Geostatistics - A Brief Introduction

Julián M. Ortiz - https://julianmortiz.com/

July 13, 2024

Summary

Geostatistics provides tools to model variables located in space (usually three dimensional) and takes advantage of their spatial structure (continuity) to improve the prediction at locations that have not been sampled and to characterize their spatial texture as a way to assess the uncertainty linked to the limited knowledge provided by the samples.

It is founded in statistical theory and shares many concepts and methods with statistical inference, pattern recognition and other related disciplines.

In this set of notes, we review the main concepts and try to provide both intuitive explanations to the different concepts and detailed implementation parameters and examples, to understand the mechanics to operate these techniques.

We will cover concepts related to probabilistic theory, statistical inference, spatial analysis, estimation and simulation. Further to these, we will explain some of the issues

linked to constraining the models with geological knowledge, extending these theories to the case of multiple variables, expanding the notions of spatial continuity to pattern statistics, link with classical statistical methods and with machine learning and deep learning techniques.

1 Introduction

The use of geostatistics is better understood if seen as a workflow where different methods are applied sequentially to achieve a specific goal. These goals can be linked to **estimating** or to **assessing uncertainty** associated to one or more variables at an unsampled location or over a volume. In the context of natural resources, we often want to understand how a reservoir, ore deposit, or aquifer behaves subject to a specific set of actions applied to it, which we call a transfer function.

For instance, in an ore deposit, we want to extract the rock, decide what goes to the processing plant and how to process it, to maximize the revenues. The optimum extraction sequence will depend on the grades and other mineralogical properties of the rock. Thus, to define the best way of mining the deposit, we need to understand how heterogeneous it is, and what is the spatial distribution of the different geological units or domains and the concentrations of elements and minerals, as well as other important properties, within these domains. The final production of a specific metal is the result of a sequence of decisions made from the early exploration, through the rock extraction, to the processing and metal recovery **Figure 1**. A proper charac-

terization of the in situ resources and then tracking of materials through the different processes is key to understand and predict the performance of every operation. And this is needed if we want to optimize the entire mining value chain.

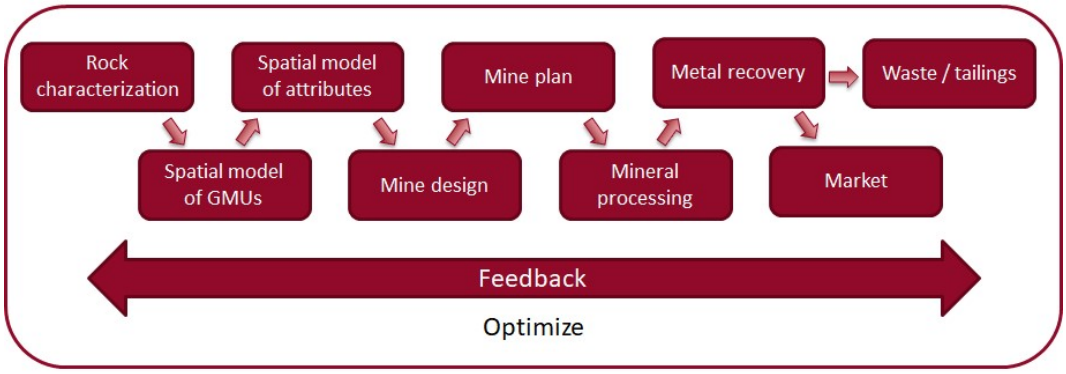

Figure 1: The steps in the mining value chain. Ideally, each stage should provide feedback to all other stages and the entire sequence could be optimized.

In the case of an oil reservoir, a well may be drilled to inject water to move the oil and extract it at a nearby producing well. The production will depend on the heterogeneity of the reservoir, which is characterized by the petrophysical properties of the different rock types found, and their spatial distribution. The notions covered in these notes apply to any variable spatially distributed, where the heterogeneity impacts a response variable. In the following pages, we will focus on mining examples, but the techniques can easily be imported to other applications in geosciences.

The typical sequence of a geostatistical study is:

- Preparation of the database
- Exploratory data analysis
- Definition of domains
- Spatial continuity analysis
- Definition of neighborhood for sample search
- Estimation or simulation
- Post-process

In all these steps, the understanding of the geological setting will constrain the modeling decisions. This means that **geostatistics cannot be applied as a black box**, but requires input of geological and mining knowledge, interpretation of the results of each step, and necessary adjustments.

