Chapter 31
When Should We Use Multiple-Point Geostatistics?

Gregoire Mariethoz

Abstract Multiple-point geostatistics should be used when there is either too little or too much information available for other types of geostatistics.

31.1 Under-Informed Versus Over-Informed Models

For a long time, the classical geostatistical framework required moderate amounts of knowledge. Too little knowledge (few hard data, poorly distributed, absence of auxiliary information), makes it difficult to infer the parameters of a covariance model. In the other extreme, too much knowledge risks revealing characteristics of the underlying field that are too complex to be represented by a handful of covariance model parameters. These two situations can be denoted respectively under-informed and over-informed models. In-between these extremes, we have the moderately informed case where it is convenient to use the covariance-based geostatistical framework, which has been—and still is—a very solid basis for building models that incorporate spatial and temporal variability.

Extreme under-informed and over-informed cases have often presented technical challenges, for which practical workarounds are used. For under-informed cases, standard geostatistical practice consists for example in including interpretative knowledge to guide variogram fitting when too few hard data are available. This is one of the reasons for the common recommendation to fit variograms by hand (e.g. Olea 1999). The question of designing spatial models for over-informed cases (i.e., when large amounts of data are available) is relatively recent, with the development of improved sensors and high-resolution numerical models that triggered the era of "big data".

The concept of multiple-point statistics (MPS) appeared in the early 1990s, initially as a means of overcoming extreme under-informed situations. The idea, at

G. Mariethoz (✉)
Institute of Earth Surface Dynamics (IDYST), University of Lausanne, Lausanne, Switzerland
e-mail: gregoire.mariethoz@unil.ch

© The Author(s) 2018

B. S. Daya Sagar et al. (eds.), *Handbook of Mathematical Geosciences*,
https://doi.org/10.1007/978-3-319-78999-6_31

the time developed by Guardiano and Srivastava (1993) under the impulsion and guidance of A. Journel, was to give the modeler improved tools to include interpretative knowledge in spatial models. The fundamental novelty of the MPS framework was to encapsulate in a training image the interpretative knowledge on the spatial structure of the modeled phenomenon. Since an image is an object most people are familiar with, it allows combining different types of expertise and data, in particular from people who are not familiar with geostatistics.

This approach naturally leads to disregarding hard data as a tiny fraction of the information to include in a model, implying that data alone are not enough. Then, an important part of the modeling work resides in the design of the training image, which can be difficult as natural images are typically not sufficiently repetitive or stationary. Unsurprisingly, the first successful applications of MPS took place in fields where data are typically few, uncertain and expensive, such as reservoir modeling, soil science or mining. In those domains, MPS is often seen as an alternative to object-based methods. Later, it was found that the concept of training image could also be used to incorporate large amounts of information in a model, and therefore address over-informed and data-rich situations, where an increasing number of applications are taking place.

31.2 MPS Versus Covariance-Based Geostatistics

These different aspects have resulted in MPS being seen as in opposition with covariance-based geostatistics. Indeed, from a traditional statistics point of view, MPS is not rigorous in many respects: for instance there is no real model inference, the uncertainty that can be estimated based on a set of MPS realizations is poorly defined, and extreme events cannot be produced beyond those found in the training image. Emery and Lantuéjoul (2014) have shown, based on thorough numerical and theoretical investigations, that MPS only produces random fields when the size of the training image tends to infinity. With a finite training image, MPS algorithms do no longer approximate a random function. Their value then lies in their capability to automatically generate realistic model realizations, but without control of the underlying statistical model. These issues make MPS methodologically close to machine learning and computer graphics. As a result, when using MPS, one often has to make compromises with random function theory and model consistency. In return, it may be possible to explore the data in new ways and obtain, in some cases, models that are more in line with the unobserved physical reality (Journel 1993).

While the hypotheses and tools used are very different, the domains of application of MPS are essentially the same as traditional geostatistics, consisting in the simulation of either conditional or unconditional random fields, mainly for geoscience applications. As such, MPS and covariance-based geostatistics can be seen as competing, and it is not very surprising that in the last decade there have been many cases of fierce debate between the promoters of these two concurrent approaches (Journel and Zhang 2006; Li et al. 2015). My view is that in fact, the

two sets of methods should not be seen as opposed, but as complementary approaches. They are complementary because they are able to solve different types of problems which can be distinguished by the nature and amount of information at hand. Seeing the covariance-based and the algorithm-based approaches as opposed can distract from the higher goal of building on the strengths of each approach. The risk has been stated by Breiman (2001) on the topic of machine learning methods: "*statisticians have ruled themselves out of some of the most interesting and challenging statistical problems that have arisen out of the rapidly increasing ability of computers to store and manipulate data*".

