Journal of Data Science 20 (4), 437–438
October 2022
DOI: 10.6339/22-JDS204EDI
Editorial

Editorial: Large-Scale Spatial Data Science

Sameh Abdulah[1], Stefano Castruccio[2], Marc G. Genton[1,3,*], and Ying Sun[1,3]

[1]*Extreme Computing Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

[2]*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA*

[3]*Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

This special issue features eight articles on "Large-Scale Spatial Data Science." Data science for complex and large-scale spatial and spatio-temporal data has become essential in many research fields, such as climate science and environmental applications. Due to the ever-increasing amounts of data collected, traditional statistical approaches tend to break down and computationally efficient methods and scalable algorithms that are suitable for large-scale spatial data have become crucial to cope with many challenges associated with big data. This special issue aims at highlighting some of the latest developments in the area of large-scale spatial data science. The research papers presented showcase advanced statistical methods and machine learning approaches for solving complex and large-scale problems arising from modern data science applications.

Abdulah et al. (2022) reported the results of the second competition on spatial statistics for large datasets organized by the King Abdullah University of Science and Technology (KAUST). Very large datasets (up to 1 million in size) were generated with the *ExaGeoStat* software to design the competition on large-scale predictions in challenging settings, including univariate nonstationary spatial processes, univariate stationary space-time processes, and bivariate stationary spatial processes. The authors described the data generation process in detail in each setting and made these valuable datasets publicly available. They reviewed the methods used by fourteen competing teams worldwide, analyzed the results of the competition, and assessed the performance of each team.

Chakraborty and Katzfuss (2022) tackled the challenging problem of high-dimensional nonlinear spatio-temporal filtering. They proposed a scalable method based on compressing and decompressing a hierarchical sparse Cholesky factorization of the filtering covariance matrix. Because the sparsity was preserved over time, their approach yielded fast computations and low memory footprints. They demonstrated the superiority of their method compared to state-of-the-art competitors on a high-dimensional and nonlinear Lorenz model.

Fahmy and Guinness (2022) proposed their implementation of the Matérn cross-covariance model for multivariate spatial random fields for large datasets in the *GpGp* package of the *R* software, a package geared to fast fitting to large datasets. The key idea of this work is to use Vecchia's approximation and an optimization algorithm with Fisher scoring. The authors considered various parameterizations of the model from the literature and tested their implementation on several multivariate spatial datasets from practical applications.

Gray et al. (2022) studied the use of deep neural networks for large-scale spatial predictions. They compared them with Gaussian processes in terms of computational efficacy and predictive power on various big spatial datasets, including Gaussian, non-Gaussian and binary data.

*Corresponding author. Email: marc.genton@kaust.edu.sa.

They concluded that neural networks are a viable approach for prediction in large spatial data problems, although more research is needed to develop appropriate uncertainty quantifications.

Pereira et al. (2022) put forward a matrix-free approach for large-scale geostatistics on Riemannian manifolds. Their proposal bridged the gap between nonstationary Gaussian random fields and random fields on manifolds such as the sphere. They provided scalable algorithms for parameter estimation and optimal prediction by kriging, and they illustrated their method on very large synthetic datasets.

Saha et al. (2022) built a scalable prediction method for spatial probit linear mixed models using nearest neighbor Gaussian processes. Their method only involved sparse or small matrix computations and could be implemented in an embarrassingly parallel manner. They demonstrated the accuracy and scalability of their approach via simulations and an analysis of species presence-absence data.

Wendelberger et al. (2022) proposed a multiresolution robust online Bayesian monitoring algorithm for spatial characteristics of gapless remote sensing big data. They demonstrated through a simulation study the superiority of their approach for detecting subtle changes in images. Furthermore, they illustrated the applicability of their method in detecting construction changes in two cities from the USA and UAE.

Daw and Wikle (2022) presented a supervised spatial regionalization method based on the Karhunen-Loève expansion and minimum spanning trees. It aimed at minimizing the *ecological fallacy* across spatial scales in order to partition spatial domains and ensuring spatial contiguity. The authors demonstrated their approach on simulated data and an ocean color dataset.

All articles in this special issue were peer-reviewed by 2–3 reviewers. We are grateful to the authors for their outstanding contributions and to the anonymous reviewers for their constructive feedback on the manuscripts. We hope that this special issue will be of interest to all data scientists.

References

Abdulah S, Alamri F, Nag P, Sun Y, Ltaief H, Keyes DE, Genton MG (2022). The second competition on spatial statistics for large datasets. *Journal of Data Science*, 20(4): 439–460.

Chakraborty A, Katzfuss M (2022). High-dimensional nonlinear spatio-temporal filtering by compressing hierarchical sparse Cholesky factors. *Journal of Data Science*, 20(4): 461–474.

Daw R, Wikle CK (2022). Supervised spatial regionalization using the Karhunen-Loève expansion and minimum spanning trees. *Journal of Data Science*, 20(4): 566–584.

Fahmy Y, Guinness J (2022). Vecchia approximations and optimization for multivariate Matérn models. *Journal of Data Science*, 20(4): 475–492.

Gray SD, Heaton MJ, Bolintineanu DS, Olson A (2022). On the use of deep neural networks for large-scale spatial prediction. *Journal of Data Science*, 20(4): 493–511.

Pereira M, Desassis N, Allard D (2022). Geostatistics for large datasets on Riemannian manifolds: A matrix free approach. *Journal of Data Science*, 20(4): 512–532.

Saha A, Datta A, Banerjee S (2022). Scalable predictions for spatial probit linear mixed models using nearest neighbor Gaussian processes. *Journal of Data Science*, 20(4): 533–544.

Wendelberger LJ, Gray JM, Wilson AG, Houborg R, Reich BJ (2022). Multiresolution broad area search: Monitoring spatial characteristics of gapless remote sensing data. *Journal of Data Science*, 20(4): 545–565.