Geostatistics - Spatial Data Analysis

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

August 3, 2024

Summary

The statistical exploratory data analysis must be completed by exploring the spatial distribution of the data. In space, stationarity needs to be assumed in order to allow for statistical inference. The behavior of the variables in space needs to be understood in order to make the stationary decision and understand its potential drawbacks. On the other hand, we also need to define domains that will be used for estimation or simulation. These domains define the set of data used to infer the variable at unsampled locations within that same volume. Therefore, a representative distribution must be available within each domain to condition the inference process. The statistical distribution of the variables must be characterized within each domain, but oftentimes, the samples are not spatially representative, biasing the statistics that can be inferred from them.

In this chapter, we introduce the formal notion of stationarity and discuss issues related to inference of the representative distribution when faced with clustered or preferentially sampled data. We briefly touch on intentionally censored data.

1 Introduction

As soon as we start looking at a set of samples to infer properties that are common to them, we enter the realm of inference. We are interested in finding the set of samples and the associated volume, where the variables behave in a consistent manner. So far, we have not defined clearly this consistency, but let's say for now, that we want the variables to hold the same properties within this volume. This calls for an abstraction in order to go beyond the data. It would be impossible to learn about an unsampled location, unless we assume something about the properties of the variables of interest at that location. The natural step is to assume that "it behaves similarly to its neighbors". When we say that, we are referring to two different aspects: first, the statistical properties are similar, and second, the spatial properties are also similar to the neighbors.

When we think about the statistical properties, we assume the value at an unsampled location should be within the range of the known samples. We would expect it to be similar to its neighbors: if these are high, we expect to find a high value. Similarly, if the neighboring samples are low. On the other hand, we would be more confident on this if the values in the neighborhood are very homogeneous,

with low variability. Why would this point be any different? On the other hand, if the samples in the neighborhood vary a lot, then we would not be as sure about what to expect at this unsampled location.

So, it is clear that we can see a relationship between the variable at the unsampled location and at other known positions in space. We also realize that we cannot predict the value without error. A natural approach to handle this uncertainty is the use of probability theory.

We will formally introduce the concepts of random variable and random function, and then we will discuss the issues of statonarity and representative sampling.

2 The random function formalism

The unknown value at an unsampled location can be modeled using the concept of a random variable. This means that at every location, the regionalized variable $z(\mathbf{u})$ is modeled as a random variable $Z(\mathbf{u})$ that takes values according to a probability distribution. Notice the notation with lower case for the true measurable variable and upper case for the mathematical object. The entire goal of the steps that are going to be presented in the next sections and chapters is to infer the characteristics of the random variable at unsampled locations.

In addition to this perfectly reasonable goal of local inference, we will also be interested in understanding the *relationship* between the different locations. The random variables $Z(\mathbf{u_i})$ and $Z(\mathbf{u_j})$ are related because there is spatial correlation. We can extend this notion to the relation be-

tween multiple points. The collection of random variables within a domain is known as a random function. The random function describes the statistical properties of the random variables within the domain and their spatial relationships.

In geostatistics, inference is aimed at characterizing the behavior of the random function. We will see that this requires a statistical and a spatial analysis.

Definitions:

- Regionalized variable: is the variable that can be measured in the real world
- Random variable: is a mathematical abstraction, where the unknown value at an unsampled location is assumed to follow a probability distribution
- Random function: is the collection of random variables in a defined domain and it must be characterized from a statistical and spatial point of view
- Domain: is the volume within which the random variables have a homogeneous (stationary) behavior. Notice that stationarity is a decision made for modeling purposes

3 Stationarity

Stationarity refers to a notion of homogeneity within a domain. Intuitively, we expect to find the same behavior in any part of a domain. When we say behavior, we refer to statistical properties and spatial properties. For example, a homogeneous domain should have the same mean value

for the variable in different parts of that domain. We would not expect to find a homogeneous value, that is, the same value in every place of the domain, but we do expect to find values within the same range. Similarly, if values show a given level of variability in one part of the domain, we expect to find the same variability elsewhere in the domain.

Stationarity can be classified as:

- First order
- Second order
- Quasi second order
- Strict stationariity

First order stationarity implies that the means of the random variables within the domain are constant:

$$m(Z(\mathbf{u_i})) = constant \quad \forall \mathbf{u_i} \in D \tag{1}$$

This could be checked from the available data by computing an average over a moving window.

If a systematic change in the average value is seen over the domain, we could say that first order stationarity does not hold.

Second order stationarity requires, in addition to the condition over the means, that the second order moments of the random variables be constant. This means that the local variances are relatively constant over the domain:

$$\sigma^2_Z(\mathbf{u_i}) = constant \quad \forall \mathbf{u_i} \in D \tag{2}$$

It also means that the two-point relationships are constant over the domain. Thus, we should expect the same

amount of similarity between points separated by a distance $\mathbf{h}$ (this is actually a vector, with magnitude and direction) in any position within the domain D. We will come back to this later, once we introduce the measures of spatial continuity.