When the available data and knowledge on the studied phenomenon allow building a random function model, using covariance-based geostatistics is usually appropriate. There are numerous examples of successful models designed in this framework for which it would be very difficult to apply MPS (e.g. Diggle et al. 1998; Goovaerts 2005). Conversely, there are applications where the use of training images and MPS algorithms are better able to address some practical questions. In the next sections, I will show two such examples where the available information is either extremely poor or extremely rich. Applying covariance-based geostatistics to these examples would likely yield unsatisfactory results. I emphasize here that for the purpose of demonstration, I am exclusively focusing on examples that are tailored for the application of MPS. Countless examples can be found for which covariance models are perfectly applicable, but it is beyond the scope of this short chapter to show them here.

31.3 Examples for Which MPS Works Well

31.3.1 MPS Can Be Used in Extreme Under-Informed Situations

An example of extreme under-informed model is the common problem of interpolating rainfall data over a given area based on a small number of rain gauges. Rainfall is an inherently intermittent and highly spatially variable process (Benoit and Mariethoz 2017). Moreover, in some cases rain gauge data can be of poor quality, and it is not uncommon to only have binary wet/dry information (as opposed to rainfall accumulation). An example of such poor dataset is shown in Fig. 31.1, with synthetic rain observations consisting of 30 rain gauges. While this case is synthetic, the setting is relatively standard in terms of data density and heterogeneity. It is quite clear that 30 observation points are insufficient to properly infer a spatial model, which is confirmed by the experimental variogram that shows no spatial structure (and wild fluctuations when the number of lags is varied).

In such a setting, the MPS approach starts by stating that the information contained in the hard data is insufficient. At best, the data points can be used for conditioning, but not for inferring any kind of structural model. Instead, one has to

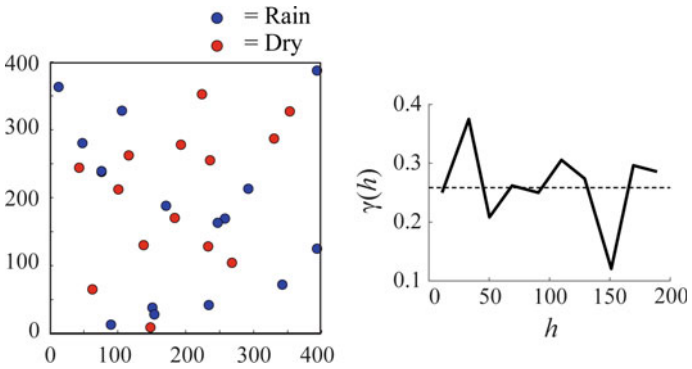

Fig. 31.1 Under-informed setting. Left: synthetic rain gauge network made of 30 points with only wet/dry information. Right: experimental omnidirectional indicator variogram of the probability of rainfall, computed on 10 lags

supplement the insufficient data by resorting to external knowledge of the modeled process. For example, one may know the type of rainfall for that specific day. Based on this knowledge, it is possible to collect radar images of rain events of the same type. Rainfall radar images, either ground-based or satellite-based, are typically collected by national weather agencies and made available to the scientific community. Then, using these representative radar images as training images, MPS can be used to generate rain fields conditioned to the gauge data.

Figure 31.2 shows the results of using two different training images to interpolate the data shown in Fig. 31.1, by considering as training image alternatively a cyclone (left) or a tropical storm (right). It is obvious here that the choice of the training image has a strong influence on the results as it determines the types of patterns found in the simulations, as well as global statistics such as the proportion of wet areas.

This example illustrates the conceptual differences between MPS and covariance-based geostatistics. These differences extend beyond the formalism or the algorithms used. While classical geostatistics infer a model based on data, MPS generates additional data based on external knowledge, in this case through the search for and the selection of an appropriate radar image.

