PREFACE

Preface: Geostatistics and hydrogeology

Philippe Renard[1] · **J. Jaime Gómez-Hernández**[2] · **Maria-Theresia Schafmeister**[3] · **Emmanouil A. Varouchakis**[4]

Geostatistics is a branch of statistics that deals with the analysis and modeling of spatial and temporal variability of natural phenomena. It is a discipline that grew at the boundaries of probability theory and earth sciences; it is now applied widely, for example, in environmental sciences, health sciences, or geographical information systems. It is also the foundation of a family of methods (Gaussian processes) that are now employed heavily in machine learning and artificial intelligence.

Since the early days of geostatistics, a close link existed with hydrogeology. Georges Matheron, who was one of the founders of geostatistics (Matheron 1962) developed these tools for the estimation of ore deposit resources. He was also one of the first scientists to propose that hydraulic conductivity in aquifers could be described using spatially correlated random fields. He used this mathematical formalism to prove that the effective hydraulic conductivity of a log-normal isotropic medium in two dimensions is the geometric mean of the local values (Matheron 1967). This was one of the starting points of the field of stochastic hydrogeology, which allowed deciphering the relations between the statistical characteristics

✉ Philippe Renard
philippe.renard@unine.ch

J. Jaime Gómez-Hernández
jaime@dihma.upv.es

Maria-Theresia Schafmeister
schaf@uni-greifswald.de

Emmanouil A. Varouchakis
evarouchakis@tuc.gr

1   Centre for Hydrogeology and Geothermics, University of Neuchâtel, 11 Rue Emile Argand, CH-2000 Neuchâtel, Switzerland

2   Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022 Valencia, Spain

3   Department of Applied Geology/Hydrogeology, University of Greifswald, Friedrich-Ludwig-Jahn-Str. 17a, 17487 Greifswald, Germany

4   School of Mineral Resources Engineering, Technical University of Crete, 73100 Chania, Crete, Greece

of aquifers and their global flow and transport properties (Dagan 1989; Rubin 2003). The connections between the two fields are therefore intimate. Some theoretical developments in geostatistics have been inspired by specific questions originating from hydrogeological applications—for example, the question of the connectivity of high-permeability zones in an aquifer is crucial to accurately predict flow and solute transport (Gómez-Hernández and Wen 1998). How to constrain the interpolation of a parameter field using indirect state variables and the flow or transport equations connecting these quantities is also a core problem that needs to be solved in hydrogeology (de Marsily et al. 1999). All of these questions triggered research in geostatistics and had impacts on other fields.

From an application perspective, there are many situations in which data that are acquired in a groundwater system need to be interpolated to understand or model the system. These data can be state variables such as piezometric levels, quality indicators, forcing terms (recharge), rock types, or petrophysical properties. Geostatistics offers a rigorous framework for analyzing the spatial variability of these data and for carrying out the interpolation while quantifying the resulting uncertainty (Journel 1989; Goovaerts 1997; Chiles and Delfiner 2012). It is therefore not surprising that geostatistics became a standard tool available in many geographical information systems or modeling packages employed by hydrogeologists.

The special issue

This special issue reflects some of the typical and current applications of geostatistics in the field of hydrogeology. It also introduces some recent developments at the intersection of these two fields. In particular, many of the questions that are at the heart of the geostatistical approach are identical to those that are treated by machine learning or artificial intelligence algorithms. The idea of learning some patterns from data sets and making statistical predictions is exactly what geostatisticians have been doing for the last 60 years. Some of the algorithms recently used in the machine learning

community are identical to what is commonly used in geostatistics. However, the rise and diversity of machine learning approaches and the dynamism of this new community has also opened new doors in the field of spatial statistics for hydrogeological applications; therefore, this special issue also explores some of these new methods.

This issue is organized into five main parts, which reflect the typical use of geostatistics in hydrogeology. The three first parts consider the problems of interpolating groundwater levels, groundwater quality, and aquifer parameters, and are key tools for understanding the structure or the behavior of an aquifer. The fourth part is more fundamental; it illustrates how geostatistical methods can be employed for investigating the behavior of heterogeneous aquifers. The final part involves the most complex methods. Their aim is to relate state variables and aquifer parameters to solve the inverse problem while quantifying uncertainty.

Last but not least, with the development of the open science movement, it has become more and more common for scientists to share their codes and their data to foster reproducibility and to challenge colleagues to obtain better results. Geostatistics was at the forefront of this movement with open-source software such as GSLIB (Deutsch and Journel 1992). This special issue is also contributing to the open science movement by providing some original data sets.