We will briefly review each one of these steps and present a summary of their relationship with traditional and advanced statistical techniques.

2 Preparation of the database

This seems like a silly aspect, but in most cases it is an important step that takes some time, but that will save us some problems down the path if we take care of all the possible problems early on.

Since the database will be used to perform numerical modeling, it has to be clean and complete. Formats need to be consistent and we need to make sure that the values make sense. This will be checked during the exploratory data analysis stage, but we can save some time if we know

what is expected for the subsequent analyses. For most geostatistical analysis, what we need is a data table where columns are variables and rows are observations (or samples). This can be seen as a point dataset, where each observation is linked to a coordinate in space.

Typical questions to be aware of are:

- Does the database comes as a set of relational tables? Do you need to desurvey the data to get coordinates of each sample point?
- Is the database a 2D or a 3D set of points? Are all the sample points within the expected volume? Are there any missing coordinates?
- Is there information in a regular grid? Are the samples scattered? Is there a drilling campaign with regular spacing?
- Is there information for all the variables? For a single variable, are all measurements done with the same analytical procedure? Are they comparable?
- Are there missing values? What are the codes for missing values?
- Are there values below (or above) a detection limit? How are they registered in the database?
- Are there duplicated samples, that is locations where more than one measurement has been taken? This is frequent in quality control and quality assurance procedures. In this case, how are we going to handle the duplicates? Are we only keeping the original measurement for modeling?
- Are there categorical variables recorded with alphanumerical codes? Are these classes exhaustive and mutually exclusive? This means there is one and only one of the classes at each location.

- Are there ordinal categorical variables, that is where the number represents intensity, but with no objective scale? This is typical in geological logging of alterations or mineralization, for example.
- ...

Be aware of the data you will be working with. Understand the variables, the meaning of the categories and codes, and make sure the data are correct. Check the ranges, the categorical codes, the anomalous values or symbols. Correct everything as early as you can. Many of the questions listed above require particular solutions that will not necessarily be covered in these notes. However, when possible, we will provide some guidance as to how to handle these kinds of issues.

3 Exploratory data analysis

Exploratory data analysis (EDA) is one of the most fundamental tasks in any geostatistical modeling exercise. It is the task where we get to know the data and understand its possible limitations, complexities, and peculiarities.

During EDA, we will spend lots of time summarizing, sorting, and visualizing the data. At this stage, we should spot important flaws on the data, we should identify incorrect measurements, outliers and the nuances of different data types. This is usually when you can spot issues with data manipulation, such as transference from one software to another, where decimal places were truncated, or columns were incorrectly formated.

Data comes in very different forms, for different projects. In classical resource estimation, data often originates in an exploration drilling campaign, that is, it consists of samples taken at particular coordinates in space, which include analytical and interpretative data. Analytical data are obtained from a chemical laboratory where concentrations of different elements (potentially tens of elements) have been quantitatively assessed. In addition to these measurements, interpretative information is available for each sample. Other measurements such as mineral proportions can be provided through different methods.

Geological attributes are characterized by looking at the lithologies present in each rock fragment belonging to the sample, the alteration type (linked to the geological genesis of the deposit), the mineralization zone, the texture or fabric, the geotechnical properties, etc. Many of these features can be deemed qualitative or semi-quantitative (which is nothing more than an euphemism for imprecise), and are prone to error, as they are often subjective interpretations: a different geologist may “label” these attributes differently. However, quantitative data (such as geochemical analyses) can be used to quantitatively cluster these features and provide a consistent representation of these “subjective” classes.

In addition to these characteristics, many other attributes can be included in the database. Mineralogical proportions (either obtained qualitatively or quantitatively), intensities of alteration and mineralization, color, joints frequency, RQD (rock quality designation), hyperspectral response, density, humidity, particle size distribution (or some related parameter), etc. Furthermore, the result of metallurgical tests are often collected for geometallurgical modeling purposes.

Typical measurements are hardness, grindability, recovery, acid and reagent consumption, etc.