Quasi second order stationarity is a convenient approximation of the second order type of stationarity. It basically relaxes the condition to local neighborhoods.

$$m(Z(\mathbf{u_i})) = constant \quad \forall \mathbf{u_i} \in N_{\mathbf{u_j}} \tag{3}$$

$$\sigma^2_Z(\mathbf{u_i}) = constant \quad \forall \mathbf{u_i} \in N_{\mathbf{u_j}} \tag{4}$$

For example, we can say that the random function is second order stationary only within the neighborhood N around a location $\mathbf{u_j}$ we are trying to characterize. This really simplifies the decision of stationarity, since a change in the local mean (or in the variability) may exist, but we can still assume that within each local neighborhood things remain constant. This decision is used in ordinary kriging later on.

Strict stationarity is the assumption that all moments of the random function remain constant over the entire domain. Moments refer to different statistics of the random function, such as the mean, the variance and the spatial continuity as measured by two-point statistics, or by pattern statistics, that is, statistics that relate multiple points at the same time. This is an unrealistic assumption, but it may be required for inference and for developing some theoretical tools.

In summary, any obvious systematic change in the properties of the random function over the domain should be seen as a warning in terms of the assumptions we need to make about the random function when we are making infer-

ence about different statistics.

4 Domaining

The decision of stationarity is key for statistical inference. Inference must be done using individuals (samples) that belong to the same population (domain). Furthermore, in space, the statistical properties of the locations used to make inference about an unsampled location should be consistent, in order to avoid bias. If the spatial properties of the variable change with location, inference becomes problematic.

In order to handle this issue, domains are defined. In the context of modeling in geosciences, these domains must show similar properties from the geological point of view, and many times they should also show consistency in regards to the response variable. For example, in mineral deposits, a domain should group locations with similar grade distribution, but also the mineralization may be relevant if performance depends on it, as in the case of metallurgical recovery.

Stationary domains are then defined, based on the geological properties, and on the statistical properties of the variable being studied.

Defining domains for resource estimation (not really a recipe, but first steps that require further iterations with geological input):

- Perform statistical analysis of relevant grades by category of geological attribute (typically lithology, alteration type and mineralization zone)
- Identify categories that show a similar behavior in terms of the statistical distribution by looking at probability plots
- Merge geological categories, in light of the statistical similarity, their spatial location and geological consistency. For example, merge different types of similar rock types together, and do not merge materials that require different processing
- Update statistics and review

Systematic trends in the value of the variable sometimes exist. For example, the concentration of an element may change systematically with depth. In these cases, we can bypass the idea of a strict stationary domain by using the idea of quasi second order stationary, that is, we can assume that for each location, statistical (and geological) properties remain homogeneous within a neighborhood. This neighborhood can be defined to make the assumption work, that is, if a strong trend exists in the data, one can use a small enough neighborhood to assume stationarity within it.

In presence of systematic trends in the attribute:

- Plot local means and variances for different directions
- Determine whether the change in mean or variance is significant in relation to the scale of the model
- Determine if within a local neighborhood of enough size to find samples for interpolation the trend is still significant:
  - If the trend is not significant: use the local neighborhoods for estimation and simulation and make sure to check the model at the end to see how well it follows the trends
  - If the trend is still significant within the neighborhood, model the trend with a smooth interpolation, and compute residuals at the sample locations. Model the residuals with conventional geostatistical tools (estimation or simulation) and add the trend back at the end. In many cases, this approach requires more advanced techniques, such as modeling jointly the trend and residual

It is clear that there is a relationship between the scale of the possible trend and the size of the neighborhood. In addition to this, the availablity of sample data is important, as we need to find enough samples within each small neighborhood to permit inference. As a general rule, if there are abundant samples, the effect of the trend will be negligible, as enough information to inform local statistics will

exist within a small neighborhood. Trends become (very) problematic when scarce data exist, for example, in early exploration stages, where drillholes are far apart.

Defining stationary domains is a decision, a very important one, since it determines what data are used to make inference about unsampled locations within that region. Furthermore, defining the local quasi stationary neighborhoods is as important as the definition of domains.

In practice, domains are dictated by the geological understanding. Statistical and spatial analyses are done to assess how appropriate a stationary model is to the data. Exploratory data analysis tools are used for this purpose. If domains are homogeneous from their geological standpoint, but show trends, this can be handled by setting the appropriate local neighborhood when making inference about each location.

In some cases, this is not sufficient and the trend may need to be modeled and removed, in the hope that the residuals obtained through this process are stationary and capture some of the spatial structure. However, this is a tricky balance and no rule of thumb will be provided at this stage.

5 Representative sampling

One of the key questions raised during the analysis of the data is whether the sample available over the domain we are studying is fair, in the sense that it represents the properties of the domain. Notice that the term sample is being used here in a statistical sense. A sample is a collection

of individuals selected from the population. In the spatial context, each location corresponds to an individual from the population (the domain).

In sampling theory, two approaches are considered fair: pure random sampling and regular sampling.