Interpolation of groundwater levels

In the first part of this special issue, three approaches are discussed for mapping groundwater levels using geostatistics or machine learning methods. **Chihi and Ben Cheikh Larbi** compare ordinary kriging and kernel ridge regression (KRR) to improve water table prediction in complex geological environments. The KRR approach is tested and evaluated in a coastal faulted aquifer system in southeast Tunisia. Both methods were able to generate plausible piezometric maps. The results of the cross-validation are better for both methods when faults are considered, and KRR is slightly superior to ordinary kriging. **Júnez-Ferreira et al.** emphasize the superiority of spatiotemporal (ST) kriging estimation of groundwater levels as compared to the classical spatial approach, especially when assessing temporal fluctuations. Based on long-term, but irregular, time series in the southern Basin of the Mexico aquifer system, the advantages and limitations of ST kriging are demonstrated. Finally, **Pavlides et al.** compare universal kriging (UK) and stochastic local interactions (SLI) to interpolate groundwater levels around three mine sites in northern Greece. They employ cross-validation to compare the methods and show that both methods perform adequately given the small sample size. UK performs slightly better; this

is interpreted to be the result of using a constant mean in the current implementation of the SLI approach.

Interpolation of groundwater quality

In the second part of this special issue, four papers present methods and case studies involving the interpolation or analysis of groundwater quality. The paper by **Palma et al.** conducts an analysis of three water quality parameters and one water quantity parameter with the ultimate objective of producing a probability map that shows the risk for aquifer deterioration. The method the authors use to produce the probability map is space-time indicator kriging. A space-time linear co-regionalization model (ST-LCM) is applied to describe the direct and cross-covariance of the four variables in space-time. **Schafmeister et al.** present an application of extension variance, together with Voronoi tessellation, to groundwater quality data in Germany. By calculating the extension variance and utilizing geostatistical assumptions, it becomes possible to estimate the probability of exceeding a given concentration threshold. Furthermore, the method is extended to include the delimitation of hydraulically defined groundwater bodies, ensuring that hydraulic boundaries are appropriately considered. **Wang et al.** present a comprehensive investigation of the spatial redox architecture, which is important, for example, to understand nitrate reduction capacity in rural areas. Data from transient electromagnetic resistivity surveys and noncollocated boreholes are integrated by means of geostatistical simulations and a subsequent statistical learning method (multinomial logistic regression) in order to predict the three-dimensional (3D) redox structure in a heterogeneous glacial aquifer in Denmark. Finally, **de Fouquet et al.** treat the problem of mapping a radioactive contaminant plume in Chernobyl, Ukraine, using kriging. The difficulty lies in the nonstationarity of the spatial statistics of the concentrations. To circumvent this common issue, the authors propose to construct numerical covariances that are estimated from a set of numerical simulations of plume migration using a solute transport model. From the ensemble of numerical results, one can derive all the covariances that are required to solve the kriging equations. It is worthwhile to note that this approach can potentially be applied to relate other quantities such as transmissivity and hydraulic data, offering a practical solution to the inverse problem.

Modeling aquifer structure and heterogeneity

One of the key features of aquifers is their internal geological heterogeneity, implying that their petrophysical properties can vary very strongly in space. **Pardo-Igúzquiza et al.** compare different variants of kriging and cokriging for

interpolating the transmissivity field of the Vega de Granada aquifer in Spain. Three types of data are used: actual transmissivity measurements, specific capacity measurements, and hydraulic heads. The authors show how cokriging can be used to combine these different sources of information. The use of an analytical covariance relating groundwater heads and transmissivity overcomes the need to infer a covariance model from a scarce data set and provides a simple and straightforward solution. To complement the lack of direct observations, another possibility is to integrate indirect geophysical data in the interpolation algorithm. Along this line of research, **Kawo et al.** address the necessity of building realistic 3D models of highly heterogeneous aquifers, such as the glacial deposits in Nebraska, USA, in order to design reliable groundwater management zones. Multiple point simulations based on electromagnetic geophysical surveys are used to generate 3D hydrofacies models that realistically reflect the spatial aquifer heterogeneity. **Manzoni et al.** use a very large data set covering most of the Po Plain of Italy to build an artificial neural network (ANN) trained on 450,000 lithology labels collected on more than 50,000 boreholes to predict the lithology of the sediments in three dimensions. The hyperparameters of the ANN are selected based on a k-fold cross-validation procedure. In addition to the exceptional data set treated in this case study, the authors focus a large part of their research on the quantification of the uncertainty with ANN. Finally, instead of generating geostatistical simulations of random fields, a new possibility is to use generative neural networks. The advantage is that the generation of realizations is usually very fast. In this perspective, **Redoloza et al.** show how a progressive growing generative adversarial network (PGGAN) can be employed to generate 2D stochastic binary facies models representing channels embedded in a matrix. One of the difficulties with this family of methods is accurately conditioning the simulations with borehole data. In the paper, the authors investigate, in detail, the relationship between the efficiency of the conditioning and the internal structure of the GAN.