31.3.2 MPS Can Be Used in Extreme Over-Informed Situations

The most common situation in geostatistics is to have a handful of data points, and based on these, to estimate the target variable on a large grid. Increasingly in recent years, the opposite situation occurs with a large number of data used to predict the value at a smaller set of locations. One prime example of such over-informed

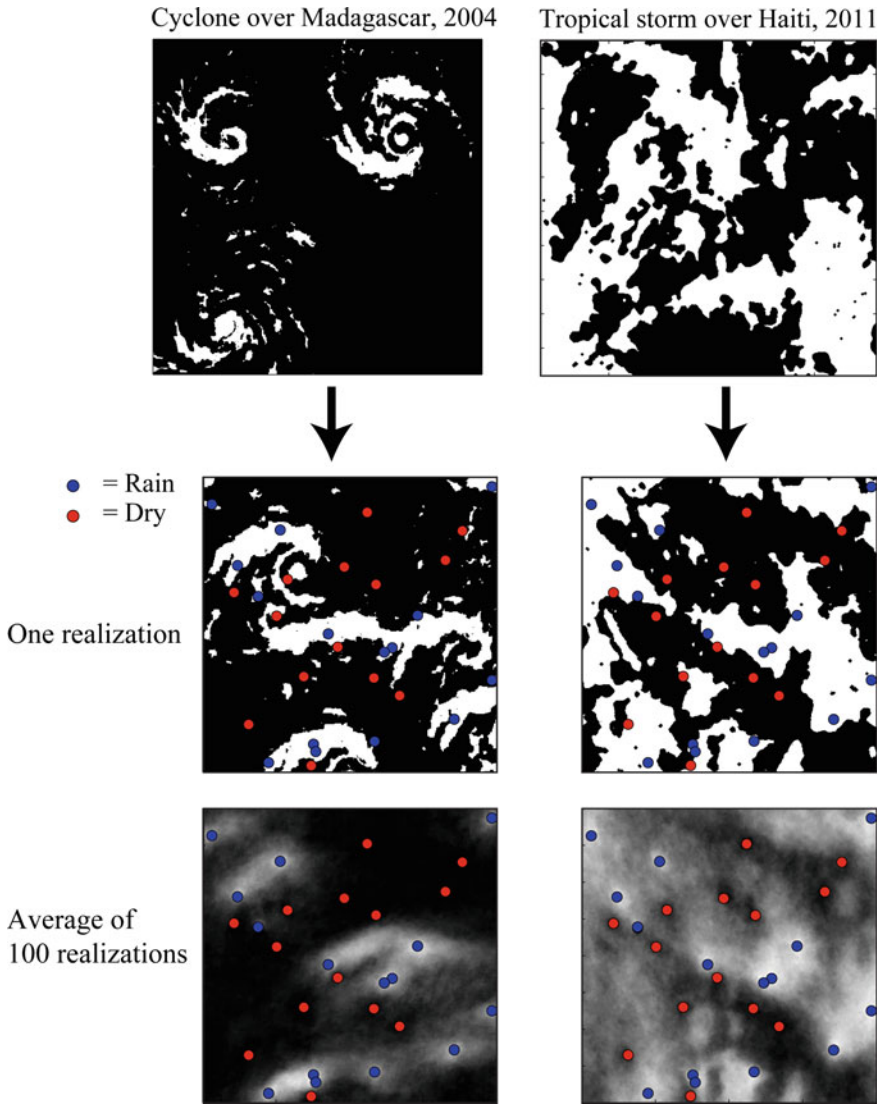

Fig. 31.2 Application of MPS for rain occurrence simulation. Left: simulation of binary rainfall based on a training image of a cyclone. Right: same setting based on a training image of a tropical storm. Size of training images: 572 × 584 pixels. Size of simulation grid: 400 × 400 pixels. The Direct Sampling MPS algorithm was used

problems is applications to satellite imagery, which typically consist in large spatial datasets (typically the entire Earth is covered at high spatial resolution) that also present a temporal aspect since the same location is imaged at regular intervals. Here we look at the Landsat 7 ETM + sensor, which has the characteristic that it