In the case of pure random sampling, each individual belonging to the sample from the population is chosen at random. In our context, this means that each location within the domain can be selected with equal probability. No specific area within the domain should have a higher probability of being sampled.

On the other hand, another way of making the sample fair, is to sample in a regular fashion, so all possible individuals may belong to the sample (since the origin of the sampling grid is randomly selected). This can be extended to a stratified sampling, that is sampling randomly within regular strata that are predefined. Although we are not going to get into details of sampling theory, as long as the variable does not show a cyclic behavior and the strata are of different size than the cycle size, this method should also provide a fair sample.

In many applications in geosciences, sampling is not regular or random. This is due to different reasons: sometimes, we cannot access all areas for sampling, but in most cases, it is because we direct sampling to areas of interest. This naturally biases the result towards the “interesting” values. In mining, we tend to sample areas of high grades and over drill them to make sure we delineate the ore properly for extraction. The same happens in the oil industry. High permeability areas are sought after for production drilling.

Therefore, in most cases, sampling will not be fair, and attention must be paid to ensure we can infer a reasonable

estimate of the statistical distribution of the variables of interest within the modeling domain. This is extremely important when using some statistical and geostatistical tools. For example, in geostatistical simulation, the reference distribution is a key parameter. The simulated results should reproduce the reference histogram provided, thus, if this histogram is biased, the resulting simulations will be biased too, and will not represent the domain they are trying to characterize. Fortunately, in the case of estimation, kriging provides an unbiased estimate and weighs the data based on their spatial importance.

5.1 Declustering

There are several approaches to correct the statistical distribution, to ensure that it does correct for spatial bias, that is for clusters of data that have been gathered preferentially in some locations.

Let us first understand what the purpose of declustering is. Consider the following simplified example (**Figure 1**).

We can see that over the domain D, 8 samples have been taken. Initially, 4 samples were taken to "explore" the domain and a high value was found at the lower right quadrant. Since this value is an interesting value, infill samples have been taken around that location (samples 5 to 8).

If we were trying to represent the statistical distribution of values in this domain, we would consider the available information and build a histogram. Basically, we need to find the frequency associated to each value of the variable z (**Table 1**).

We can now group the values into the bins for the his-

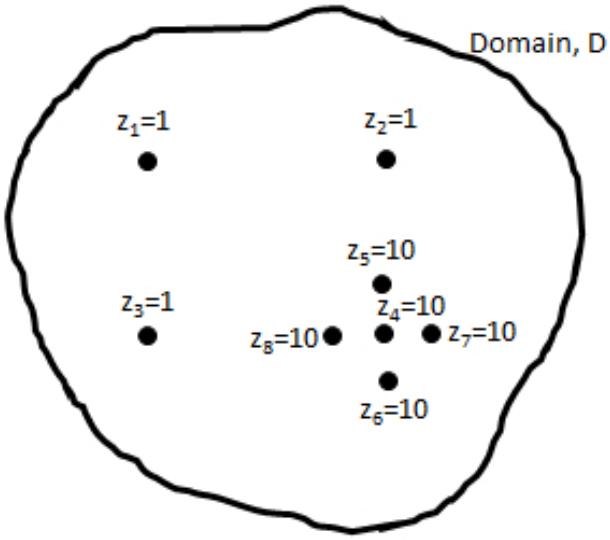


Figure 1: Clustered data

Sample	Value	Frequency	Weight
z_1	1	1	1/8
z_2	1	1	1/8
z_3	1	1	1/8
z_4	10	1	1/8
z_5	10	1	1/8
z_6	10	1	1/8
z_7	10	1	1/8
z_8	10	1	1/8

Table 1: Sample values, frequencies and weights in histogram calculation for clustered data

togram, and add the frequencies and the weights (**Table 2**), so statistics can be computed for that distribution.

The average value is:

Value	Frequency	Weight
1	3	3/8
10	5	5/8

Table 2: Cumulative frequencies and weights in histogram calculation for clustered data

$$\begin{aligned} \bar{z} &= \sum z \cdot f(z) \\ &= 1 \cdot \frac{3}{8} + 10 \cdot \frac{5}{8} = 6.625 \end{aligned} \tag{5}$$

And the variance[1] is:

$$\begin{aligned} \sigma_z^2 &= \sum (z - \bar{z})^2 \cdot f(z) \\ &= (1 - 6.625)^2 \cdot \frac{3}{8} + (10 - 6.625)^2 \cdot \frac{5}{8} = 18.984 \end{aligned} \tag{6}$$

Therefore, during the calculation of any statistics, we assign a weight to each value, linked to its frequency $f(z)$.

The histogram is built assigning an equal weight of 1/8 to each one of the 8 samples (**Figure 2**).

Now, looking at the spatial configuration, we realize sampling is preferential and we should "compensate" for this fact. A very simple approach would be to interpret that the domain is really represented in four quadrants (see **Figure 3**) and that the lower right quadrant has been over sampled.