No matter what the variables are, these must be treated numerically in order to become relevant information for building the model. In the case of numerical values coming from continuous variables, processing is relatively straightforward. For numerical values that are categorical, where no order relation exists among the codes, one option is to code these variables as indicators valued 1 if the category is present at one location or 0 if not [1]. With this approach, a single categorical variable with K categories, will become a K dimensional vector of (binary) indicators. Ordinal variables (variables with a relative order between the categories), images, color, texture are much more difficult to incorporate into the models, although there are ways to "quantify" them.

EDA proceeds usually by pooling all the data together, and computing some summary statistics, building some basic visualizations (both in two and three dimensions), and then splitting the data by value range or categories, by applying filters. Further to this, correlations between variables are studied by visualizing scatter plots, and computing (linear) correlation coefficients. Proportions of categorical variables are also computed. Finally, the variation of the attributes in space is also explored at this stage, by creating spatial trend plots to see how means and variances change with spatial coordinates. This gives a first idea to support the decision of stationarity. Stationarity is necessary to pool together the data for statistical inference and will be discussed in detail later.

[1] these are sometimes callled dummy variables in machine learning

This process can become tedious and complex. Keeping a systematic and orderly procedure will help us expedite the process and clarify our conclusions. Using scripts to run repeated processes over multiple variables, or applying filters is an important skill that is needed.

Recipe for EDA:

- Identify all your variables
- Ensure they are ready for statistical analysis
- Summarize the statistics of all your variables
- Visualize all your variables
- Understand relationships between variables
- Understand how different categories have different behaviors
- Understand how variables change in different zones of the space

The main takeaway of EDA is becoming familiar with the data and having a preliminary understanding of the key relationships between variables. Having summaries, visualizations and preliminary statistics will facilitate interpretation of the results as we move forward.

4 Definition of domains

The notion of a domain is quite vague. Depending on the purpose of the model, domains will be defined to determine the volumes over which data are pooled together and used for estimation or simulation of blocks contained in those same domains.

Notice that this requires two processes:

- The first one is a clustering process, where the sample data are analyzed to define "populations". In most cases, this is done by the exploration geologist based mostly on a combination of geological attributes (consistency of geological properties) and spatial continuity (geological volumes are continuous). Several new approaches have been proposed in recent years using machine learning methods for clustering, that are either combined with spatial criteria or modified, so that the distance metrics used therein are linked to spatial continuity and anisotropy.

- The second process is that of inferring the extent of the domains beyond the samples. This is usually an interpolation problem, where every point in the spatial domain is assigned one of the groups (clusters or domains) defined in the previous step. Of course, this stage has uncertainty in the inference of the extent of these domains. Uncertainty in these volumes should be accounted for and carried downstream for decision making.

For resource estimation, we look for volumes where the properties of the rock are similar, in terms of the features that control the grade. Typically, geological logging will include the lithology, mineralization zone and alteration type. Sometimes, the structures are also featured. The relevant grades are compared within these different units. For example, we check the grade distribution in different lithologies, to see if some lithologies concentrate the samples with

higher grades. In that way, we can identify groups of geological characteristics that determine whether grades will be high or low. Of course, this is a very ad hoc decision, and no hard rules can be defined as how to proceed to define the domains. Recall that these domains are our construct of the grouping that the data has in this geological context, but these are not “real”; they are just useful to constrain the models we build. This actually tends to confuse people that come from teh machine learning community a lot, since there is no clear way we can validate the labels defined by the clustering method. The quality of the domaining is thus subjective and we will only see the consequences of it at the end of the modeling process, when we validate or reconcile the data with production information.

A general rule is that domains will depend on the geological characteristics that distinguish between mineralized and non-mineralized rock. Furthermore, different degrees of intensity of the mineralization can be considered, or the fact that mineralized rock may require different metallurgical processes to recover the metal or element of interest. In this case, we say our domains are geometallurgical. In some cases, the hardness of the materials fed to the processing plant will be relevant, as this may determine the recovery and tehrefore the quality and cost of the final product. So, domains really depend on the purpose of the model.