Impact of aquifer heterogeneity on hydrogeological processes

While the papers presented in the previous part report on methods that can be used to describe geological heterogeneity, the two papers that are grouped in this section use those methods to investigate the impact of geological heterogeneity on hydrogeological processes. The first paper, by **Chen et al.**, highlights the importance of mineralogical rock heterogeneity, especially when considering the transport of dissolved radionuclides within granitic host rocks for high-level radioactive repositories. They demonstrate the impact of the spatial correlation structure of reactive mineral facies (RMF)

on scale-dependent sorption coefficients based on data from rock samples of the Beishan site, in northwest China. The second paper, by **Pannone**, provides an in-depth analytical derivation of the dispersive behavior of a solute plume in 3D geological formations displaying power-law-type variograms of log-hydraulic conductivity. This kind of variogram is typical of aquifers having no characteristic scales (correlation length) of heterogeneity but rather having heterogeneity patterns that evolve with the scale of observation. Under these conditions, the author is able to predict how the exponent of the variogram influences the statistics of the center of mass of a plume of solute and its macrodispersion. A comparison with tracer experiments in Cape Cod, USA, seems to confirm the validity of the theory.

Inverse and data assimilation methods to identify aquifer parameters

In the last section, all the papers aim to identify aquifer parameters from the measurements of state variables. The relation between these two quantities involves a set of partial differential equations and therefore it is necessary to solve an inverse problem to identify the parameters. While the inverse problem is as old as quantitative hydrogeology, obtaining reasonable solutions efficiently is still a major challenge and an area of very active research. In the first paper, **van Leer et al.** investigate whether it is feasible to employ an inverse method, not to identify directly the hydraulic conductivity values, but rather to identify the parameters of the covariance function of the hydraulic conductivity of an aquitard using pumping test data. Their aim is therefore not only to identify the specific heterogeneity of their aquitard, but also the type of heterogeneity that would be compatible with the pumping test data. They show that such a procedure can work but that large uncertainties remain. **Pereira et al.** show that geostatistical methods can also be used to improve aquifer characterization by enhancing the results obtained from electrical resistivity tomography (ERT). The method that they propose better characterizes small-scale variability and uncertainty than traditional geophysical inversion methods. **Lauzon and Marcotte** consider the problem of jointly identifying a map of hydrofacies and the spatially varying hydraulic conductivities within the hydrofacies using transient data. They model the hydrofacies' categorical distribution using a pluriGaussian geostatistical method and the hydraulic conductivity using more classical multiGaussian fields. These two distributions are simulated using a spectral turning band method. By optimizing the underlying phase vectors, the method can drastically reduce the dimensionality of the problem and minimize the misfit function; 2D and 3D synthetic examples demonstrate the applicability of the approach. **Todaro et al.** consider a rather similar type of problem (identifying the sediment type

and the hydraulic parameters), but they invert concentration data that were obtained during a tracer test in an experimental sandbox. They compare the direct application of the ensemble smoother with multiple data assimilation (ES-MDA) to the hydraulic conductivity field and its application coupled with a truncated Gaussian model. The parameter updates are applied to pilot point values. They conclude that ES-MDA coupled with a truncated Gaussian model outperforms the standard ES-MDA. The last paper in this special issue considers an even more complex situation. The aim is not only to identify the hydraulic conductivity field but also to locate the initial spatial distribution of a dense nonaqueous phase liquid (DNAPL) contamination. To solve this problem, **Shi et al.** compare the performances of two geostatistical algorithms and one deep-learning inversion algorithm. The numerical experiments are carried out on a synthetic case. The conclusion of this study is that the deep learning approach consistently outperformed the two other approaches.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Chiles J-P, Delfiner P (2012) Geostatistics: modeling spatial uncertainty, 2nd edn. Wiley, Chichester, England

Dagan G (1989) Flow and transport in porous formations. Springer, New York

de Marsily G, Delhomme J-P, Delay F, Buoro A (1999) 40 years of inverse problems in hydrogeology. C R Acad Sci Ser IIA Earth Planet Sci 329:73–87

Deutsch CV, Journel AG (1992) GSLIB geostatistical software library and user's guide. Oxford University Press, New York

Gómez-Hernández JJ, Wen X-H (1998) To be or not to be multi-Gaussian? a reflection on stochastic hydrogeology. Adv Water Resour 21:47–61

Goovaerts P (1997) Geostatistics for natural resource evaluation. Oxford University Press, New York

Journel A (1989) Fundamentals of geostatistics in five lessons. American Geophysical Union, Washington, DC

Matheron G (1962) Traité de géostatistique appliquée [Treatise on applied geostatistics]. Technip, Paris, France

Matheron G (1967) Éléments pour une théorie des milieux poreux [Elements for a theory of porous media]. Masson and Cie, Paris, France

Rubin Y (2003) Applied stochastic hydrogeology. Oxford University Press, New York