[1] Notice that here we are using the estimator for the variance for large samples, which uses $1/n$ instead of $1/(n-1)$. The point of the example is understanding the weights assigned to each "squared difference".

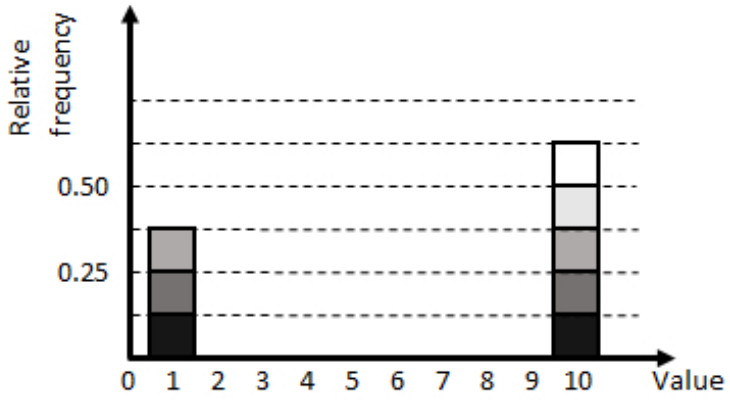


Figure 2: Histogram of clustered data

In that case, each zone would represent roughly a quarter of the entire domain.

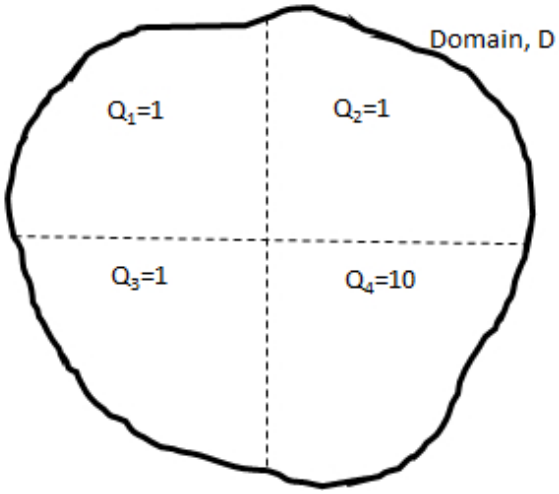


Figure 3: Interpreted quadrants in the domain

Equivalently, we can consider penalizing the weight assigned to samples that are redundant. For that, we can use

the same quadrants as a reference. Therefore, each quadrant will be assigned 1/4 of the total weight, and samples within the quadrant will be evenly weighted. This would lead to the weight configuration displayed in **Figure 4** and the corresponding histogram (**Figure 5**).

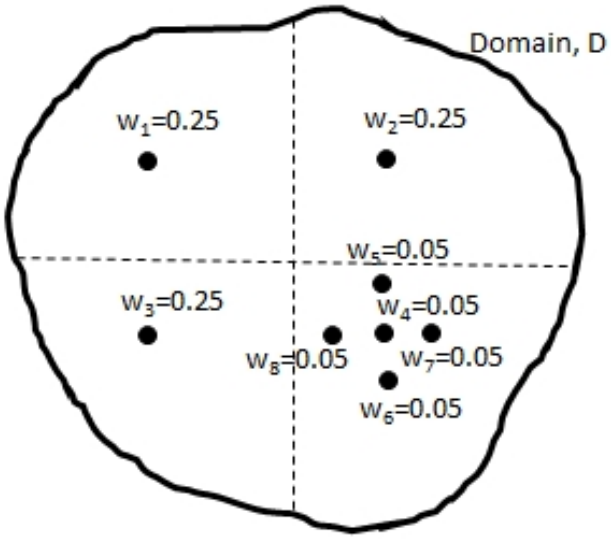


Figure 4: Weights assigned to each sample

The declustered statistics would be:

$$\begin{aligned} \overline{z_{decl}} &= \sum z \cdot f(z) \\ &= 3 \cdot \left(1 \cdot \frac{1}{4}\right) + 5 \cdot \left(10 \cdot \frac{1}{4} \cdot \frac{1}{5}\right) = 3.25 \end{aligned} \tag{7}$$

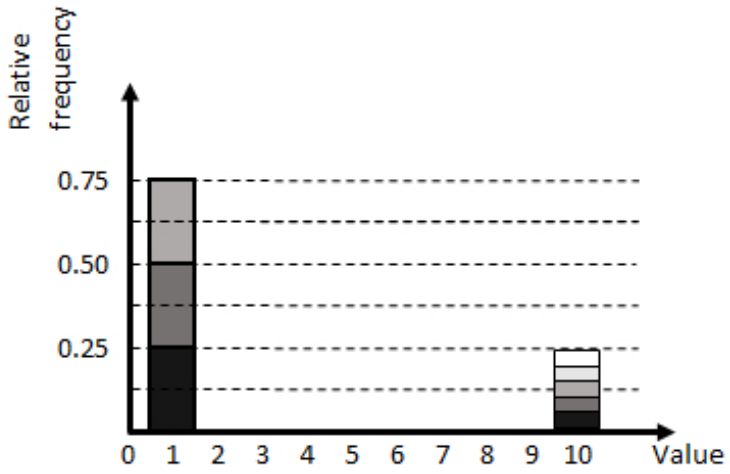


Figure 5: Declustered histogram

$$\sigma^2_{z_{decl}} = \sum (z - \overline{z_{decl}})^2 \cdot f(z) \quad (8)$$
$$= (1-3.25)^2 \cdot \frac{3}{4} + (10-3.25)^2 \cdot \frac{1}{4} = 15.1875$$

