Multiple-point geostatistics: Stochastic imaging with training images by G. Mariethoz and J. Caers

J. Jaime Gómez-Hernández
Research Institute of Water and Environmental Engineering
Universitat Politècnica de València, Spain
Tel.: +34 96 387 9614, email: jgomez@upv.es

I was especially pleased when I was asked to review this book, because I was a graduate student at Stanford when multiple-point geostatistics (MPS) was born back in the late 80's of the past century. Indicator geostatistics had been a major breakthrough, but André Journel felt the need to go beyond the bivariate characterization offered by indicator variograms. He came up with the idea of computing local conditional probability distributions for a binary variable accounting for more than two-point variograms by scanning for high-order moments through a training image. Although André is never credited when referring to the history of MPS, he was truly its father. The initial idea became the subject of the master thesis of Felipe Guardiano, with Mohan Srivastava in charge of all the implementation subtleties. That was 25 years ago. Much action has taken place since then, multiple-point geostatistics has matured into a discipline of its own, and today we welcome a comprehensive book on the subject that has been put together by Gregoire Mariethoz and Jef Caers.

The reader is confronted with a 378-page book and an accompanying website with complementary material. A book that is oriented to the practical aspects of MPS, not to create any new theory, a book about practice and the solution of real problems. Beware theoreticians, welcome engineers! And, although the generality of the MPS framework is recognized, the book centers in applications to the physical sciences, not to other areas in which spatial statistics are also used, and for which MPS could be applicable.

The book is divided in three major parts: concepts, methods and applications. In the first part a case study, unrelated to the physical sciences that will underlie the rest of the book, is used to motivate the reader on why the need for MPS, the fundamental concepts are then introduced, and the case study is revisited to show, indeed, the need for MPS when dealing with complex spatial patterns. In the second part a thorough review of all methods, algorithms and tools needed to perform MPS in a variety of contexts are described. And in the third part, three real applications are presented, in the fields of petroleum engineering, mining engineering and climate modeling.

The book is written with a very direct prose, trying to explain the practical implications of the different decisions that a modeler has to take in his quest for producing a representation of a partially known reality, whether you call it estimation or simulation. In this respect, the book is truffled with tips and recommendations, as well as lengthy explanations why many modeling decisions are taken. The book cannot be read on its own, the reader must have previous knowledge about geostatistics, uncertainty analysis and stochastic simulation to fully understand the benefits of MPS. The authors make numerous reference of the standard geostatistical approach based on two-point statistics; therefore, previous knowledge on classic geostatistics is a must.

Part I, on concepts, sets the stage.

Chapter I.1, with practicality in mind, poses a problem (somehow unrelated to the geosciences, not statistical at all), in which the goal is to estimate, from limited information, the topography of a certain area through which hikers will cross to provide them with sufficient (but not in excess) food supplies. This problem will be revisited at the end of this part to support the need to use MPS to model complex spatial patterns.

Chapter I.2 introduces standard estimation based on random function theory, that is, kriging. This chapter is representative of the type of discussions that are included throughout in the book. It begins with a very lengthy and appropriate discussion about the assumptions of stationarity and ergodicity, the fact that there is a unique reality, and that to construct a model many assumptions must be made that cannot be corroborated with data (the authors make clear that stationarity and ergodicity are assumptions, never hypothesis). The chapter goes on presenting the concepts of stationarity, unbiasedness, loss functions, local stationarity, drifts, trends; it develops the equations of kriging, with some discussion about how to infer the covariances and the variograms from data, and the dangers of estimating these from the residuals obtained after a least-square fitting of a trend. The chapter ends with a warning that no statistical model is good for universal kriging.

Chapter I.3 builds on the conclusion from the previous chapter to provide a framework to perform universal kriging without resorting to a random function model, much as Journel did with his "deterministic geostatistics". The concept of training image (TI), although mentioned several times before, is presented for the first time in a formal way: a TI is an image, which provides an analog to the field of study, and which is "deemed representative" of the spatial variability within the study area. Now, the big model decision is not stationarity or ergodicity, but rather to decide whether a TI is deemed representative.

One of the problems when presenting a new framework is nomenclature. This chapter introduces a large number of new elements. The authors have made a tremendous effort to be consistent throughout the entire book, and they have succeeded; however, they have introduced so many new terms and concepts that the reader may become overwhelmed with all the new symbols, acronyms and abbreviations. This wealth of new terminology may preclude some readers to use the book to address a specific problem that can be solved with the tools presented in a specific chapter, because they may stumble with symbols and acronyms for which no definition is readily available. The authors are urged to include, in the next printing of the book, a nomenclature section with all the symbols used in the book and their meaning, and possibly, the page in which they are first introduced.