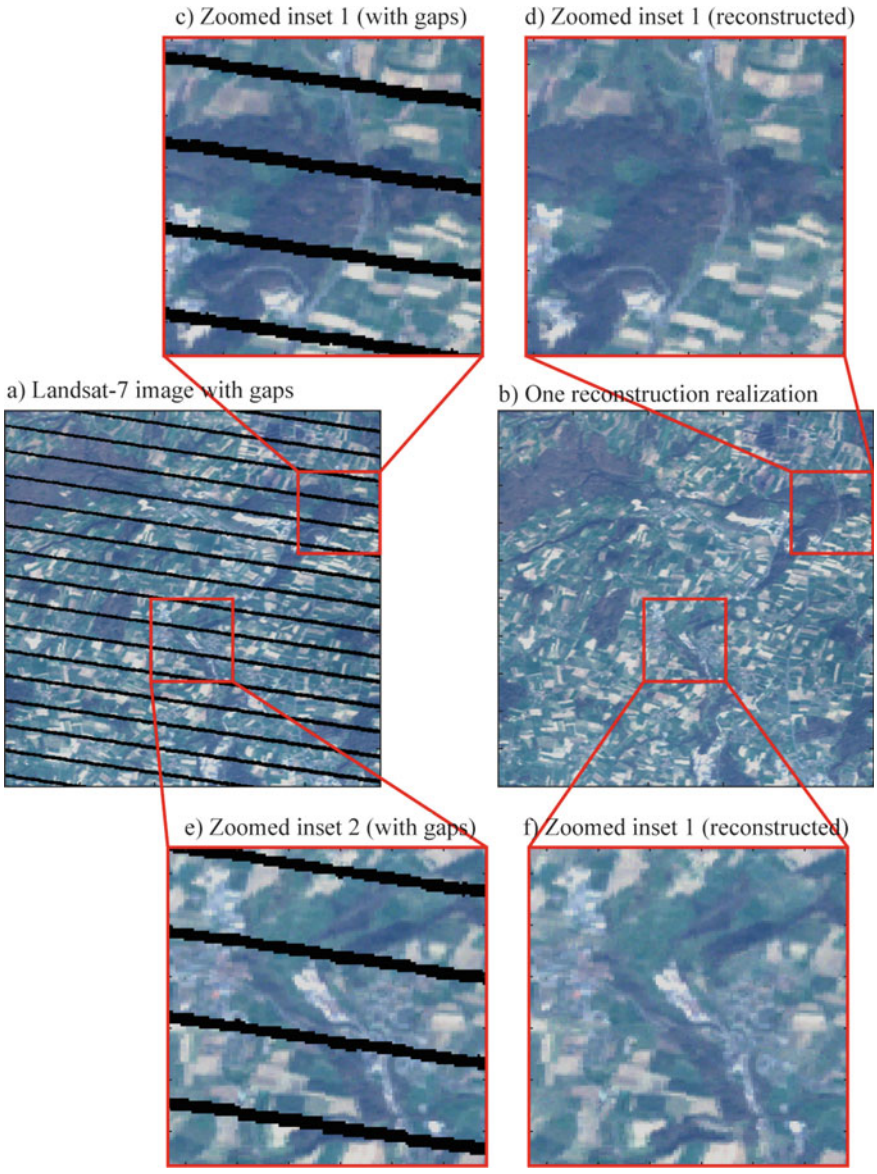

Fig. 31.3 MPS applied to gap-filling of a 5-band Landsat 7 image. Scene acquired on March 22, 2017 in Western Switzerland. Image size: 500 × 500 pixels. The Direct Sampling MPS algorithm was used. Image shown in natural colors

partially failed in 2003, and since then the images it acquires present gaps (as shown on Fig. 31.3a). The goal here is to fill these gaps with simulated values. In such an image, the regions to reconstruct typically represent about 20% of the

domain, the rest consisting of conditioning data. These data contain not only local information, but also very rich structural information such as the type of land surface features (fields, forests, cities), the connectivity of the different objects (roads, water bodies), and their spatial arrangement (see details shown in Fig. 31.3c, e).

The application of covariance-based geostatistics is in this case difficult, not because of challenges related to model inference and identification (as in Fig. 31.1), but because standard simulation techniques, such as Sequential Gaussian Simulation or turning bands, will likely result in artifacts that are clearly visible to the eye. Indeed, the complex land surface information cannot be entirely represented by covariance models which are typically represented by a small number of parameters. Furthermore, although interpolation artifacts are sometimes obvious to the eye, they are typically undetectable by standard statistical metrics because these metrics are based on covariance (or two-point statistics) and cannot identify complex patterns such as connectivity, for which the human eye is very well suited. It can of course be argued that there are applications where these complex properties do not matter; but if they do, the covariance-based framework is inappropriate (Zinn and Harvey 2003).

