The Importance of Geostatistics in the Era of Data Science

Atkinson, P. (Accepted/In press). The Importance of Geostatistics in the Era of Data Science. *Mathematical Geosciences*, 1-2.

Published in:
Mathematical Geosciences

Document Version:
Peer reviewed version

The Importance of Geostatistics in the Era of Data Science

(MATG geoENV2018 Special Issue) https://doi.org/10.1007/s11004-020-09858-1

Jennifer McKinley, School of Natural and Built Environment, Queen's University Belfast, UK

Peter M. Atkinson, Lancaster Environment Centre, Lancaster University, UK

Introduction

Since the development of geostatistics in the 1960s, geostatistical techniques have been applied widely, from early advances in estimating mineral and oil/gas reserves (David 1977; Journel and Huijbregts 1978) to a range of environmental problems including in soil science, hydrology, ecology, climatology and oceanography, more recently (Cressie 1993; Goovaerts 1997). This expansion in the use of geostatistics stems, crucially, from the spatially (and temporally) explicit nature of the methods which geostatistics provides for the appropriate analysis of Earth science properties, which vary in space and time

Geostatistical techniques were developed to enable geoscientists to characterise spatial properties and operate on those properties. Characterisation, differently to classical (non-spatial) statistics, includes the spatial dependence of properties and their joint distribution where the data are multivariate. It can, commonly, include the spatial support (the size, geometry and orientation of the space on which observations are made) and the principled handling of different data types, including compositional data (Talebi et al. 2019). Operations centre on prediction via kriging (the best linear unbiased predictor) and its variants, and the modelling of prediction uncertainty, but also include extension to simulation and modelling of (joint) spatial uncertainty, spatial regularization (change to the spatial support of the stochastic process), and optimization of spatial sampling design, amongst others. Bayesian inference importantly has allowed estimation of model uncertainty.

With the emergence of the era of *big data science*, data-oriented approaches that capitalise on the richness of data available have been preferred over model-based approaches (including geostatistics) which have been perceived to be limited in flexibility and too computationally intensive (and, thus, too slow) to handle big datasets. This drive towards data-oriented approaches has undoubtedly brought tremendous innovation in the field of machine-learning and, more recently, deep learning; e.g., Zhang et al. 2019, Demyanov et al. 2020). It is also worth noting that innovation has occurred within the field of geostatistics, for example, in the development of non-stationary models which fit better locally to rich spatial datasets, and in the image-based training of the multiple-point statistics approach which replaces two-point (spatial covariance) representations of spatial dependence (Mariethoz and Caers, 2015).

However, one has to be careful not to "throw the baby out with the bathwater": the learning that has been achieved over a nearly 60-year period in the field of geostatistics and, in particular, the principled approaches that have been hard-earned, need to be sewn together with machine-learning and deep-learning approaches that rely less on models and which, thus, may fail to characterise explicitly the true spatial (and temporal) nature of environmental properties. At a minimum, non-adherence to these principles should be declared explicitly so that users are at least aware where machine learning and deep learning approaches fall short.

An example is given by treatment of the spatial support (the subject of two of the papers in this special issue). Geostatistics allows for characterisation of the spatial (temporal) support within the

stochastic model itself, thus, supporting manipulation of the model, once fitted, to explore the effect of the support on prediction and other operations. Machine learning approaches rarely, if ever, do this. And yet, we know that *all* environmental data, with almost zero exceptions, have a spatial (and temporal) support. One may wonder how did the community forget to consider this fundamental property of environmental data when developing machine learning approaches for their fields of study? The importance of this neglect is laid bare when one considers that attempts at spatial data integration (currently a hot topic) *without* consideration of the spatial support are necessarily flawed. This is just one example where geostatistics has a lot to offer to data science. There are many others, some of which are neatly demonstrated by the papers in this special issue (see the paper by López et al., in this issue for an example of integrating geostatistics and machine learning).