The authors continue presenting a version of ordinary kriging with TI, in which variogram estimation/evaluation is replaced by the calculation of sums of products from the TI. Then, they present the simple kriging version, and move on to analyze the impact of the choice of search neighborhood, the size of the training image, the nature of the training image and other aspects. An interesting result is that since the products used are two-point products, two TIs with similar histograms and variograms but with very different textures will result in very similar estimates. Then, the authors move into the introduction of non-stationary models, stating that, for the model to be non-stationary, the TIs must be locally representative; one way to achieve this is by the introduction of auxiliary variables. Some examples are given, but unlike for the stationary case, the description is not as exhaustive and the reader is referred to later chapters to fully understand how to work with non-stationary MPS models.

It is important to note that in universal kriging with TI, conditioning is separated from model specification (that is, data are not used to build the model, in this case, the training image) and therefore there could be some inconsistency between data and training image if lots of data are present. An issue that will be discussed in depth later in the book.

Chapter 1.4 deals with stochastic simulation based on random function theory. It starts by explaining in which context one would need stochastic realizations instead of a single estimation map, and then carry on to present several stochastic simulation techniques: first, sequential Gaussian simulation, then direct sequential simulation, and finally pluriGaussian simulation.

Chapter 1.5 presents stochastic simulation without random function theory. It starts by presenting the direct sampling approach in the context of producing a paragraph that may look like it is written in French but that, in fact, is a random realization of words and punctuation signs that resemble true French. The new paragraph simply replicates the patterns observed in a text by Flaubert. Then, it shows how direct sampling would perform in the Walker Lake data set. These two demonstrations show how stochastic simulation can be performed using a TI instead of a formal random function model. The algorithm itself will be explained in detail later in the book.

Next, the extended normal equation is formulated in detail. This equation is the foundation of all MPS algorithms. The extended normal equation, at the end, reduces to counting how many patterns are found in the TI close to a given data event (a specific spatial pattern of point values). From an implementation point of view, there are two ways to address this counting, one by scanning the image each time that a count is needed, and another by precomputing all potential counts a priori. The first approach needs CPU-time; the second one needs random access memory (RAM).

There is an interesting, albeit anecdotic, discussion of the bias introduced in any sequential simulation by the fact that conditioning data are always visited first in the simulation path, before the chapter ends with an excursion into the computer graphics literature to review some texture generating algorithms that can also be applied in MPS: patchwork and image quilting.

Chapter I.6 closes the first part of the book by returning to the case study posed in the first chapter. Conditional realizations of the study area are generated using three different approaches: multiGaussian simulation, direct sampling and image quilting. The authors conclude that, in practice, no model can be validated since there is never a "true" reference with which to compare the results, even though, in the case at hand, the multiGaussian model is poorer than the other two models. They also argue that the underlying multivariate random function model is not primordial, but rather the ensemble of realizations, and therefore those realizations should have the characteristics we wish in terms of continuity, texture, etc. And finally, they remind that when the realizations have to be fed to a non-linear transfer function, the results will most likely depend on the higher-order moments of those realizations, and, consequently, the modeler should use those methods that best capture those moments.

Part II, on methods, provides the tools.

Chapter II.1 starts with a clear statement: it is not about theoretical developments, but about how to implement ways to borrow information from training images; because, at the end, the user needs a computer code, which is made of algorithms, and these algorithms are built based on decisions which may give rise to different interpretations of the same concept.

In this introduction the concept of TI is reinforced, and some of its potential pitfalls already pointed out, such as the possible incoherence between image and conditioning data, or the need that TIs must constrain the images only up to a certain point (otherwise the realizations will be verbatim replicas of the TI or of portions of it).

Chapter II.2 is the longest chapter in the book; it lies out all the elementary blocks necessary to build an MPS algorithm. The chapter starts with a discussion on how data, for one or several variables, can be provided to the modeler, differentiating between scattered data and gridded data. Then, it continues discussing the concept of neighborhood, a key concept in MPS since it does not only fixes the spatial extent within which data will be searched, but also determines the order of the spatial statistical to be considered. (Contrary to traditional geostatistics, there is no equivalence to a unique or global neighborhood in MPS.) Next concept discussed is that of a "data event", which is the combination of the locations of a neighborhood and the values at those locations (a concept that is valid for one or several variables). There is an important remark at this point in the book: concepts in MPS are easy to understand and intuitive, but they could be very difficult to implement, especially for the multivariate cases.