A very simple example for domaining is the difference between oxides and sulfides in porphyry copper deposits. Oxides and sulfides require different processes to recover a metal, thus characterizing where the contact between oxide and sulfide mineralization is, determines two domains. In reality, things are not that simple and a transition zone where mixed oxides and sulfides exist will complicate our

decision. Furthermore, within the sulfides, different levels of enrichment can be found, hence different mineralogies are present in the rock, which determines different processing conditions. Therefore these main mineralogical units will also be divided in distinct domains. Finally, rock where the metal has been leached (typically gravels) is often found laying over the oxides. This zonation is typical for example in porphyry copper deposits, and similar behaviors are expected in different types of deposit. This knowledge helps guiding our decision about how to define the domains. Unfortunately, no two deposits are alike, so the definition of domains must be well thought in every case, and will usually suffer some revisions and updates as new information is gathered. In general each genetic model of ore deposit will provide guidance regarding which distinct domains should be recognized, but as previously said, these are not definitive and should be defined looking at the specific conditions of the deposit.

Conventionally, a single deterministic interpretation is done, which locks the volumes of different ore types, but the actual types will be unveiled during production and should be used to feed back into the interpretation and update the models. It is clear that the use of deterministic models is risky, since the risk associated to the volumetric quantification of different materials is not accounted for. In reality, the volumes and tonnages extracted for each domain will vary, as the deterministic model is not accurate. This may have significant impact in the economic of the projects.

Now, once the domains have been defined, it is expected that within each domain, all points behave in the same manner. This is quite a stretch, but it stems from the fact that these domains will be used to perform statistical inference

and because of this, we must pool together samples for inference (otherwise we cannot go beyond the data). Thus, we do not want to "mix oranges and apples". All points within a domain should have the same (statistical) properties. We will kater see what this means when we introduce the notion of stationarity, but for now, let's just say that wherever we look in the domain, we should see the same properties for the distribution of the variable we are modeling. In particular, one would expect to see the same average value and same dispersion around that average, as well as the same spatial relationships between points (the same "texture" of values) within the different areas of the domain.

5 Spatial continuity analysis

A key consideration in geostatistics is to measure and take advantage of spatial continuity. This builds from the very intuitive idea that things that are close, should be similar. When looking at grades or concentrations, for example, considering that these concentrations come from a geological process, it makes sense to assume that grades will be similar when looking at two points a short distance apart, and that they will become more different (dissimilar), as that distance increases. This geological process can be a deposition from a sedimentary process, or a hydrothermal flow.

This intuitive notion is conventionally captured mathematically through three measures of spatial continuity that we will review in more detail later.

The three conventional measures of continuity are:

- The correlogram
- The (spatial) covariance
- The variogram (or semi-variogram)

We shall introduce these measures later on in detail, but for now, it is important to state that the estimation techniques that will be introduced (and in fact, most of the geostatistical techniques) will call for these functions to find the estimates or the distribution parameters, when assessing uncertainty.

These measures capture the relationship between pairs of locations (hence, they are called two-point statistics). However, in some cases it is necessary to look at the relationship between multiple points to capture patterns that are not "seen" by these two-point measures. These are tools that are still under development, but that show a great promise to improve the models and that are very closely linked to pattern recognition, image analysis, and computer vision.

The fact that spatial correlation exists, prevents us from using traditional statistical tools. Most of these classical statistical tools assume samples are drawn independently, meaning that there is no correlation between them, when in fact this correlation exists (unless samples are really far apart). Hypothesis testing and many traditional techniques must then be adapted to account for this correlation.

6 Definition of neighborhood for sample search

Predicting the expected value and also the expected uncertainty associated to the value at an unsampled location are the two main questions addressed by geostatistical techniques. Samples taken in the domain are used as information to make inference.

The question then arises: which samples should I use to predict the value of the variable I am interested in, at a particular location?

In theory, all samples within the domain are relevant, since we assume that the properties of the variable are constant over the domain (stationarity). However, intuition suggests that we should use the closest samples, since, as we discussed earlier, these are the ones more correlated to the location; as samples are farther they become more dissimilar to the value at the location we are trying to predict, hence less relevant. So, for most estimation and simulation techniques, we define a limited number of samples laying on a neighborhood of the location under consideration. Notice that the selection must not be based on distance only, but rather on correlation. When the rock has been folded after the mineralization has occurred, one should consider geodesic distances. Many new algorithms have been proposed to deal with this by using a locally varying anisotropy field. This means that the spatial correlation changes orientation and potentially also intensity (spatial range) depending on the location. This is already a departure of the stationarity assumption made earlier, but the aim is to better represent the reality.

In resource estimation, we are interested in building a block model, therefore, we need to estimate the grade at each one of those blocks, so the neighborhood needs to be defined for each block. Some basic rules of thumb are applied to define how far and which samples are considered to estimate the grade at each particular location.