This special issue is based on research contributions presented at geoENV2018, the 12th International Conference on Geostatistics for Environmental Applications. GeoENV conferences, which are held biennially, have become established as a leading forum for scientists across a broad range of disciplines to share their experiences on the application of geostatistics to *environmental* problems (with mining geostatistics and petroleum geostatistics covered in other fora).

The topics covered in this issue, including aquifer modelling, remote sensing, renewable energy, soil geochemistry, contaminated land and risk assessment and monitoring fish habitats, demonstrate the diversity and practical utility of geostatistical methods. The approaches explored within this special issue fall into three areas: spatial and temporal modelling, multivariate geostatistics and Bayesian inference.

Spatial and temporal modelling

The suitability of geostatistics to deal with big data in the form of remote sensing data is demonstrated by Qunming Wang, Xiaohua Tong and Peter Atkinson. A geostatistcial filter, based on a downscaling-upscaling approach using area-to-point kriging (ATPK), was applied to enhance image quality by alleviating the point spread function (PSF), which affects all remotely sensed imagery. The PSF is a consequence of measurement and defines the spatial support of imagery to be centre-weighted and larger in spatial extent than a pixel, thus, causing over-smoothing in the raw data themselves. The novel approach presented aims to return the imagery to a perfect “square-wave” PSF (i.e., no over-smoothing) without new measurement. The approach was tested on simulated and real datasets and was found to reduce the PSF effect substantially. The approach is entirely general.

While geostatistics was originally concerned with developing principled approaches to handling spatial data, increasingly the focus has been on space-time modelling **(**Kyriakidis and Journel 1999; Bailey and Krzanowski, 2012). The paper by Pierre Petitgas, Didier Renard, Nicolas Desassis, Martin Huret, Jean-Baptiste Romagnan, Mathieu Doray, Mathieu Woillez and Jacques Rivoirard focuses on optimising space-time models for monitoring ecosystem conservation, and more specifically changes in the spatial variability of fish habitats through time. The research uses geostatistical multivariate min-max autocorrelation factors (MAFs) to determine the spatial components that are the most consistent or coherent through time, and provides a time-series of their amplitudes. When applied to the spawning distributions of sardines in the Bay of Biscay, the use of MAFs enabled groups of typical distributions to be identified, with different occurrence probabilities for different time periods.

Multivariate geostatistics

In geosciences, there are many forms of multivariate datasets. In the paper by Julian Ortiz, Willy Kracht, Giovanni Pamparana and Jannik Haas, geostatistical tools are used within an integrated framework to model the variability of rock properties, climatic parameters and demand management

of a semi-autogenous grinding (SAG) mill energy system (used in mineral processing). The application of geostatisical simulation and stochastic optimization to the integration of renewable energy and mining processing operations demonstrates the flexibility and suitability of the geostatistical toolbox to model multivariate spatial uncertainty for complex challenges such as those related to advancing our progress towards delivering green energy.

Acknowledgement of the nature of the data used and deploying a suitable approach that honours the multivariate and constrained nature of any data are important considerations (Aitchison and Egozcue 2005; Schaeben et al. 2007). Spatial geochemical distributions are used in numerous applications, from environmental baseline determination to development of thresholds to inform health guidelines. However, the compositional nature of geochemical data imposes several limitations related to both closure and the inherently multivariate relative information conveyed by these data, especially when subsequent statistical methods are used (McKinley et al. 2016; Tolosana-Delgado et al. 2019). The paper by Carlos Boente López, Saki Gerassis, Maria Teresa Albuquerque, Javier Taboada and José Luis Rodríguez Gallego, uses geochemical data to investigate the suitability of local and regional soil screening levels (SSLs). The paper highlights the limitations of working with restricted protocols set by regulatory bodies in the development of SSLs from regional databases. The authors deploy a methodology based on a Bayesian network analysis, followed by a stochastic sequential Gaussian simulation approach. The use of a supervised Bayesian machine learning approach with the use of local SSLs was found to outperform the regional SSLs and reduce the uncertainty of the geostatistical predictions. The research demonstrates the opportunity to adopt an integrated machine learning and geostatistical approach.

