Math Geosci (2022) 54:459–465
https://doi.org/10.1007/s11004-022-09998-6

SPECIAL ISSUE

Special Issue: Geostatistics and Machine Learning

Sandra De Iaco[1] · **Dionissios T. Hristopulos**[2] · **Guang Lin**[3]

Received: 14 February 2022 / Accepted: 15 February 2022 / Published online: 21 March 2022

Abstract Recent years have seen a steady growth in the number of papers that apply machine learning methods to problems in the earth sciences. Although they have different origins, machine learning and geostatistics share concepts and methods. For example, the kriging formalism can be cast in the machine learning framework of Gaussian process regression. Machine learning, with its focus on algorithms and ability to seek, identify, and exploit hidden structures in big data sets, is providing new tools for exploration and prediction in the earth sciences. Geostatistics, on the other hand, offers interpretable models of spatial (and spatiotemporal) dependence. This special issue on *Geostatistics and Machine Learning* aims to investigate applications of machine learning methods as well as hybrid approaches combining machine learning and geostatistics which advance our understanding and predictive ability of spatial processes.

Keywords Geostatistics · Statistical learning · Machine learning · Spatial process · Gaussian process regression

✉ Sandra De Iaco
sandra.deiaco@unisalento.it

Dionissios T. Hristopulos
dchristopoulos@ece.tuc.gr

Guang Lin
guanglin@purdue.edu

[1] Department of Economic Sciences, Sect. of Mathematics and Statistics, University of Salento, Lecce, Italy

[2] School of Electrical and Computer Engineering, Technical University of Crete, 73100 Chania, Greece

[3] Department of Mathematics & School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA

1 Introduction

This special issue explores connections between *Geostatistics and Machine Learning*, and their applications in spatial data processing and modeling. Applications of machine learning in the geosciences have become quite popular in recent years following the development of tools such as random forests and deep learning. A search in the databases of the IAMG journals *Mathematical Geosciences* and *Computers and Geosciences* with the keyword "machine learning" returns 105 and 319 hits, respectively. The majority of these contributions are dated after 2016, a fact which indicates the accelerating interest in machine learning for the analysis of spatial data. In the following paragraphs of this section we present a partial and undoubtedly biased account of some links between geostatistics and machine learning. We also briefly report on some recent developments in machine learning which we believe are relevant for the geosciences. Finally, we touch on remaining methodological and computational challenges.

The application of machine learning to earth science data has been spearheaded by Mikhail Kanevski and coworkers (Demyanov et al. 1998; Kanevski et al. 2004, 2009; Kanevski and Demyanov 2015). In recent years, following the increasing interest in machine learning, several review papers have discussed its potential uses in geosciences and remote sensing (Dramsch 2020; Karpatne et al. 2019; Lary et al. 2016; Shen et al. 2021). A nontechnical account which focuses on the challenges related to the extraction of information from earth science data sets and the opportunities created by machine learning is given in Maskey et al. (2020).

Modeling of spatial data in the earth sciences usually involves one of the following three fundamental problems: (i) the classification problem which concerns predicting the class label for categorical data; (ii) the regression problem which is related to the prediction of continuous data; and (iii) the problem of probability density function estimation for uncertain processes (Williams and Rasmussen 2006; Kanevski et al. 2009). These problems can be addressed by means of geostatistical methods or machine learning models and algorithms, or in terms of combined solutions. Both machine learning and geostatistics provide powerful frameworks for spatial data processing. Combinations of these two approaches can lead to flexible and computationally efficient spatial models, as some of the papers in this special issue highlight.

As mentioned in the abstract, certain machine learning methods share concepts with geostatistical approaches. For example, geostatistical interpolation by means of optimal linear estimation (kriging) and Gaussian process regression are both based on the theory of Gaussian random fields (Gaussian processes) (Adler and Taylor 2009; Yaglom 1987; Chilès and Delfiner 2012; Williams and Rasmussen 2006). Positive definite functions (i.e., "covariance functions" in geostatistics and "covariance kernels" in machine learning) play a key role in problems of interpolation, classification, clustering, and simulation, whether these are treated in the geostatistical or in the machine learning framework. Machine learning, however, also includes methods which are based on algorithms (procedures consisting of specific steps) instead of explicitly defined mathematical models.