Considerations for defining the search neighborhood:

- Anisotropy must account for the fact that correlation may be greater in one direction than in others
- Minimum number of samples to ensure we achieve a reliable estimate
- Maximum number of samples to limit the computation time and avoid over smoothing
- Use of octants or maximum number of samples per drillhole to ensure we interpolate rather than extrapolate in any particular direction

These tricks in the definition of the neighborhoods for sample search have been drawn over the years by experience, but we can find a reasonable theoretical explanation for most of them.

7 Estimation or simulation

The model can represent the expected value (a prediction) of the variable at every location, or it can try to capture the heterogeneity or variability of its spatial distribution, trading off the local accuracy for a better representation of the spatial variability. This is achieved using stochastic simulation

techniques.

In order to introduce these concepts, a simple example can be considered.

A set of samples obtained from a drilling campaign over a deposit, has been logged and 7 rock types have been characterized and are displayed in **Figure 2**. In addition to these logged rock types, copper grades are available in every sample.

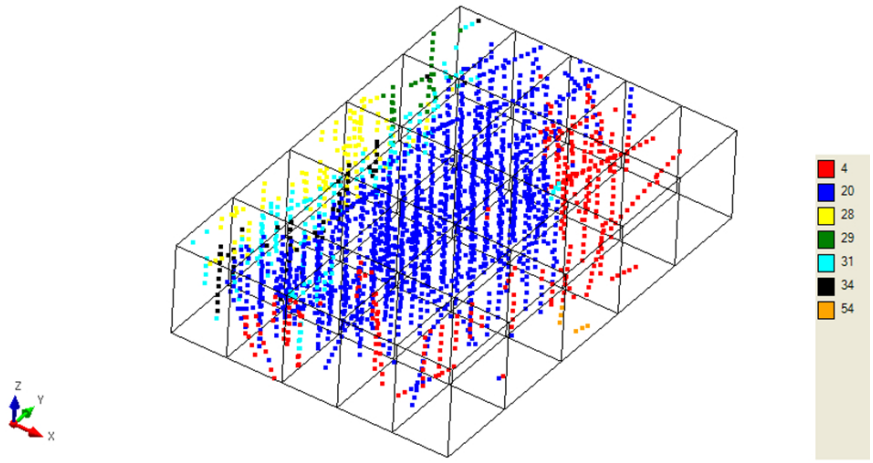

Figure 2: Example of samples

The extent of the geological units, defined here as the different rock types, is interpreted by the geology team. There are many approaches for this that will be discussed later, but for now, we will assume the following model is the "geological model" that the team has come up with (**Figure 3**). This model is important, as often the extent of categorical domains (in this case the rock types) controls the spatial distribution of the continuous variable of interest. Furthermore, as we fix the extent of these domains, we are locking

the volumes and tonnages in each unit. In an ore deposit, the lithological units, the alteration types, or the mineralization zones (or a combination of these) control the extent of grades of metals with economic value.

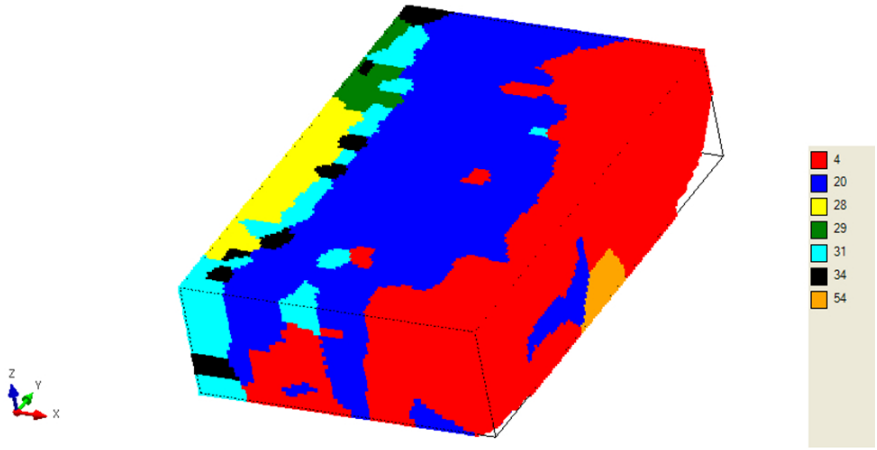

Figure 3: Geological model resulting from the interpretation of the available information

