Geostatistics - Data Analysis

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

July 27, 2024

Summary

Putting together a clean database, dealing with issues such as duplicate and missing data, are among the first tasks when developing a numerical model.

Then, the data are analyzed to understand the nature of different variables, their statistical distributions, their spatial distributions, statistical problems linked to spatial clustering, preferential sampling, censored distributions, relationships between variables, and stationarity.

The main goals of this stage are to create a clean consolidated database to work with, and to perform the basic exploratory analyses required to understand the data and the possible problems we will face during the subsequent modeling stages.

This stage is the preamble to the definition of the domains for modeling.

1 Introduction

In this chapter, we will divide the data analysis challenges into two different problems. First, we need to ensure the database is usable. Second, we need to analyze the information to become familiar with the variables and their specific issues.

1.1 Consolidating the database

We need to put all variables in a common database, so we can infer statistical and spatial distributions as well as relationships between the variables.

The following issues need to be addressed during this stage:

- Coordinates are complete and correct
- Duplicate data (at a particular coordinate) are identified. This refers to having more than one value for a variable at a particular location
- Data below (or above) a detection limit are identified and handled properly
- Zeros are confirmed as such and they are not missing values
- Codes for missing values are understood
- Categorical variables codes are understood
- The support of the measurements is understood. This refers to the volume over which the sample has been taken

Typical mistakes such as typos when entering coordinates or values are easily detected by inspecting maps or summary statistics. Duplicate data originates from quality assurance and quality control processes. Usually the samples are analyzed (primary record - original) and then some are reanalyzed (secondary record - duplicate), because there was a problem or because it was part of the routine quality checks. If an original observation is known to have a problem, then the duplicates can be used, granted it is confirmed as correct. In case a routine duplicate is available, the original should be used. Notice that it is not a good idea to average original and duplicate, as this changes the statistical characteristics of the result. The average is smoother than both the original and the duplicate, so it will be inconsistent with the other samples (those where only the original value is available).

Regarding the support of the samples, we will discuss the need for compositing later on. However, one should realize that, if samples have been analyzed over different support, that is, the volume represented by the value is different (for example a 1m long core compared to a 10m long core, when measuring a grade), then these cannot be deemed the “same” variable, since the statistical and spatial properties of the variables change with the change of support. Larger supports (more volume) will tend to be smoother in their statistical distribution and this will also carry some consequences in terms of their spatial properties. Make sure you are aware of the support of the samples, and if reporting sample statistics where the samples have varying supports, it is a good idea to state this. It often occurs that when sampling waste, a larger support is used (to save money).

We will not get into the details of how to fix issues found during this stage as this truly depends on the history of the database. Just make sure, you have checked the database before starting you analyses.

1.2 Exploratory data analysis

Once the database has been consolidated and is clean, the process of exploratory data analysis (EDA) begins.

A series of statistical and graphical analysis are performed to get a better understanding of the data, both statistically and in regards to their distribution in space.

The process typically includes the application of the following analyses to understand the statistical distribution of the relevant variables:

- Basic statistics of each continuous variable
- Proportions for categorical variables
- Report of missing values, zeros, values below detection limits
- Statistics filtered by categories for each categorical variable
- Histograms and probability plots of each variable, globally and by category
- Statistics regarding support of the samples (sample length or volume)
- Scatter plots for all the pairs of variables, globally and by category

In addition to these tasks, one needs to understand the relationship between variables and also in space. This makes the process quite challenging, especially when multiple continuous variables are available and several categorical attributes exist, as we need to explore all the cross relationships and filter by one or more of the categorical variables.

The spatial analysis of the data will involve:

- Maps of each variable in planview and sections
- Trend plots of each continuous variable
- Trend plots of proportions of categorical variables

2 Univariate statistical measures

Basic statistics must be reported for continuous and for categorical variables.

2.1 Continuous variables

A typical report of statistics for a continuous variable should include:

- Minimum and maximum
- Mean and median
- Standard deviation and variance
- Coefficient of variation
- Lower and upper quartile

Recall the formula for the mean or average:

$$m = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{1}$$

and the variance, which measures the dispersion of the samples:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m)^2 \tag{2}$$

Notice that the unbiased estimator of the variance considers a denominator of $n-1$, but for large samples, this does not really make a difference, considering that these statistics are purely descriptive.