It is clear from this example that declustering is needed to obtain statistics that are more representative of the domain, but we will never be sure whether these corrected statistics actually match the reference distribution of the domain, unless we can exhaustively access all the domain locations.

We can see that the idea behind declustering is to modify the weight assigned to each sample when computing statistics or when building the histogram (which amounts to the same). Samples that are more redundant should get a lower weight, while samples that represent a larger volume, should get a larger weight.

Notice that the change really should depend on the spatial continuity of the variable. Think about a case where the variable at different locations shows no correlation. This means that knowing the value at one location does not inform about neighboring locations. The only thing we know for sure is that all these locations belong to the same domain. In that case, the histogram of the variable within that domain can be obtained by pooling together all the available sample, no matter where they are in space, since no sample is more redundant than the others, because there is no spatial correlation. All samples should have an equal weight in that case, so no correction for clusters is needed.

In a case with a significant spatial correlation (like the one shown in the example before), the declustering weights should penalize more significantly redundant samples. However, the tools available for declustering are today purely geometric, and do not take the spatial continuity into account. They are based on the idea of volume of influence to modify the weight.

When clusters of samples are preferentially located in high valued areas, the declustered distribution will have a lower mean. The declustered distribution assigns a non-uniform weight to the samples. The weights are determined by heuristic methods. Two main approaches are used for declustering: polygonal and cell declustering.

Polygonal declustering

Polygonal declustering amounts to assigning a weight to each sample that is proportional to the volume (or area) of influence.

This is achieved in 2D by defining a Voronoi tessellation, that is, a set of convex polygons are defined by intersecting half-planes defined by the line perpendicular to the line connecting each pair of samples in the plane, and that passes through the midpoint of that segment. In 3D, the same idea works by defining polyhedra using half-spaces, that is planes perpendicular to the midpoint of the line connecting pairs of samples in space.

Implementation of polygonal declustering can be easily achieved through a numerical approximation: a regular grid of points is defined over the volume, and the closest sample is found for each point in the grid. The relative frequency of points associated to each sample defines its weight. Notice that this can be done with arbitrarily high resolution, by increasing the density of the grid. Also, it is important to mention that the boundaries of the volume over which the weights are calculated can also be defined, so it works nicely with geological volumes, inside which we want to decluster the samples available.

Polygonal declustering is highly sensitive to the size of the domain, since samples in the boundary may get a higher weight if the domain is enlarged.

Cell declustering

Cell declustering works over domains without boundaries clearly defined.

The main idea is to define a regular grid of “cells” and assign an equal weight to each informed cell. Inside the cell, the samples found share the total weight of the cell uniformly. This is similar to the introductory example pre-

sented before.

The result depends on the cell size and on the origin of the grid of cells.

The cell size should be set to account for the regularity of the underlying grid (if available), aiming at having most of the time one sample per cell, and expecting to find more than one sample per cell in areas with clusters of samples.

5.2 Censored data

A brief note on censored data. In some instances, a variable may be intentionally censored. For example, it is measured only when the primary variable has a worthy value. In those cases, depending on the relationship between the two variables, this will bias the distribution and may even miss completely some range of possible values of the censored variable.

A real example occurs, when total copper is measured and soluble copper is only analyzed if the total copper value is higher than a given threshold. In these cases, the representative distribution of the censored variable cannot be recovered using declustering.

The censored bivariate distribution along with the (non-censored) distribution of the primary variable can be used to correct the bivariate weights and infer the corrected distribution of the secondary variable.

6 Example

Continuing the example presented earlier, we will analyze the distributions to determine reasonable domains for estimation and simulation. Then, we will apply declustering and discuss the result. If we look at the spatial distribution of the samples, coded with the rock types, we can see the spatial location of the different rock types (**Figure 6**).