This interpretation is perfectly consistent with the available samples. For simplicity, it was built using a nearest neighbor approach, in this case, but in practice, the geological interpretation is a much more arduous exercise. Now, we know this is just an interpretation, therefore, the true extent of the geological units may be different. **Figure 4** shows a representation of the true unknown spatial distribution of the rock types. This is an intentionally exaggerated example, but it is evident from a simple visual inspection that our model does not represent the heterogeneity of the true spatial distribution of rock types. However, the general location of the units is correct. Notice that the true distribution and the model, both are perfectly consistent with the

available sample data.

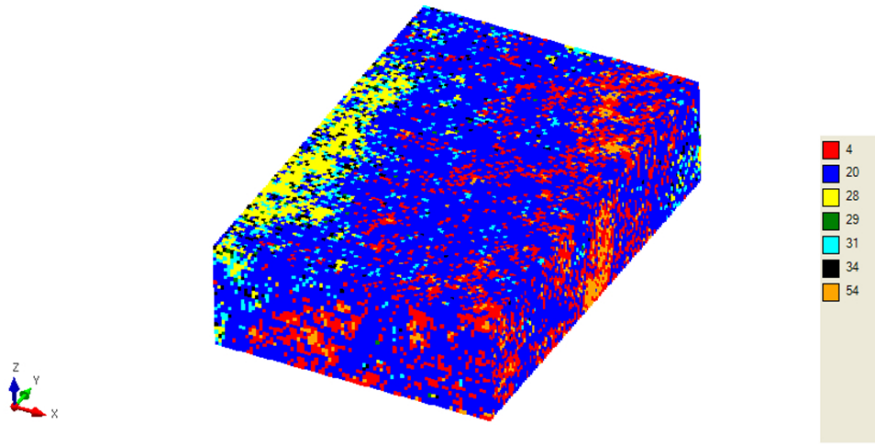

Figure 4: True (unknown) distribution of geological units (rock types)

So, clearly, one goal is to predict as accurately as possible the rock type in each location. The geological model achieves this reasonably well. Another goal, however, is to understand the heterogeneity in the spatial distribution of the units, to constrain the characterization of the grades.

This can be achieved by using simulation. Simulation builds alternative models of the spatial distribution of the categories, honoring the sample data at their locations, but also capturing the spatial continuity of the categories. Reproduction of the global proportions of the different categories in the domain is also a requirement for these methods. Models will match sample data at their location and will reflect more variability in areas far from data, since the uncertainty is larger at those locations.

Figure 5 shows the reference sample locations over a plan view and compares the interpretation with the actual distribution of the rock types.

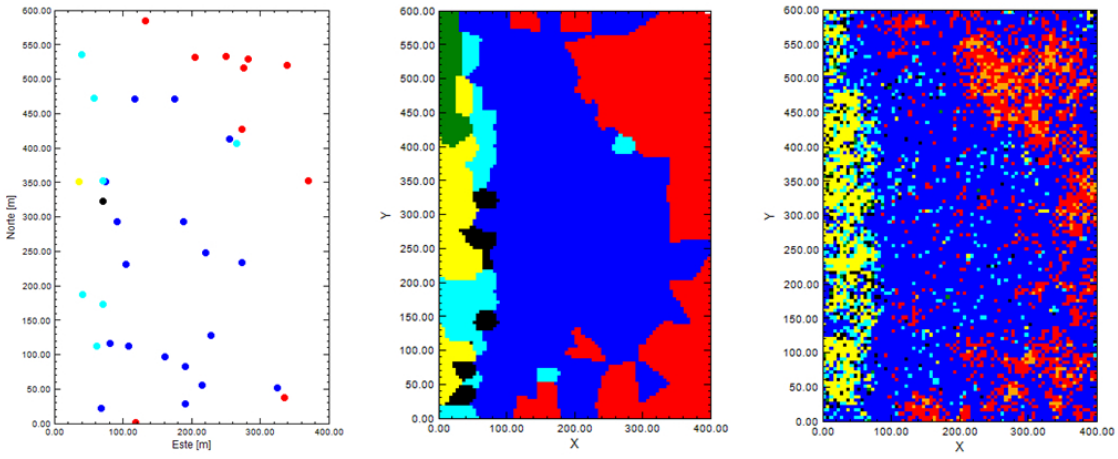

Figure 5: Comparison over a plan view of the sample locations (left), and the interpreted (center) and true (right) distributions of the rock types.