The application of the extended normal equation requires finding and counting specific data events from a TI. The book now discusses the two possible approaches to extract this information from the TI: the raw approach, in which the TI is kept in memory and each time a data event needs to be evaluated, the TI is searched; and the tree-storage approach, in which the TI is thoroughly scanned before the beginning of the simulation and all potential data events are identified and counted, storing, at the end, a search tree with all this information. Raw storage relies on CPU power, whereas tree search relies on RAM availability.

The notion of convolution is introduced to formalize the computation of the local conditional probability distribution given a data event in the case of the raw storage approach, with a natural extension to continuous variables. Next, the tree storage is described in detail; tree storage is especially suited for categorical variables, although the authors also point out to some extension, based on the use of prototypes, for its application to continuous variables. List storage is presented as a less RAM demanding alternative.

This part of the chapter ends with a discussion on how to retrieve/store information from a TI when simulating by patches. Very detailed information is given on how to proceed in this case: the need for clustering similar data events, how to compute similarities using filters, how to assign a score to each pattern, how to classify the patterns using multidimensional scaling or how to extract the cluster from this scaling. At the end, a few pattern clusters are identified, each of which is represented by a prototype, and which can be used to build cluster-histogram of patterns.

An alternative approach to store patterns would be using a parametric approach, and the next part of this chapter goes in detail in how high-order statistics can be computed from cumulants, derived from a TI, and how these cumulants can be used to computed any local conditional probability needed.

The next section of this chapter enters into the discussion of distances both for categorical and continuous variables, a key element to decide how similar two data events are. The Manhattan, Euclidean, and Hausdorff distances are presented. The possibility of comparing data events after translating or normalizing their values, or up to some transformation of their node coordinates is also discussed, together with the discussion of other possible transformations for specific data types. Finally, several distances for the comparison of two distributions are mentioned.

Next, the issue of how the path visiting the nodes of a realization in the context of sequential simulation should be constructed is analyzed, with two main alternatives, to use a random path or a unilateral path. Each one has its advantages and disadvantages, which are discussed, together with variants. An important point brought up next is the need to use multiple nested grids in order to capture the heterogeneity patterns at all scales.

Conditioning to data is next. There is a distinction between hard data and other types. Hard data can be assigned to the grid nodes, although there may be an issue of support. Global proportions can be enforced using a servo-system approach. When indirect data can be translated onto a conditional probability, this probability can be aggregated to the conditional probability given the hard data using different models, such as odd ratios, or the tau model.

Chapter II.3 is probably the most interesting chapter of the book since it describes, in a very clear way, the different algorithms currently available for MPS simulation. There are many algorithms but they have a lot of similarities. All algorithms are presented in a colored box with three components: the inputs needed, a pseudo-code of the algorithm, and the output produced. The chapter starts with the presentation of what the authors call the archetypal MPS algorithm, and a warning on the need to perform some sensitivity analysis of the algorithmic parameters to evaluate how they influence the quantification of the final uncertainty. (Tools for this latter task will be presented later in the book.)

The algorithms are divided into pixel based and pattern based. The pixel-based ones are, in chronological order, ENESIM, SNESIM, Direct sampling, and Simulated Annealing. The pattern-based ones, also in chronological order, are SIMPAT, FILTERSIM, Patchwork simulation, CCSIM, and IQ. For all algorithms, some comments on their weaknesses and strengths are given, and then, all algorithms are compiled in a table where, at a glance, the reader can know the ability of each algorithm to deal with categorical and continuous variables, to handle multivariate simulations, to use hard data, soft probabilities or non-stationarity, plus an estimate of its CPU performance.

This chapter ends with some advice on how to post process the realizations, a task needed when there are inconsistencies between the values in a data event and the TI, inconsistencies that may lead to the simulation of spurious values at some locations.

Chapter II.4 falls a little bit outside the main theme of the book. It presents the theory behind Markov chain models, and Markov mesh models, and discusses how they could be applied in MPS.

Chapter II.5 discusses at depth how to use non-stationary models in MPS. It starts with a discussion on what non-stationarity means to follow with a description of different approaches to inject non-stationarity into the realizations, and how to control it.

First, the authors focus on the use of stationary TIs to generate non-stationary realizations, one way is by using a zonation of the realization, with each zone to be simulated with a different TI; another way is by using data event transformations such as rotations and affinities, specifying locally in the realization the affinity factor or the rotation angle to apply; another alternative is by using soft probabilities (derived, for instance, from a seismic map) which are later aggregated with the probabilities obtained after the solution of the extended normal equations.