A key issue that both machine learning and geostatistical approaches need to address in the face of big earth data sets is the scaling of the required computational resources

with the data size N. For example, regression and classification tasks with spatially dependent data require the inversion of dense covariance (Gram) matrices, an operation which has a computational complexity of $\mathcal{O}(N^3)$. This scaling affects both kriging and Gaussian process-based methods, and it is prohibitive for very big data sets (Chilès and Delfiner 2012; Hristopulos 2020; Williams and Rasmussen 2006). The problem can be alleviated by means of different methods such as the stochastic partial differential equation (SPDE) approach that relies on a sparse solution basis (Lindgren et al. 2011, 2021; Vergara et al. 2022), the stochastic local interaction approach (Hristopulos 2015; Hristopulos et al. 2021) that exploits sparse expressions for the precision (inverse covariance) matrix, or composite likelihood methods that break down the calculation of the likelihood in terms of smaller subsets of the data (Bevilacqua et al. 2012).

Neural networks are a cornerstone of modern machine learning. These models can be trained to discover features which are hidden in high-dimensional data sets. Neural networks comprise a large number of parameters which need to be tuned, so overfitting is a likely problem. However, this is avoided within the Bayesian framework by assigning probability distributions (instead of single values) to the weights of the connections between different neurons. Bayesian neural networks have the ability to capture cross-correlations and are therefore potentially useful in problems that involve data with spatial or spatiotemporal dependence. A Bayesian neural network contains a number of hidden layers where information is processed. "Deep neural networks" involve a high number of such layers. Surprisingly, the limit of an infinitely deep neural network is a Gaussian process (Neal 1996). Neural networks that are not infinitely deep can capture correlations between different output variables. This feature can lead to improved spatial prediction in the case of multivariate data sets (Wilson et al. 2011). A well-known spatial data set from the Swiss Jura Mountains comprises measurements of soil concentration for seven toxic metals (Goovaerts 1997). The Gaussian process regression network (GPRN) developed by Wilson et al. (2011) predicted cadmium concentration more accurately (i.e., with lower mean absolute error) than co-kriging. Improved Gaussian process regression models for multivariate problems (called "multi-output" in machine learning jargon) have since been developed; these include the Gaussian process autoregressive regression model (GPAR) (Requeima et al. 2019) and the multi-output Gaussian processes (MOGPs) (Bruinsma et al. 2020). To our knowledge, the strength of these methods has not yet been investigated in earth sciences applications.

An important topic in geosciences is the spatiotemporal modeling of dynamic environmental processes. Reliable conceptual and quantitative models are necessary to achieve improved understanding, to better forecast potential environmental hazards, and to quantify uncertainties. This general problem can be pursued by means of two different approaches. The first one involves data-driven spatiotemporal prediction (e.g., regression and classification) by means of geostatistical and machine learning methods. One of the open problems is the formulation of epistemically adequate and computationally efficient methods of characterizing spatiotemporal dependence (Christakos 2000; De Iaco et al. 2001, 2002; Cappello et al. 2018; Hristopulos and Agou 2020; Cappello et al. 2020; Porcu et al. 2021). Addressing this problem requires the construction of space-time covariance functions or precision operators which are mathematically well defined and capture the dynamically generated correlations of

realistic space-time systems. The issue of proper definition of covariance functions (i.e., functions that satisfy permissibility conditions) is well known in mathematics and statistics (De Iaco and Posa 2018), but it is not always recognized in the applied sciences literature. This oversight can lead to the use of non-permissible covariance models which result in numerical instabilities. Computational efficiency requires the implementation of methods that can alleviate the problem of numerical inversion of very large matrices resulting from extended space-time domains.

A different approach to spatiotemporal modeling involves the solution of partial differential equations that model specific earth processes (e.g., transport of contaminants in the groundwater, or geophysical fluid dynamics). Machine learning is providing new tools, such as physics-inspired neural networks (PINNs) (Karniadakis et al. 2021; Yang et al. 2021), for this type of problems. In the PINN framework, deep neural networks are trained using a combination of data and constraints imposed by the physical laws. This hybrid framework gives more weight to the model of the system when the data are sparse, but progressively shifts focus to the data when the latter are abundant. PINNs can be used for both forward and inverse as well as high-dimensional problems.

In the next section, we present a short introduction to the six articles of this special issue on *Geostatistics and Machine Learning*. The topics covered in these papers represent an eclectic selection of practical problems and machine learning approaches used to tackle them. The contributions of the special issue also present fertile combinations of machine learning and geostatistical methods tailored to address problems that involve spatial dependence.

Summary of Articles in this Special Issue

The paper titled "A comparison between machine learning and functional geostatistics approaches for data-driven analyses of solid transport in a pre-Alpine stream" by Oleksandr Didkovskyi et al. focuses on predicting the probability of pebble movement in streams using two different approaches: the machine learning method of gradient boosting decision trees (based on the computationally efficient XGBoost algorithm) and the geostatistical method of functional kriging. Both approaches take into account geometrical features of pebbles and the stream flow rate as input variables. The performance of the two methods is compared in terms of the accuracy with which they classify the motion (or lack of mobility) of pebbles. The probability of movement has a highly nonlinear dependence on the morphological features and the stream's flow rate and is thus difficult to predict using physics-based methods. In spite of the quite different perspectives of XGBoost and functional kriging, analysis of the results shows that both methods perform similarly well and can provide useful modeling frameworks for sediment transport.

