TEST (2019) 28:338–341
https://doi.org/10.1007/s11749-019-00642-w

DISCUSSION

# Comments on: Data science, big data and statistics

**Marc G. Genton**[1] · **Ying Sun**[1]

Published online: 8 April 2019


**Mathematics Subject Classification** 62M30 · 62H30

## 1 Introduction

We would like to start by congratulating the authors for a very timely and stimulating paper. They have provided thought-provoking ideas on Data Science and Big Data, and on how Statistics must play a major role in these new areas. We focus our discussion on two points that have caught our attention and interest: visualization and computations for new sources of information.

### 1.1 Visualization for new sources of information

Traditionally, Statistics has dealt with scalar and vectorial observations. However, as noted by the authors, advances in technology have greatly facilitated the collection of large-scale high-dimensional data in many research fields. Among various types of high-dimensional data, spatiotemporal data and functional data have been particularly popular. Classical statistical methodologies face many challenges for such datasets because they often contain massive amounts of observations, non-Gaussian features, and they may exhibit complex spatiotemporal dynamics. In recent years, functional data analysis techniques have been increasingly used for space–time data, where data are typically considered to be functional but are correlated in time and/or space. However, the data structure can be even more complicated, for instance, functional data may be multivariate and observed along time from different geographical locations.

This comment refers to the invited paper available at: https://doi.org/10.1007/s11749-019-00651-9.

✉ Marc G. Genton
marc.genton@kaust.edu.sa

Ying Sun
ying.sun@kaust.edu.sa

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

Although there have been extensive developments in both fields of spatial statistics and functional data analysis, many problems related to visualization, model diagnostics and complex real data applications still remain unexplored.

As a popular tool for exploratory data analysis, visualization has proved its value in Statistics, yet it remains a challenge for large complex datasets. Sun and Genton (2011) first proposed the functional boxplot as an analogue to the classical boxplot to address functional data visualization. It has been extended and applied to different types of high-dimensional datasets and has become a valuable tool in exploratory data analysis and model diagnostics. For example, Sun and Genton (2012) proposed adjusted functional boxplots for spatiotemporal data visualization and outlier detection, while Huang and Sun (2019) used functional boxplots for visualization and assessment of spatiotemporal covariance properties. Xie et al. (2017) developed a geometric framework to decompose and visualize functional data based on amplitude, phase and vertical translation. Functional boxplots for multivariate curves were proposed by Dai and Genton (2018a) based on directional outlyingness of multivariate functional data (Dai and Genton 2019). Other graphics for multivariate functional data visualization and outlier detection can be found in Dai and Genton (2018b).

The authors also pointed out the importance of analyzing heterogeneous data. We strongly agree that heterogeneity is a standard feature for Big Data and methods and tools for heterogeneous data cannot be overemphasized. For cluster analysis, we would like to discuss two aspects when investigating complex spatiotemporal phenomena: cluster detection in spatial data and time series clustering. In spatial regression analysis, heterogeneity refers to spatial units that show distinctive patterns. Typically, the goal is to identify important predictor variables or regimes of predictors that lead to certain response patterns (Lee et al. 2017; Gangnon and Clayton 2004). The mixture models and clustering algorithms reviewed by the authors are valuable if one can devise them for spatially indexed large datasets. When a large collection of high-frequency time series data is available, the dataset is often treated as a series of short stationary time series. Time series clustering can be used to identify homogeneous temporal regimes. One approach is to use spectral features of times series for classification or clustering purposes. The general goal is to identify clusters of time series. Within each cluster, the time series show similar spectral densities or oscillatory patterns. Recently, many new and efficient clustering algorithms have been developed for spectral density functions (Euán et al. 2018, 2019). However, visualization of the identified time series clusters is not trivial. We would like to link the functional data visualization tools to the clustering setting. For example, the functional boxplot can be used to visualize spectral densities from each cluster and summarize their distinct features. Euán and Sun (2019) treated directional distributions of the ocean waves as functional data and constructed a directional functional boxplot to display the main directional distribution of the wave energy within a cluster.

Finally, we second the authors' claim that visuanimations (Genton et al. 2015), the embedding of animations and movies in papers, will play a major role in describing new sources of information. In addition, the development of apps for the visualization (stereoscopic view) and dissemination of results from Data Science and Big Data analyses, is an important step forward; see Castruccio et al. (2019) and the discussions

therein. This may prove especially useful to visualize network data and their evolution over time.

### 1.2 Computations for new sources of information

New sources of information bring large datasets and associated computational issues. This necessitates new frameworks to deal with these challenges. For example, spatiotemporal Big Data arising from the output of climate models require new principles for statistical inference (Castruccio and Genton 2018). Similarly, Big Data from fMRI experiments to study spatiotemporal brain activation and connectivity demand scalable multi-resolution models (Castruccio et al. 2018). Even the functional boxplot requires fast ranking of curves and images (Sun et al. 2012a) to be applicable to very large datasets.