In contrast, applying MPS to this gap-filling problem is straightforward. The MPS approach used here for the simulation of gaps is the one presented by Yin et al. (2017a, b). Each color channel is co-simulated and no auxiliary variables are used. Contrarily to the data-poor case, there is no need here to infer, construct or hypothesize a training image. The training image is given by the 80% of the domain that is known. While the training image size is far from infinity, it is a little closer to the ideal situation outlined by Emery and Lantuéjoul (2014). The gap-filling results (Fig. 31.3b, d, f) present very few visual artifacts. In certain places, it is possible to see that some reconstructed elongated features are discontinuous (e.g. the road near the center of Fig. 31.3d). However in most cases it is difficult to distinguish the reconstructed and the original areas (e.g. in Fig. 31.3f).

31.4 Conclusion

Often the debate around MPS and covariance-based approaches has been centered on the dichotomy between multiGaussianity or non-multiGaussianity of the variable to simulate (Gómez-Hernández and Wen 1998). The choice of a simulation approach or algorithm should certainly be driven by the nature of the variable of interest: is it non-multiGaussian? is it non-stationary? is it channelized? do these characteristics matter for a given problem? I argue here that the question of the amount of information at hand is also a critical factor to consider when choosing which simulation framework to use, and this question has often been overlooked. It may make sense to also base this choice on the quantity of information available: do I have a conceptual model? do I have enough hard or soft data to infer a covariance? do I have so much data that I am able to detect non-multiGaussian behavior?

To summarize, one can say that different tools are available, and those should be chosen according to the problem to be solved. While no example with moderate amount of information has been shown in this chapter, it is understood that it is generally the realm of covariance-based geostatistics. Under-informed situations are always going to be difficult because there are important modelling choices to make. For over-informed cases, relatively few assumptions are needed and, with some precautions, it can be possible to rely on algorithms such as MPS.

Better defining the role of MPS in the galaxy of existing spatial modeling tools can potentially help narrowing areas where future MPS research should focus. So far, there has been a strong emphasis on the development of simulation algorithms. The different algorithms available can reproduce spatial features with various degrees of faithfulness, they may need different computing resources or may offer specific options. While developments in MPS are still needed (in particular regarding training image selection and manipulation, as well as parametrization), the simulation algorithms are becoming quite mature. Moving beyond the dichotomy between covariance-based geostatistics and MPS can enable the development of new hybrid approaches. For example, using distance-based (also known as convolution-based) MPS algorithms can be seen as bootstrapping the training image. However, the link with bootstrapping theory (e.g. Davison and Hinkley 1997) has not yet been fully explored. Similarly, the MPS framework is currently unable to simulate extreme values. Combining MPS with more standard statistical approaches may open new fields of applications, in particular in domains such as climate science, hydrology or earth surface observation where increasingly rich space-time datasets are now available.

References

Benoit L, Mariethoz G (2017) Generating synthetic rainfall with geostatistical simulations. Wiley Interdiscip Rev Water 4(2):e1199-n/a

Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16(3):199–231

Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge University Press

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. J R Stat Soc Ser C Appl Stat 47(3):299–325

Emery X, Lantuéjoul C (2014) Can a training image be a substitute for a random field model? Math Geosci 46(2):133–147

Gómez-Hernández JJ, Wen XH (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. Adv Water Resour 21(1):47–61

Goovaerts P (2005) Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using poisson kriging. Int J Health Geogr 4(31):1–33

Guardiano F, Srivastava M (1993) In: Soares A (ed) Geostatistics-Troia. Kluwier Academic, Dordrecht, pp 133–144

Journel A, Zhang T (2006) The necessity of a multiple-point prior model. Math Geol 38(5): 591–610

Journel AG (1993) Geostatistics: roadblocks and challenges. Geostatistics Troia '92, vol 1, pp 213–224

Li L, Romary T, Caers J (2015) Universal kriging with training images. Spat Stat 14:240–268
Olea RA (1999) Geostatistics for engineers and earth scientists. Springer, New York
Yin G, Mariethoz G, McCabe M (2017a) Gap-filling of landsat 7 imagery using the direct sampling method. Remote Sens 9(1):12
Yin G, Mariethoz G, Sun Y, MCabe MF (2017b) A comparison of gap-filling approaches for landsat-7 satellite data. Int J Remote Sens. https://doi.org/10.1080/01431161.01432017.01363432
Zinn B, Harvey C (2003) When good statistical models of aquifer heterogeneity go bad: a comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields. Water Resour Res 39(3):WR001146

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.