The paper titled "Bayesian deep learning for spatial interpolation in the presence of auxiliary information" by Charlie Kirkwood et al. focuses on feature learning in a geostatistical context, by showing how deep neural networks can automatically learn the complex high-order patterns by which point-sampled target variables relate to gridded auxiliary variables, and in doing so produce detailed maps. This work demonstrates how both aleatoric and epistemic uncertainty can be quantified in the

deep learning approach via a Bayesian approximation known as Monte Carlo dropout. Numerical results indicate the suitability of Bayesian deep learning and its feature learning capabilities for large-scale geostatistical applications.

The paper titled "Surface Warping Incorporating Machine Learning Assisted Domain Likelihood Estimation: A New Paradigm in Mine Geology Modelling and Automation" by Raymond Leung et al. introduces the use of machine learning to support a Bayesian warping technique applied to reshape modeled surfaces on the basis of new geochemical observations and spatial constraints. This helps to improve the identification of boundaries of different spatial domains for grade estimation in mining, which represents a complex problem, set in a Bayesian framework. A strength of the manuscript is the assessment of the effectiveness of a range of classifiers. Indeed, the machine learning performance is computed for neural network, random forest, gradient boosting, and other classifiers in a binary and multi-class context. The manuscript represents progress in this evolving field, and further research will continue to address the problems presented.

The paper titled "A Hybrid Estimation Technique Using Elliptical Radial Basis Neural Networks and Cokriging" by Matthew Samson and Clayton V. Deutsch focuses on a hybrid machine learning and geostatistical algorithm to improve estimation in complex domains. The hybrid estimation technique integrates both elliptical radial basis neural networks and cokriging. Elliptical radial basis function neural networks (ERBFN) take advantage of nonstationary functions to generate geological estimates. An ERBFN does not require the assumption of stationarity, and the only input features required are the spatial coordinates of the known data. The proposed hybrid estimation considers the machine learning estimate as exhaustive secondary data in ordinary intrinsic collocated cokriging, taking advantage of kriging's exactitude while including the nonstationary features modeled in the ERBFN. The numerical results demonstrate that this hybrid method can greatly improve mineral resource estimation.

The paper titled "Stochastic Modelling of Mineral Exploration Targets" by Hasan Talebi et al. focuses on the topic of mineral prospectivity mapping and proposes a method that can handle various types of uncertainties. The authors propose a multivariate stochastic model which can be used for prediction and uncertainty quantification of mineral exploration targets. The model combines multivariate geostatistical simulations with a spatial machine learning (random forest) algorithm. The latter incorporates information from higher-order spatial statistics. The proposed approach is tested using a synthetic case study with multiple geochemical, geophysical, and lithological attributes. The new hybrid (geostatistics/machine learning) method demonstrates enhanced detection capabilities and thus provides a promising tool for investigating mineral prospectivity.

The paper titled "Robust Feature Extraction for Geochemical Anomaly Recognition Using a Stacked Convolutional Denoising Autoencoder" by Yihui Xiong and Renguang Zuo focuses on an optimized deep neural network for the recognition of multivariate geochemical anomalies, especially in the presence of missing values. In particular, the authors propose a stacked convolutional denoising autoencoder (SCDAE) to extract robust features and decrease the sensitivity to partially corrupted data. The corresponding parameters, which include the network depth, number of convolution layers, number of pooling layers, number of filters, and their respective sizes

(i.e., the number of convolution kernels, convolution kernel size, number of pooling kernels, and pooling kernel size), and the sliding stride, are optimized using trial-and-error experiments. The performance of the optimal SCDAE architecture in recognizing multivariate geochemical anomalies, based on the differences in the reconstruction errors between sample populations, is discussed through a case study regarding the mineralization in the southwestern Fujian Province. They also show that SCDAE has a better feature representation capacity than both the stacked convolutional autoencoder and stacked denoising autoencoder for geochemical anomaly recognition with different corruption levels. The robustness of the SCDAE encourages its application to various geochemical exploration scenarios, especially when there are incomplete or missing data.

Acknowledgements GL gratefully acknowledges the support from the National Science Foundation (DMS-1555072, DMS-1736364, CMMI-1634832, and CMMI-1560834), and the Brookhaven National Laboratory Subcontract 382247, ARO/MURI grant W911NF-15-1-0562, and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program DE-SC0021142.

Funding Open access funding provided by Universitá degli Studi di Milano within the CRUI-CARE Agreement.