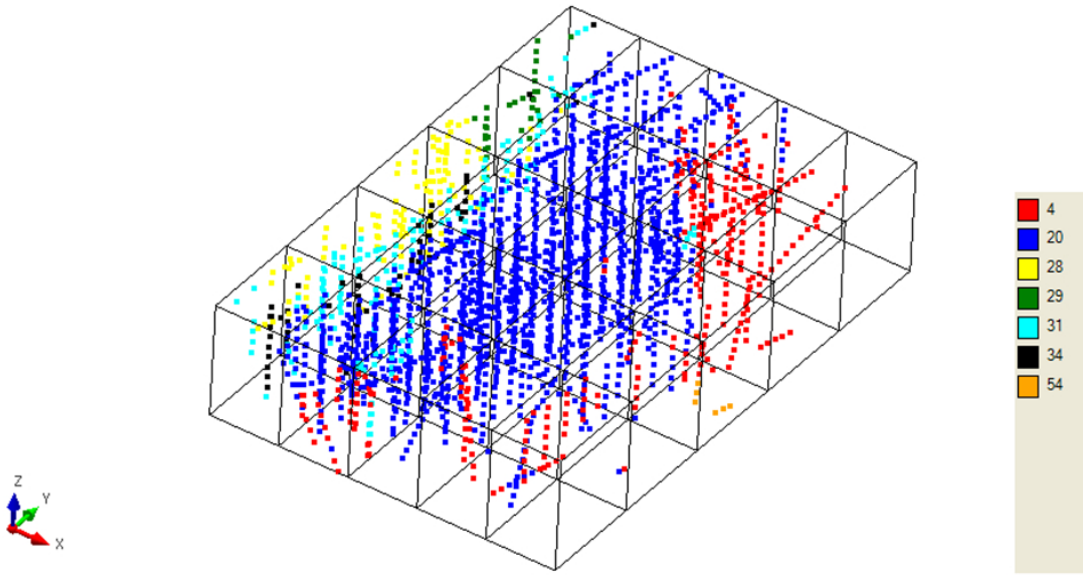
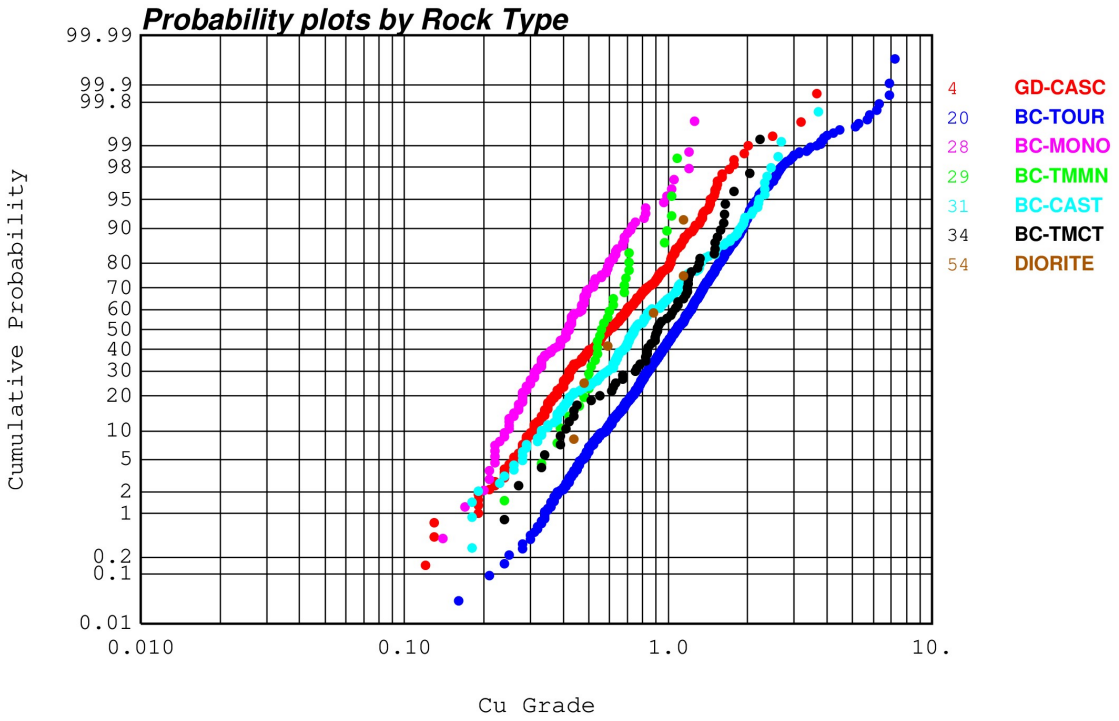


Figure 6: Location of samples coded by rock type

We compute the probability plots of the copper grade for the different rock types (**Figure 7**).

The description of the rock types is provided in **Table 3**. It can be seen that Unit 20 is the Tourmaline Breccia (in blue in **Figure 7**), the most relevant unit, with higher grades and volume within the deposit. The other units are smaller, with lower grades and surround the Tourmaline Breccia.



ID	REC #	MIN	MAX	MEAN	STD
4	348	0.120	3.650	0.707	0.430
20	1635	0.160	7.240	1.196	0.661
28	120	0.140	1.260	0.460	0.221
29	33	0.240	1.080	0.614	0.203
31	172	0.180	3.710	0.941	0.607
34	62	0.240	2.230	0.973	0.443
54	6	0.440	1.140	0.778	0.292

Figure 7: Probability plots of copper grade for the rock types

From the EDA done previously, we know that the sample information is a combination of vertical drilling and some inclined drillholes, with a composite length of 12m. The vertical drilling forms a regular triangular grid, where the spacing of the drillholes is about 30 to 35m.