Figure 6 shows two realizations of a stochastic simulation to reproduce the spatial variability of the true distribution. It can be seen by visual inspection that the simulated models look like the real distribution, although all three are different.

The same can be done with continuous variables. We can predict or estimate the value at every location, which is going to smooth out some of its variability. Simulation can provide alternate models that, again, trade off the local accuracy to capture the spatial heterogeneity.

Figure 7 shows a plan view with estimated grades. This can be compared to three realizations obtained from simulation shown in **Figure 8**.

From this simple example, it is easy to see that in Geosciences a model can be built to account for the characteristics of categorical and continuous variables, using estima-

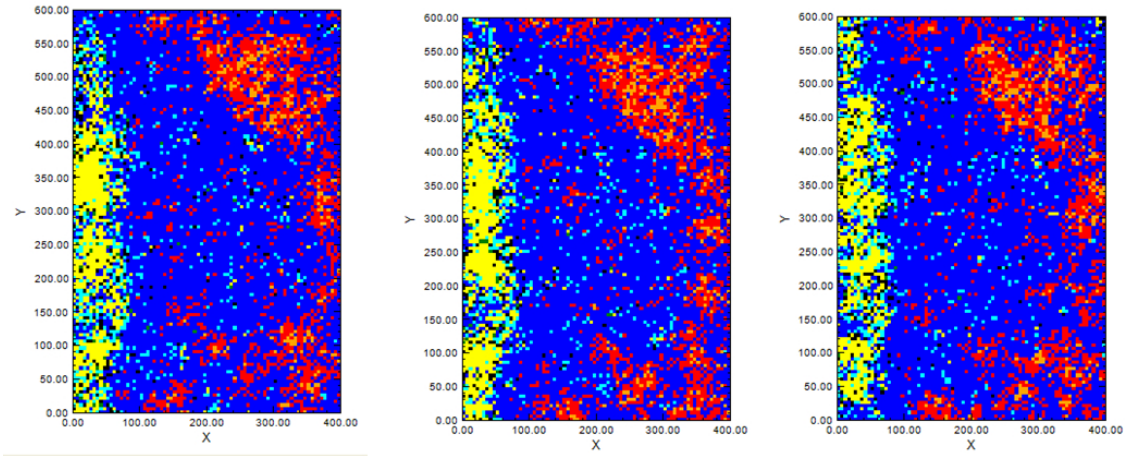

Figure 6: Two realizations obtained by stochastic simulation (left and center) compared with the true distribution of rock types (right).

tion (prediction) or simulation approaches.

Four combinations can be used to generate the final model:

- Categories are predicted (deterministic) and continuous variables are estimated (deterministic): this leads to a single model that does not account for uncertainty.
- Categories are predicted (deterministic) and continuous variables are simulated (stochastic): this leads to multiple models that account for uncertainty in the distribution of the continuous variables.
- Categories are simulated (stochastic) and continuous variables are estimated (deterministic): this leads to multiple models that account for uncertainty in the distribution of geological units, but not of the continuous variables.
- Categories are simulated (stochastic) and continuous variables are simulated (stochastic): this leads to multiple models that account for uncertainty in the distribution of geological units, and of continuous variables.

When simulating continuous variables, these are constrained to the extent of the domains, that is the statistical and spatial distributions of the continuous variables are inferred for each domain defined by the geological model. Boundaries between domains are usually considered hard, that is information from one domain is not used to estimate a block in another domain. In some instances, however, boundaries are soft, that is the information from one domain is shared to infer blocks in another domain, usually up

to a maximum distance.

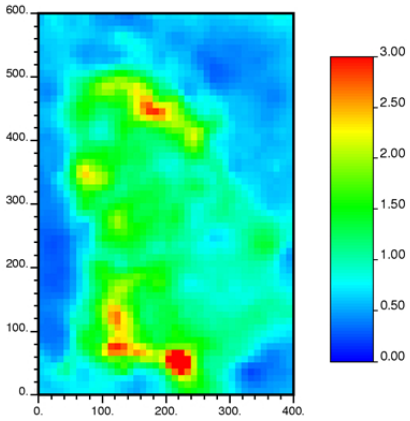

Figure 7: Block model of the grades obtained by estimation.