References

Adler RJ, Taylor JE (2009) Random fields and geometry. Springer, Berlin

Bevilacqua M, Gaetan C, Mateu J, Porcu E (2012) Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. J Am Stat Assoc 107(497):268–280

Bruinsma W, Perim E, Tebbutt W, Hosking S, Solin A, Turner R (2020) Scalable exact inference in multi-output Gaussian processes. In: Daumé H, Singh A (eds) Proceedings of the 37th international conference on machine learning, volume 119 of Proceedings of Machine Learning Research, PMLR, pp 1190–1201

Cappello C, De Iaco S, Posa D (2018) Testing the type of non-separability and some classes of space-time covariance function models. Stoch Environ Res Risk Assess 32:17–35

Cappello C, De Iaco S, Posa D (2020) covatest: an R package for selecting a class of space-time covariance functions. J Stat Softw 94(1):1–42

Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, New York

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, Oxford

De Iaco S, Myers DE, Posa D (2001) Space-time analysis using a general product-sum model. Stat Probab Lett 52(1):21–28

De Iaco S, Myers DE, Posa D (2002) Nonseparable space-time covariance models: some parametric families. Math Geol 34(1):23–42

De Iaco S, Posa D (2018) Strict positive definiteness in geostatistics. Stoch Environ Res Risk Assess 32:577–590

Demyanov V, Kanevsky M, Chernov S, Savelieva E, Timonin V (1998) Neural network residual kriging application for climatic data. J Geogr Inf Decis Anal 2(2):215–232

Dramsch JS (2020) 70 years of machine learning in geoscience in review. Adv Geophys 61:1–55
Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, NY
Hristopulos DT (2015) Stochastic local interaction (SLI) model: Bridging machine learning and geostatistics. Comput Geosci 85(Part B):26–37
Hristopulos DT (2020) Random fields for spatial data modeling. Springer, Dordrecht
Hristopulos DT, Agou VD (2020) Stochastic local interaction model with sparse precision matrix for space–time interpolation. In: spatial Statistics 40:100403, space-time modeling of rare events and environmental risks: METMA conference
Hristopulos DT, Pavlides A, Agou VD, Gkafa P (2021) Stochastic local interaction model: an alternative to kriging for massive datasets. Math Geosci 53:1907–1949
Kanevski M, Demyanov V (2015) Statistical learning in geoscience modelling: novel algorithms and challenging case studies. Comput Geosci 85:1–2
Kanevski M, Kanevski MF, Maignan M (2004) Analysis and modelling of spatial environmental data, vol 6501. EPFL Press, Lausanne
Kanevski M, Timonin V, Pozdnukhov A (2009) Machine learning for spatial environmental data: theory, applications, and software. EPFL Press, Lausanne
Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L (2021) Physics-informed machine learning. Nature Reviews Nat Rev Phys 3(6):422–440
Karpatne A, Ebert-Uphoff I, Ravela S, Babaie HA, Kumar V (2019) Machine learning for the geosciences: challenges and opportunities. IEEE Trans Knowl Data Eng 31(8):1544–1554
Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geosci Front 7(1):3–10
Lindgren F, Bolin D, Rue H (2021) The spde approach for gaussian and non-gaussian fields: 10 years and still running
Lindgren F, Rue H, Lindström J (2011) An explicit link between gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J R Stat Soc Ser B (Stat Methodol) 73(4):423–498
Maskey M, Alemohammad H, Murphy K, Ramachandran R (2020) Advancing AI for Earth science: a data systems perspective. Eos 101
Neal RM (1996) Bayesian learning for neural networks, vol 118. Springer, New York
Porcu E, Furrer R, Nychka D (2021) 30 years of space-time covariance functions. WIREs Comput Stat 13(2):e1512
Requeima J, Tebbutt W, Bruinsma W, Turner R E (2019) The gaussian process autoregressive regression model (gpar). In: Chaudhuri K, Sugiyama M (eds) Proceedings of the twenty-second international conference on artificial intelligence and statistics, volume 89 of Proceedings of Machine Learning Research, PMLR, pp 1860–1869
Shen C, Chen X, Laloy E (2021) Editorial: Broadening the use of machine learning in hydrology. Frontiers in Water 3
Vergara RC, Allard D, Desassis N (2022) A general framework for SPDE-based stationary random fields. Bernoulli 28(1):1–32
Williams CKI, Rasmussen CE (2006) Gaussian processes for machine learning. MIT Press, Cambridge, MA
Wilson A G, Knowles D A, Ghahramani Z (2011) Gaussian process regression networks. arXiv preprint arXiv:1110.4411
Yaglom AM (1987) Correlation theory of stationary and related random functions, vol I. Springer, New York
Yang L, Meng X, Karniadakis GE (2021) B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. J Comput Phys 425:109913