging with no transformation achieved the best predictions of the test data in our internal competition.

Our local kriging method performed well in sub-competitions 2a and 2b (with RMSEs around 3% worse than the top-ranked team). Therefore, we give a brief overview of our setup: We used ordinary kriging (Cressie 1993, pp. 119–123) with a Matérn semivariogram model, which was fitted via weighted least squares to an empirical semivariogram with 50 bins and a maximum separation distance of $\sqrt{2}/2$. In sub-competition 2a, all training data within a $\sqrt{2}/20$ radius of a prediction location were used to predict at that location. In sub-competition 2b, the nearest 1,000 data points were used.

3. DISCUSSION

In this section, we discuss a few key aspects of the competition and give suggestions that may be useful for future initiatives of a similar nature.

3.1. DATASETS

A primary contribution of this work is the availability of the competition datasets, which can be used to benchmark the performance of new approximation methods for large spatial datasets. Thanks to Huang et al. (2021) we now have 20 datasets accompanied by the lowest RMSEs achieved by researchers worldwide when predicting hold-out data using a variety of methods and approaches. Further, two of these datasets are non-Gaussian (generated via the Tukey g-and-h transformation).

Yet, 20 datasets that exhibit several similar properties are a long way from what is needed to genuinely quantify a method, or a group's, ability, to predict at unobserved spatial locations in a general setting. If one wishes to generalize results from a competition to the real world, the benchmark datasets should present a variety of issues that arise in practice. A few realistic datasets have been proposed for competitions. (e.g., Wikle et al. 2017; Heaton et al. 2019), but there remains a need for a widely accepted database of large spatial benchmark datasets that are realistic and also well-documented. Perhaps the most important consideration missing in this competition is that of non-stationarity, which is ubiquitous and challenging in practice. Another missing consideration is the experimental design: data that are regularly spaced and data that are clustered, so that there are large gaps of missing data, should also be considered. Finally, data that are multivariate and spatiotemporal could also be easily considered in a competition where out-of-sample prediction determines the main performance criteria.

3.2. THE *ExaGeoStat* SOFTWARE

The paper of Huang et al. (2021) hinges on using the *ExaGeoStat* software (Abdulah et al. 2018) to generate the datasets used in the competition and to do inference and prediction in competitions 1a and 1b. *ExaGeoStat* acts as a 'gold standard' and is used to benchmark the results of the participants.

The *ExaGeoStat* software has many attractive features that could revolutionize spatial and spatiotemporal statistics in practice as we know it. It was pivotal in this competition, as it allowed us to assess our approximation methods to spatial prediction against exact methods for very large spatial datasets in a variety of settings. There are, however, a few barriers to entry for the spatial statistician. First, this software is of greatest value when high-performance computing (HPC) facilities are available to the user. Second, compiling and installing *ExaGeoStat* in a distributed-processing environment requires considerable HPC expertise. Demonstrations showcasing implementation with, for example, Docker, or a cloud platform, would be very helpful for non-specialist users to begin to experiment with the software.

3.3. ASSESSMENT

The ranking of the submissions in sub-competitions 1b, 2a, and 2b was based on the RMSE. However, this assessment metric only considers point predictions and does not account for the uncertainties associated with those predictions. From our experience, predictions using non-Gaussian or non-stationary models often do not yield point predictions that are materially different from predictions using a standard Gaussian model in terms of RMSE, especially when there are no big (spatial) gaps in the data and when the signal-to-noise ratio is high. For example, in our internal competition for sub-competitions 2a and 2b, we found that the use of a non-Gaussian model with the Tukey g-and-h transformations did not result in better predictive performance in terms of RMSE, while Fuglstad et al. (2015) showed that the use of a non-stationary model might not lead to better point predictions, even when the data are indeed from a non-stationary process. This is unlikely to be the case, however, when assessing the predictive distribution in its entirety: here, considerations of non-stationarity and non-Gaussianity can lead to substantial differences, (e.g., Zammit-Mangion et al. 2021). Ideally, when comparing the predictive performance of different models in a statistical setting, one uses scoring rules that take into account the predictive distribution and not just the prediction. One such scoring rule is the continuous ranked probability score (CRPS; Gneiting and Raftery 2007).

The quality of parameter estimation in sub-competition 1a was assessed using two metrics: the mean loss of efficiency (MLOE) and the mean misspecification of mean square error (MMOM). Both of these metrics do not account for uncertainties in the estimates. However, we also acknowledge that uncertainty in these parameters is often of less direct interest than uncertainty on the process of interest.

Computing time is possibly one of the most important aspects when dealing with large spatial datasets, yet the time required to do prediction or estimation was not assessed in the competition. This is likely due to the fact that different groups all have different computing hardware and environments, making comparison difficult, if not impossible. Heaton et al. (2019) got around this issue by re-running all the code on the same HPC, while a system showcasing an R Software back-end for a competition of this nature is shown in https://hpc.niasra.uow.edu.au/ctf/. Providing a common testbed for computing is also important for equity reasons, as many practitioners may not have access to high-end computing resources. A lack of resources may preclude participants from adopting certain algorithms or methods in their submission to a competition.

ACKNOWLEDGEMENTS

Q.V. and J.J. were each supported by a University Postgraduate Award from the University of Wollongong, Australia. A.R.P. and A.Z.-M. were supported by the Australian Research Council (ARC) Discovery Project DP190100180. A.Z.-M. was also supported by the ARC Discovery Early Career Research Award DE180100203.

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

REFERENCES

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

Cressie N (1993) Statistics for Spatial data. Wiley, New York, NY

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

Finley AO, Datta A, Cook BD, Morton DC, Andersen HE, Banerjee S (2019) Efficient algorithms for Bayesian nearest neighbor Gaussian processes. J Comput Gr Stat 28:401–414

Fuglstad G-A, Simpson D, Lindgren F, Rue H (2015) Does non-stationary spatial data always require non-stationary random fields? Sp Stat 14:505–531

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

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 Agri, Biol Environ Stat 24:398–425

Huang H, Abdulah S, Sun Y, Ltaief H, Keyes D. E., Genton M. G., (2021). Competition on spatial statistics for large datasets. *J Agri, Biol Environ Stat*. https://doi.org/10.1007/s13253-021-00457-z

Huang H, Blake L. R., Hammerling D. M., (2019). Pushing the limit: A hybrid parallel implementation of the multi-resolution approximation for massive data. arXiv:preprint,1905.00141

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution Gaussian process model for the analysis of large spatial datasets. J Comput Graph Stat 24:579–599

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci 30:683–691

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Wikle C. K., Cressie N, Zammit-Mangion A, Shumack C, (2017). A Common Task Framework (CTF) for objective comparison of spatial prediction methodologies. In: Statistics Views, Wiley, Chichester, UK. 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 112:1236–1249

Zammit-Mangion A, Cressie N (2021) FRK: An R package for spatial and spatio-temporal prediction with large datasets. J Stat Softw 98(4):1–48

Zammit-Mangion A, Ng TLJ, Vu Q, Filippone M (2021) Deep compositional spatial models. J Am Stat Assoc. https://doi.org/10.1080/01621459.2021.1887741