Next, the authors continue describing how to use non-stationary images to generate non-stationary realizations. Although it is possible to use zonation here, too, the best alternative is to incorporate control maps, which are continuous variables defined both in the TI and in the realization, which are, essentially a continuous version of zonation.

Chapter II.6 deals with multivariate modeling with training images. Conceptually, the problem to solve is the same as for a single variable, but now we need as many training images as there are variables. The authors mention that only raw storage approaches have been used for relatively large problems and continue with a demonstration in a synthetic case. The authors also describe an alternative approach by interpreting multivariate modeling as a filtering problem, whereby each variable has a TI, plus a filtered version of it (for instance, obtained by geophysics), and the filtered version of the realization is also available; the simulation problem is now cast as finding the realizations that are consistent with the filtered information.

Chapter II.7 deals with the important problem of how to construct a TI. Choosing a TI is the most critical modeling decision in MPS, this is why the authors devote a whole chapter to study alternatives on how to construct them. The methods discussed include object-based methods such as Boolean models; process-based methods, in which the TI is built by modeling the physical processes underlying the phenomena to simulate; process-imitating methods, in which a numerical algorithm is built to imitate the structures observed in nature, much in the same way as the process-based method but with no physical equations solved. Many of the TIs generated in this way are difficult to use because they are clearly non-stationary.

A critical problem discussed next is how to build 3D TIs from 2D ones. The book discusses how to use probability aggregation to combine several 2D orthogonal TIs, or how to assemble 2D orthogonal data events to obtain 3D events. In some fields, the data density is so high that the data set can be used directly as a TI.

The most challenging problem is the construction of multivariate TIs, although, the authors point out that, in most cases there is a primary variable, and the TIs for the other variables can be derived from the primary-variable TI applying some physical model.

The chapter ends with a reference to the attempts of creating some TI databases, such as FAKTS and CARBDB.

Chapter II.8 describes how to validate the data versus the TI and the type of quality controls that should be performed. As indicated in the book, when a model is derived from data, it is impossible that there is any inconsistency between data and model; however, in MPS, the TI is not built from the data, and, therefore, there could be inconsistencies between TI and data that are not obvious, either because of the complexity of the data or because of the density of the hard data.

The most interesting part in relation to TI validation is the technique described to assign probabilities to training images conditioned to the hard data. These probabilities can be later used to rule out improbable TIs and to distribute the number of realizations generated from each postulated TI proportionally to their probabilities.

The chapter goes on with quality control; the first check is to compare histograms, variograms, connectivity functions or other statistics between realizations and TI. The issue of verbatim copy of big patches of the TI onto the realizations is analyzed through the use of coherence maps, and examples are given.

The authors then argument that the realizations should be checked by measuring their spatial uncertainty. The aim is having small within-realization variability (i.e., realizations stay close to the TI) but large between-realization variability (i.e., realizations span the space of uncertainty). These checks can be done by comparing summary statistics calculated on the realizations and on the TI.

A final check suggested by the authors is that of consistency for conditioning, in the sense that conditioning data should exert some effect beyond the point where they are. An expensive method to do it is proposed.

Chapter II.9 introduces the problem of inverse modeling with training images. It is the second longest chapter in the book; the subject is of sufficient entity to deserve a book by itself. The chapter starts acknowledging that, up to here, all conditioning was direct and without iterations, but there are certain types of data which require formulating and solving an inverse model. Although most of the material is general, the authors warn that they will focus in the problem of inverse modeling in the subsurface and the use of geophysical and dynamic data.

The chapter starts with an introduction about inverse modeling to conclude that it can be formulated, in the standard Bayesian framework, as finding the posterior distribution of model parameters given the data, as proportional to the product of a prior distribution of the model parameters times the likelihood of the model.

The authors then present a prior distribution, which is not parametric but rather algorithmic and is given implicitly by the collection of realizations generated from a given TI by a given algorithm. The main problem is that there is a big chance that the stated prior is inconsistent with the data. The authors insist on the importance of building a prior that is as informative as possible and not based on data; this will likely yield to the use of several TIs, that will have to be later compounded to build a prior distribution as wide as possible, while being consistent with data.

The chapter continues with a description, at times too terse, of several algorithms to sample the posterior distribution, and therefore, generate realizations that are conditioned to this type of data that are related to the model parameters through a complex non-linear forward model.