The standard deviation is simply the square root of the variance, and is useful since it is measured in the same units as the original samples:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - m)^2} \tag{3}$$

The coefficient of variation measures the dispersion through the dimensionless ratio between the standard deviation and the mean:

$$CV = \frac{\sigma}{m} \tag{4}$$

Some other statistics may also be reported. For example, in some instances the skewness and kurtosis are included to give an idea of the shape of the distribution, however, it is better to include a graphical display of the distribution, such as a histogram.

In cases of distributions with very long tails, other quantiles can be reported, such as the 99% or other deemed relevant. Also, in these cases, using a logarithmic scale in the axis of the variable whyen building a histogram or probability plot, helps visualization.

These statistics should be seen as raw statistics, since no consideration has been given of how representative these are in the spatial domain.

2.2 Categorical variables

Categorical variables should be summarized by associating each code to a meaning (a name for a particular geological feature, or a level for an ordinal variable, such as an intensity), and to a proportion (obtained as the count for that category divided by all the observations).

3 Univariate graphical displays

In addition to a mere report of the statistics, the sample distribution can be depicted with the following graphical displays:

- Histogram, displaying absolute frequency (counts), relative frequencies (proportions), considering arithmetic or logarithmic scales
- Cumulative histogram
- Boxplot
- Probability plot, considering arithmetic or logarithmic scales

4 Visualization

Finally, maps of the variables with adequate labels and colors should be displayed. The idea is to visualize the spatial distribution, have a preliminary idea of spacing, spatial configuration, clusters, trends, etc. These plots are typically done in plan views and cross sections, with a specific “corridor” thickness, that is, only a band of coordinates is visualized on each view. In addition to those, three dimensional visualizations, where the data can be rotated to be inspected, helps understanding orientations and trends.

A set of representative plan views and cross sections should be created.

5 Bivariate statistical measures

Bivariate relationships are relevant because they can help in interpretation, for example, by understanding a correlation between variables, that may be linked to a geological phenomenon. Also, relationships between continuous and categorical variables are important. Conditional distributions capture the dependence between a continuous variable and a specific category of a discrete variable. This notion can be expanded (and will be used later) to pairs of continuous variables. The statistical distribution of one variable can be conditioned by a value (or a range of values) of another continuous variable. We briefly review the summary statistics and graphical displays typically used to analyze the relationships of pairs of variables.

5.1 Continuous variables

For pairs of continuous variables, the pairwise relationship can be summarized by characterizing the correlation. There are several ways of doing this:

- Covariance: this statistic is an extension of the notion of variance in the univariate case. It computes the average product of the difference of each variable with respect to its mean.
- Linear correlation coefficient (Pearson's): it is computed as the covariance between the two variables divided by the product of the standard deviations of each variable.
- Rank correlation coefficient (Spearman's): it is defined as the Pearson's correlation coefficient for the ranked variables.

The formula for the covariance between variables Z and Y is:

$$Cov_{ZY} = \frac{1}{n}\sum_{i=1}^{n}(z_i - m_Z)\cdot(y_i - m_Y) \tag{5}$$

The linear correlation coefficient is:

$$\rho_{ZY} = \frac{Cov_{ZY}}{\sigma_Z\cdot\sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m_Z)\cdot(y_i - m_Y)}{\sigma_Z\cdot\sigma_Y} \tag{6}$$

The rank correlation coefficient is:

$$\rho_{R_{ZY}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_{z_i} - m_{R_Z})\cdot(R_{y_i} - m_{R_Y})}{\sigma_{R_Z}\cdot\sigma_{R_Y}} \tag{7}$$

where the values R_{z_i} and R_{y_i} are the corresponding rankings of z_i and y_i, respectively, m_{R_Z}, m_{R_Y}, σ_{R_Z} and σ_{R_Y} are their corresponding means and standard deviations.

5.2 Categorical variables

Analyzing relationships among categorical variables can be done by building a table of frequencies (absolute or relative) to highlight matching categories from two variables. This is a typical analysis done for example, when comparing lithological types and alteration types, or when comparing after geological modeling, the degree of coincidence between the modeled units and the logged units at the samples.