Following the ideas described for declustering, and considering that we do not know at this stage the boundaries of

Code	Description	
4	Cascade Granodiorite	GD-CASC
20	Tourmaline Breccia	BC-TOUR
28	Monolith Breccia	BC-MONO
29	Tourmaline-Monolith Breccia	BC-TMMN
31	Castellana Breccia	BC-CAST
34	Tourmaline-Castellana Breccia	BC-TMCT
54	Diorite	DIORITE

Table 3: Rock code descriptions

unit 20, which we will model separately from the other rock types, we can apply cell declustering.

To show the effect of this method, we first present the change in the mean as a function of the cell size, parameterized by the size of the cell over the X direction (East), and considering a cell with anisotropy to match the "regular sampling" found in the EDA. This means that the cell should be 35 by 35 by 12m, in the X, Y and Z directions, respectively.

We try 50 cell sizes ranging from 5 to 500m (for the X dimension of the cell). The weights obtained are the average over 50 random origins of the grid used for declustering.

Figure 8 shows the change in the declustered mean with the cell size. This shows that the resulting statistics are highly sensitive to the cell size selected for declustering. Since we know that there is an underlying regular sampling grid in our data, we can use it to define the declustering weights. Any areas with denser sampling will penalize the weight assigned to the samples. Areas with scarce sampling will impose a higher weight to those samples.

By running the cell declustering algorithm (using a cell size of 35m), the corrected histogram and statistics can be obtained (see **Figure 9**).

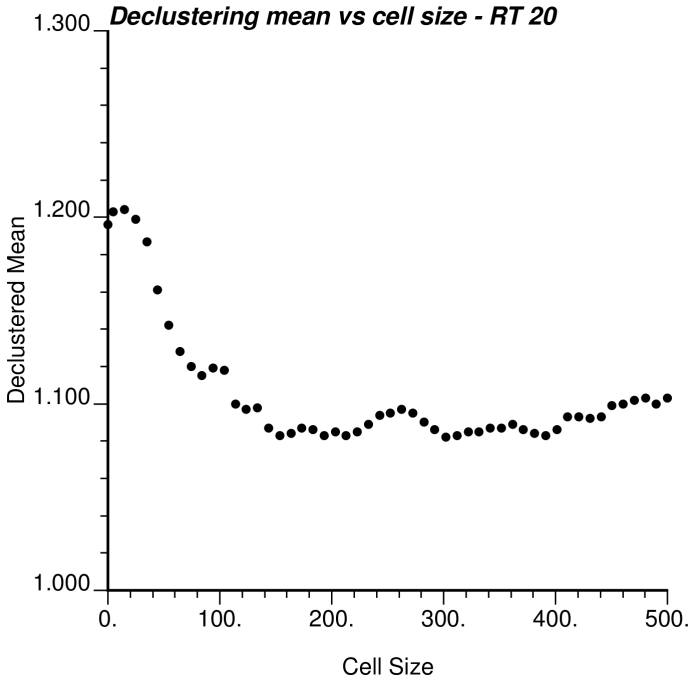


Figure 8: Declustered mean of copper grade in rock type 20 as a function of cell size

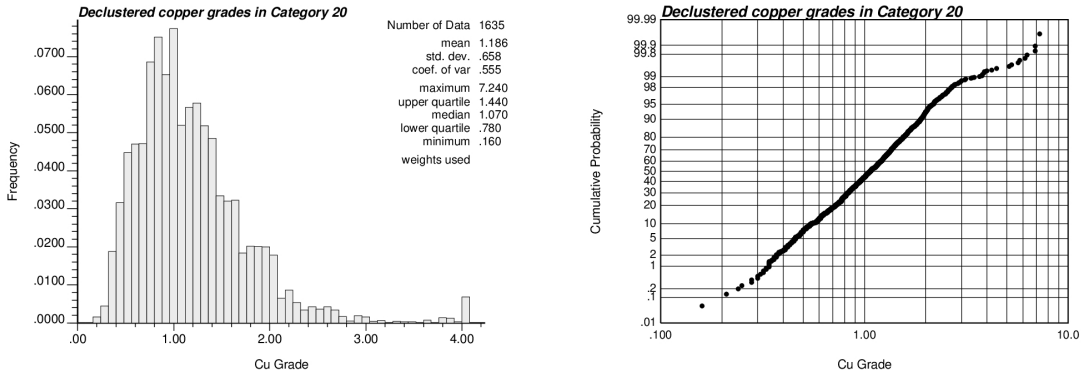


Figure 9: Histogram and probability plot of declustered copper grade for rock type 20

Finally, we can compare the declustered histogram with the raw distribution using a quantil-quantile plot (see **Figure 10**).