When the computations for Big Data become prohibitive, one can use approximated models that allow for exact computations, or devise computational methods to approximate the exact solution, or take advantage of modern computational resources such as high-performance computing (HPC). Take the widely used Gaussian process model as an example. Gaussian process models face tremendous computational challenges for Big Data, such as fitting a Gaussian process model to large spatial datasets in geostatistics, solving multivariate input regression problems in computer experiments, fitting nonparametric regression models via Gaussian processes in Bayesian analysis, performing regression and classification tasks in machine learning and so on. The major computational cost comes from evaluating the high-dimensional Gaussian likelihood for statistical inferences on unknown covariance structures. It requires the Cholesky decomposition of the covariance matrix to find exact solutions. However, for unstructured and dense covariance matrices, it is prohibitive with current memory and computation resources. In the past decade, many efforts have been made to improve the computational efficiency via approximated models, many of which have been reviewed by Sun et al. (2012b), such as covariance tapering and Gaussian Markov random fields. Recent developments include multi-resolution methods (Nychka et al. 2015), nearest-neighbor approaches (Datta et al. 2016) and hierarchical low rank approximations (Huang and Sun 2018). There is also a rich literature on taking advantage of computational tools in numerical analysis to approximate the exact covariance matrix (Abdulah et al. 2018a; Baugh and Stein 2018). Recently, HPC has been firstly introduced to calculate the maximum likelihood estimator exactly and perform kriging prediction for exascale spatial data (Abdulah et al. 2018b). These significant contributions greatly benefit from interdisciplinary research, especially in the Big Data era; hence, collaborations are essential. We believe that bridging different communities will make significant impacts to statistical computing.

## References

Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018a) Parallel approximation of the maximum likelihood estimation for the prediction of large-scale geostatistics simulations. In: IEEE Int Conf Clust Comput, pp 98–108

Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018b) ExaGeoStat: a high performance unified software for geostatistics on manycore systems. IEEE Trans Parallel Distrib Syst 29:2771–2784
Baugh S, Stein ML (2018) Computationally efficient spatial modeling using recursive skeletonization factorizations. Spat Stat 27:18–30
Castruccio S, Genton MG (2018) Principles for statistical inference on big spatio-temporal data from climate models. Stat Probab Lett 136:92–96
Castruccio S, Ombao H, Genton MG (2018) A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data. Biometrics 74:823–833
Castruccio S, Genton MG, Sun Y (2019) Visualising spatio-temporal models with virtual reality: from fully immersive environments to apps in stereoscopic view. J R Stat Soc A Stat 182:379–387
Dai W, Genton MG (2018a) Functional boxplots for multivariate curves. Stat 7:e190
Dai W, Genton MG (2018b) Multivariate functional data visualization and outlier detection. J Comput Graph Stat 27:923–934
Dai W, Genton MG (2019) Directional outlyingness for multivariate functional data. Comput Stat Data Anal 131:50–65
Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J Am Stat Assoc 111:800–812
Euán C, Sun Y (2019) Directional spectra-based clustering methods for visualizing patterns of winds and waves in the Red Sea. J Comput Graph Stat. https://doi.org/10.1080/10618600.2019.1575745
Euán C, Ombao H, Ortega J (2018) The hierarchical spectral merger algorithm: a new time series clustering procedure. J Classif 35:71–99
Euán C, Sun Y, Ombao H (2019) Coherence-based time series clustering for statistical inference and visualization of brain connectivity. Ann Appl Stat (**to appear**)
Gangnon RE, Clayton MK (2004) Likelihood-based tests for detecting spatial clustering of disease. Environmetrics 15:797–810
Genton MG, Castruccio S, Crippa P, Dutta S, Huser R, Sun Y, Vettori S (2015) Visuanimation in statistics. Stat 4:81–96
Huang H, Sun Y (2018) Hierarchical low rank approximation of likelihoods for large spatial datasets. J Comput Graph Stat 27:110–118
Huang H, Sun Y (2019) Visualization and assessment of spatio-temporal covariance properties. Spat Stat. https://doi.org/10.1016/j.spasta.2017.11.004
Lee J, Gangnon RE, Zhu J (2017) Cluster detection of spatial regression coefficients. Stat Med 27:110–118
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
Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20:316–334
Sun Y, Genton MG (2012) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23:54–64
Sun Y, Genton MG, Nychka D (2012a) Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? Stat 1:68–74
Sun Y, Li B, Genton MG (2012b) Geostatistics for large datasets, Chap 3. In: Porcu E, Montero JM, Schlather M (eds) Space-time processes and challenges related to environmental problems, vol 207. Springer, Berlin, pp 55–77
Xie W, Kurtek S, Bharath K, Sun Y (2017) A geometric approach to visualization of variability in functional data. J Am Stat Assoc 112:979–993