The methods described are rejection sampling, spatial resampling, Metropolis sampling, sequential Gibbs sampling, and pattern frequency matching. All these methods, although theoretically sound, are very CPU intensive, and, in practice, they are replaced for stochastic search methods, whose aim is to generate realizations that are consistent with the prior model and that are also conditional to the data, in the sense that the solution of the forward model matches them. The authors issue a word of caution: the stochastic search methods may generate realizations, each of which is acceptable on its own, but the whole ensemble of realizations may not sample exhaustively enough the space of uncertainty of the posterior.

The stochastic search methods described are the probability perturbation method, the gradual deformation and the neighborhood algorithm. Sometimes it is difficult to follow all the details of the algorithms from the book itself, but enough references are given for the interested reader.

An alternative way to perform a stochastic search is by finding a way to parameterize the MPS realizations and then generate realizations of the parameters. The authors first discuss principal component analysis (PCA) decomposition, and, then kernel PCA.

Chapter II.10 ends part II with a discussion on how to accelerate the generation of realizations by parallelization of the algorithms. After a general discussion of the types of parallel architectures and the types of issues to be taken into account, they present three levels of parallelization: at the realization level, at the patch level, and at the node level; each level with an increased degree of sophistication. The chapter ends with some references to the potential of using the graphic processing unit (GPU) for parallelization purposes.

Part III, on applications, is illustrative.

The book ends with three chapters presenting practical applications of MPS to the fields of petroleum engineering, mining and climate modeling. As important as the examples themselves are, I found much more interesting the workflows presented, especially for the petroleum engineering case.

Chapter III.1 presents the problem of building multiple MPS realizations of a reservoir in which the construction of a new offshore platform is being considered. The objective of the study is well established, and the amounts of data available are thoroughly described. Three geological scenarios are proposed by the geologists, which will yield three TIs generated by Boolean modeling. The issues of coordinate transformation, and the missing scale between data points and cell volumes are commented. The authors go in great detail to describe how they calibrate the target proportions to the seismic information, how the conditional simulations are performed in two steps, and the quality control checks done. The authors also discuss the type of sensitivity analysis carried out to reach some conclusions such as that the choice of TI has an important impact in the final realizations, but the ratio of vertical to horizontal permeability does not. Then, they explain how the probabilities associated to each TI are obtained, to finalize with an application of the probability perturbation method by regions to generate history-matched realizations. It is apparent that those history-matched realizations should be the ones used to appraise whether to construct or not the new offshore platform, but the reader is left out without that information.

Chapter III.2 presents an example application of MPS in mining. This chapter is co-authored by Pérez, Ortiz and Boucher. The authors explain that MPS could help in rapid modeling and updating of complex geological features, and they focus on MPS methods that address the problem of improving an existing deterministic model. The challenge in mining is to deal with plenty of data and to use MPS at the mine scale with models of many blocks. After presenting several models with different degree of conditioning data and validation data, the authors conclude that MPS can be applied in mining, but in my opinion, the real purpose of the entire exercise was obscure and difficult to understand.

Chapter III.3 presents an application in climate modeling. It is an application of downscaling of a global circulation model (GCM) to extract climatic variables at a smaller scale. The exercise has several interesting components: it is multivariate, it requires non-stationary modeling, and it adds the time component. The authors explain how they combine exhaustive data from the past at both the 100 km by 100 km scale, and the 50 km by 50 km scale of three climatic variables, to predict those same variables at 50 km by 50 km scale into the future using the past as training image, plus the predictions, at 100 km by 100 km, from the GCM.

Each chapter ends with its own bibliography. And the book ends with an index of terms.

In addition to the book, the website http://www.trainingimages.org contains additional resources including files containing many of the training images, which can be freely used for research purposes. Although in most cases it is quite obvious which is the format of the files, the authors may wish to consider adding a small explanation about it, as well as how the split zip files should be concatenated to get a working zip file from which the training image file could be extracted. The web page also contains links to research codes and an updated list of bibliographic references.

In summary, this was a long review for a great book tightly packed with plenty of information on multiple-point geostatistics. This book is a must for any researcher or practitioner interested in putting into practice MPS. The deceptive simplicity of MPS is well covered in the book, with numerous indications of potential mistakes in which a user looking for a black-box computer code may incur. All corners of MPS are covered, with extensive descriptions of methods, tools and algorithms. It is without any doubt the reference book on multiple-point geostatistics today, and I presume that it will remain as such for a long time in the future.

Acknowledgements. This review was completed during a sabbatical stay of the author with the Kansas Geological Survey, Kansas University, Lawrence, KS, USA, which was funded by the Spanish Ministry of Education, Culture and Sports through grant PRX14/00501.

488