Discussion on Competition for Spatial Statistics for Large Datasets

Joseph GUINNESS

Competitions serve many purposes. They focus our efforts to complete a task to the best of our ability, force us into agreement on rules for fair assessment of performance, reward teamwork and constructive interaction, and teach us humility in triumph and defeat. The Competition on Spatial Statistics for Large Datasets was no exception. Thank you to the organizers for running this invigorating event. I'm looking forward to seeing descriptions of the other entrants since I believe that our field can learn a lot about what works and what doesn't from competitions like this one. We learned that GpGp had some inefficiencies in how the predictions were calculated, which we fixed for the competition and included in the next update to the official CRAN version of the package.

Our GpGp and GpGp-quick submissions both used the GpGp R package (Guinness et al. 2021) for estimating parameters and performing predictions. GpGp uses Vecchia's Gaussian process approximation (Vecchia 1988), which has a tuning parameter for the number of neighbors to use in the approximation, with the size of the neighbor sets controlling a trade-off between speed and accuracy. The GpGp submission used 30 neighbors for fitting and 50 for prediction, while GpGp-quick used 20 for fitting and 30 for prediction. In addition, we reduced the computing time for fitting by subsampling the data, 30,000 data points for GpGp and 10,000 for GpGp-quick. All available observations were used for prediction. In sub-competitions 1a and 1b, we used GpGp's "matern_isotropic" covariance function, and for 2a and 2b, we used one of GpGp's nonstationary covariance functions, "matern_nonstat_var," which requires specifying some basis functions for modeling a spatially varying variance. For all datasets, we assumed the mean was an unknown constant.

Here are a few thoughts on the competitions and their results. It seems like there were several entrants in 1a and 1b that produced essentially the same results. There was some variation in the estimates of the variance and range, but it is known that estimates of these two parameters are correlated in the Matérn model (Zhang 2004). The estimates of the smoothness parameter were very close among the well-performing methods. Can we finally retire the idea that the smoothness parameter is difficult to estimate? I will reference your

J. Guinness (✉) Department of Statistics and Data Science, Cornell University, Ithaca, USA
(E-mail: *guinness@cornell.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 26, Number 4, Pages 619–620
https://doi.org/10.1007/s13253-021-00465-z

figure the next time someone makes that claim. The nugget was well estimated in almost every case.

On the prediction problems, the top entrants performed very similarly. I suppose one could argue whether or not a 1% loss of efficiency is meaningfully different than a 2% loss of efficiency. The RMSEs were basically the same for the top teams. It was a bit surprising to see that the losses in efficiency were higher when the model contained a nugget, knowing that the nuggets were well estimated, and the existence of the nugget puts a lower bound on prediction accuracy. Perhaps there is a good explanation for that.

It was great to see datasets that were obviously not from stationary Gaussian processes. We tackled it by fitting a nonstationary Gaussian process, but others appeared to use stationary non-Gaussian models.

The prediction problems were probably a bit too easy, in the sense that the observation locations were very dense relative to the spatial ranges, which is probably why most of the reasonable methods performed similarly. While this competition had a diversity of models, future organizers might consider including a diversity of observation settings, including for example space-time locations following the pattern of a polar orbiting satellite.

Understanding the difficulty of evaluating the timing, I would still be curious to see plots of accuracy (vertical axis) versus self-reported timing (horizontal axis) for each of the teams for each of the datasets. Such plots are useful for evaluating the trade-off between speed and accuracy.

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

REFERENCES

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

Vecchia AV (1988) Estimation and model identification for continuous spatial processes. J R Stat Soc Ser B (Methodol), 50(2):297–312

Zhang H (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J Am Stat Assoc 99(465), 250–261