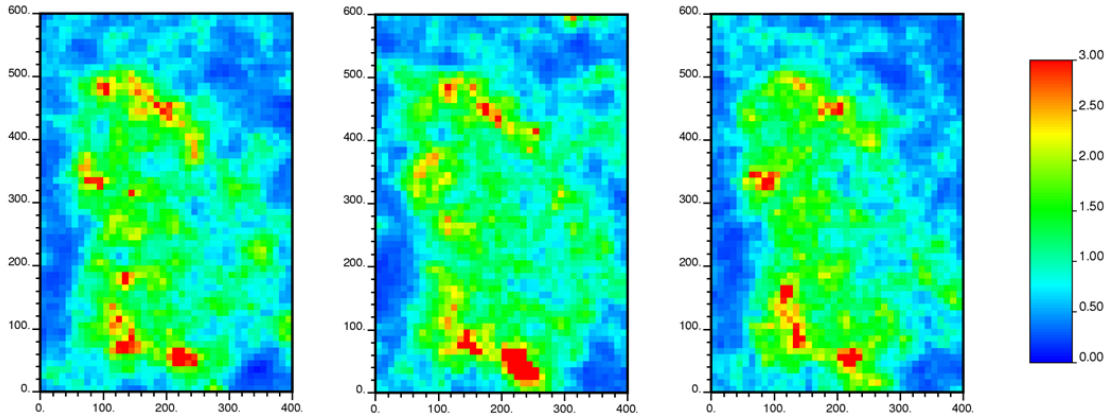

Figure 8: Three realizations obtained by stochastic simulation showing the distribution of block grades.

8 Post-process

The *in situ* model represents the statistical and spatial distribution of the properties we are interested in as they are on the ground. However, in most cases, we are interested in their response after a specific process is applied (the transfer function). For example, in mining, blocks are extracted, a selection is done to decide whether to process them or not and, if processed, a series of physical and chemical processes are applied to the materials to recover the metals of interest. In petroleum, the oil in place is recovered by injecting water and recovering the oil that flows through the permeable rocks. Similar examples can be found in problems related to groundwater and environmental applications.

The prediction of these processes is the ultimate goal of our modeling. We should be able to model and predict the response of these processes as well.

The estimated or simulated models generated using geostatistical techniques will normally need some kind of post-processing to be input into these process models. Typically, the model will need to be defined at a specific support, so a change-of-support is needed. In other cases, a threshold is applied to classify the materials into different categories (ore, low grade stock, waste).

We already discussed that estimated models do not reproduce the variability of the true variable, because they tend to smooth and only represent the trends found in the true distribution, missing the short range variability. It is easy to imagine that, if subject to the post-process discussed earlier, for example, a change of support, it will not represent correctly the true distribution of block values (since it

does not capture all the variability). Simulation, on the other hand, does represent the statistical and spatial heterogeneity of the true variable. Therefore, post-processing each realization should give a good representation of the behavior of the true variable. This makes simulation more difficult to apply, as each realization must be post-processed (increasing the work required).

9 Bibliographical notes

There are many textbooks on the subject of geostatistics and, in particular, in mining applications. An easy to read introduction is "An Introduction to Applied Geostatistics" by E.H. Isaaks and R.M. Srivastava [4]. "Geostatistics for Natural Resources Evaluation" by P. Goovaerts [3] is another excellent textbook with more advanced topics. Two excellent sources of additional information and case studies are the book by M.E. Rossi and C.V. Deutsch, "Mineral Resource Estimation" [5], and the book "Applied Mining Geology" by M. Abzalov [1]. The most comprehensive book in the matter is "Geostatistics Modeling Spatial Uncertainty" by J.P. Chilès and P. Delfiner [2].

References

[1] Abzalov, M. *Applied Mining Geology*. Springer, 2016.

[2] Chilès, J. P., and Delfiner, P. *Geostatistics Modeling Spatial Uncertainty*. John Wiley & Sons, New York, 1999.

[3] Goovaerts, P. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

[4] Isaaks, E. H., and Srivastava, R. M. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

[5] Rossi, M. E., and Deutsch, C. V. *Mineral Resource Estimation*. Springer, 2014.

Index