Bayesian inference

The paper by Luc Steinbuch, Thomas Orton and Dick Brus explores the use of a Bayesian inference framework to account for uncertainty in area-to-point kriging (ATPK). This is a welcome research topic because the bulk of ATPK research has been undertaken using classical (e.g., maximum likelihood) inference frameworks. Using a case study on aggregated crop yields in Burkina Faso the research highlights the importance of model selection for ATPK.

Andrea Zanini, Marco D'Oria, Maria Giovanna Tanda and Allan Woodbury apply a Bayesian inverse approach coupled with Akaike's Bayesian Information Criterion (ABIC) to estimate a complex heterogeneous 2D aquifer system. The results obtained for a confined aquifer case study demonstrate the robustness of the inverse approach.

Summary

In summary, Earth surface and sub-surface properties are "special" in that they are structured in space and time, as explained above. Geostatistics has been developed continuously to create stochastic models that explicitly represent and are, thus, appropriate to these space-time properties. Therefore, while it is acknowledged that data science is bringing great innovation in methods, including in machine learning and deep learning, there is also huge potential, and work to do, to develop methods that are explicitly adapted to the true nature of environmental properties. Many of the solutions (and the problems that drove them) can be found in the expanding literature on geostatistics, including in this special issue.

References

Aitchison, J., J. Egozcue, J. (2005) Compositional Data Analysis: Where Are We and Where Should We Be Heading?. Math Geol 37, 829–850. https://doi.org/10.1007/s11004-005-7383-7

Bailey, T.C., Krzanowski, W.J. (2012) An Overview of Approaches to the Analysis and Modelling of Multivariate Geostatistical Data. Math Geosci 44, 381–393 (2012). https://doi.org/10.1007/s11004-011-9360-7

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

David M (1977) Geostatistical ore reserve estimation. Elsevier, Amsterdam, p 36

Demyanov, V., Gloaguen, E. & Kanevski, M. (2020) A Special Issue on Data Science for Geosciences. Math Geosci 52, 1–3. https://doi.org/10.1007/s11004-019-09846-0

Goovaerts, P. (1997) Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, 512 p.

Journel, A.G. and Huijbregts, C.J. (1978) Mining Geostatistics. Academic Press, New York.

Kyriakidis, P.C., Journel, A.G. (1999) Geostatistical Space–Time Models: A Review. Mathematical Geology 31, 651–684. https://doi.org/10.1023/A:1007528426688

Mariethoz, G. and Caers, G. (2015) Multiple-Point Geostatistics: Stochastic Modeling with Training Images, John Wiley and Sons: London.

McKinley JM, Hron K, Grunsky EC, Reimann C, de Caritat P, Filzmoser P, van den Boogaart KG, Tolosana-Delgado R (2016) The single component geochemical map: fact or fiction? J Geochem Explor 162:16–28

Schaeben H, Pawlowski-Glahn V, Olea RA. (2007) Geostatistical analysis of compositional data. Math Geol 39, 435–437. https://doi.org/10.1007/s11004-007-9105-9

Talebi, H., Mueller, U., Tolosana-Delgado, R. et al. (2019) Geostatistical Simulation of Geochemical Compositions in the Presence of Multiple Geological Units: Application to Mineral Resource Evaluation. Math Geosci 51, 129–153. https://doi.org/10.1007/s11004-018-9763-9

Tolosana-Delgado, R., Mueller, U. & van den Boogaart, K.G. (2019) Geostatistics for Compositional Data: An Overview. Math Geosci 51, 485–526. https://doi.org/10.1007/s11004-018-9769-3

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A. , Hare, J. and Atkinson, P.M. (2019) Joint deep learning for land cover and land use classification, Remote Sensing of Environment, 221; 173-187.