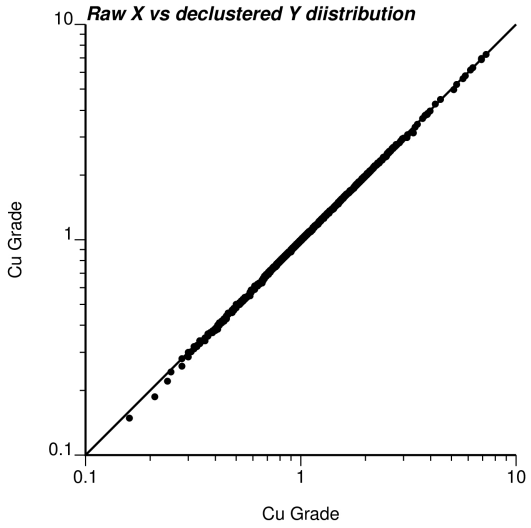


Figure 10: Quantile-quantile plot comparing the raw and declustered distributions (logarithmic scale used)

The spatial distribution of declustering weights is displayed in **Figure 11**. It is clear that samples clustered together get a lower weight, while isolated samples get a higher weight.

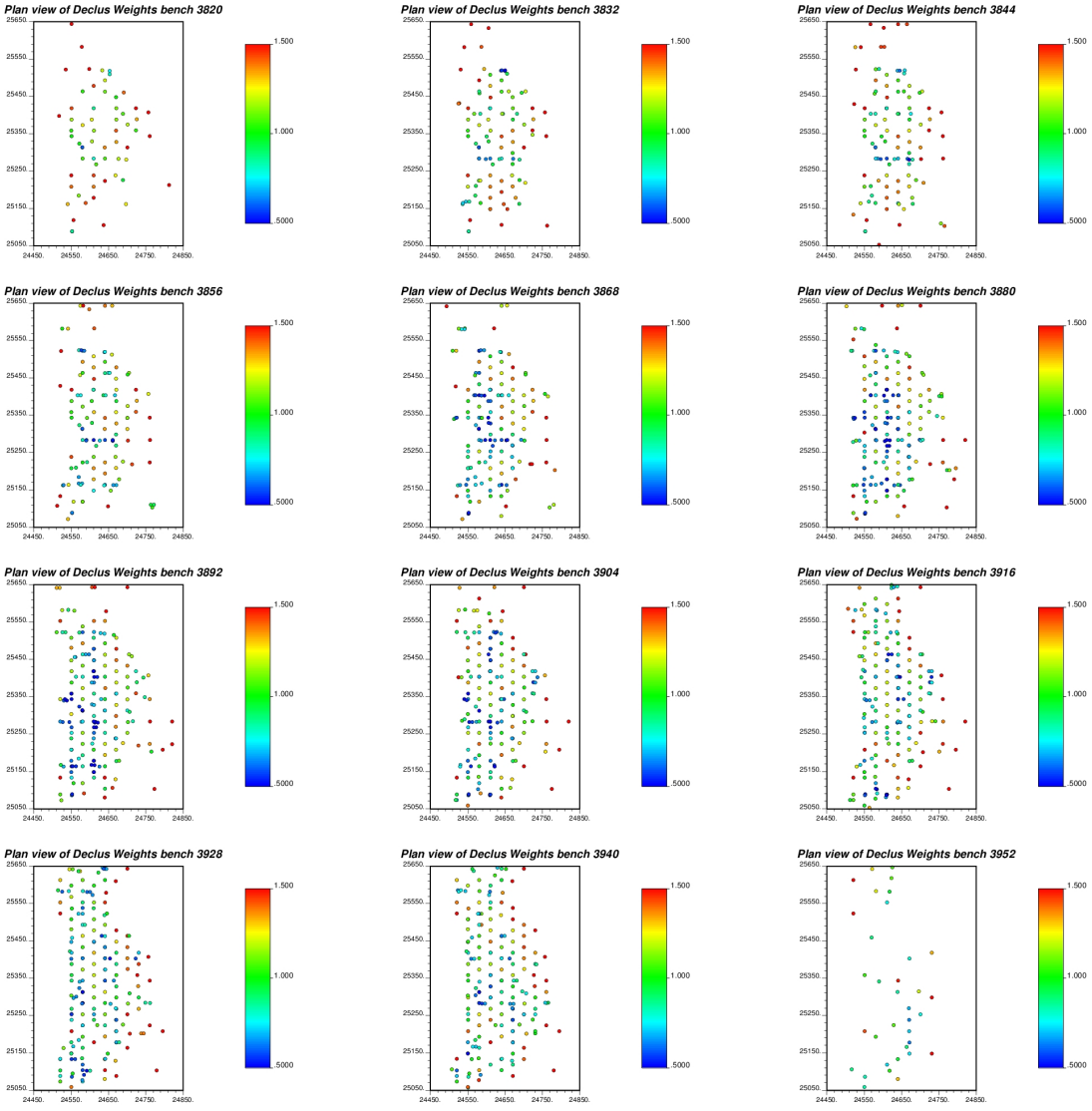


Figure 11: Plan views displaying declustering weights by bench

Index