6 Bivariate graphical displays

Bivariate graphical displays are useful to see the type of relationship existing between the variables (either continuous or categorical).

Typical plots are:

- Scatter plots
- Regression
- Conditional means

A scatter plot presents the relationship between two variables by plotting the pairs in a conventional $X - Y$ plot. Through visual inspection of the plot, it is easy to detect linear or non-linear relationships, positive or negative correlation, anomalous data, constraints in the relationship, and variability increase for some ranges of one of the variables (known as heteroscedasticity).

Scatter plots can be used to plot continuous variables versus coordinates, to detect trends in the data, however the cloud of points tend to be too noisy, therefore it is recommended to compute means for ranges of coordinates. These plots are known as conditional means. Finally, adding a regression line (or a piece-wise regression) helps understanding the linear relationship between the variables.

7 Example

We will now review the different tools and their application, by introducing an example database that will be used to illustrate some of the analyses and methods discussed.

We will consider two databases belonging to a deposit located in the Central Andes, and belonging to the Andina Division of Codelco Chile. A diamond drilling campaign generated samples that were composited to 12m constant length intervals. A denser blasthole database is also available, since the project has been mined out. these samples are drilled vertically and grades are reported at bench height, which is 12m. The deposit is characterized by a breccia complex within a porphyry copper mineralization, and has seven main units:

- Cascade granodiorite: located in the eastern and southern parts of the area. It hosts the breccia complex
- Diorite: also located in the eastern area
- Tourmaline breccia: located in the central part of the area. It is the dominant unit. The breccia is composed

of granodiorite clasts in a matrix cement dominated by tourmaline and sulphides (chalcopyrite, pyrite, molybdenite and some bornite)

- Other smaller breccias outcropping mostly in the western and southern areas:
  - Castellana breccia (rock flour breccia)
  - Monolith breccia
  - Tourmaline-Monolith breccia
  - Tourmaline-Castellana breccia

Only analyses for the drillhole database are shown at this point.

Basic statistics are obtained and are reported in **Table 1**. Only relevant values are displayed. For example, for coordinates, we really only care about the range, so only the minimum and maximum are reported (in addition to the number of valid data). For rock types, also only the range and the number of valid data are presented. For copper grades, it can be seen that the dataset ranges from 0.12% to 7.24%, so there are no zeros and all values seem to be above a detection limit.

	East	North	Elevation	Cu Grade (%)	Rock Type
Number of Data	2376	2376	2376	2376	2376
Mean				1.054	
Std. Dev.				0.645	
Coef. of Var.				0.612	
Maximum	24849.0	25648.9	3950.0	7.240	54
Upper Quartile				1.330	
Median				0.940	
Lower Quartile				0.625	
Minimum	24450.0	25052.2	3820.0	0.120	4

Table 1: Basic statistics for drillhole dataset

Histograms and probability plots also help gaining understanding about the data. **Figures 1** and **2** show these figures for all the variables.

Histograms of coordinates show cycles of high frequency for some coordinates, due to the nature of the drilling campaign, where lines of drillholes are drilled at particular positions. This is clearer when inspecting the maps (see later).

The histogram of copper grades looks fairly well-behaved, with the bulk of the samples in the low grade range, with a mode around 1.0%Cu and a tail to the right with very low frequency, all the way up to the maximum. Notice that the column displayed at 4.0%Cu accumulates all the values above that grade. The histogram is slightly asymmetric, which suggest a lognormal shape.

The histogram of rock type codes show 7 classes with samples. Here one needs to be careful and check the data using filters to ensure only 7 categories exist. It is clear that category 20 is the most frequent, accounting for nearly 70% of the samples.

Probability plots allow to see the cumulative distribution. The *Y* axis of the plot is the cumulative frequency, but the scale represents the quantiles of a normal (or Gaussian) distribution. Extreme lows and highs are somehow magnified, so it is easy to visualize the tails of the distribution. It is easy to see that if the variable has a Gaussian distribution, when plotted in a probability plot, the cumulative curve will be a straight line. However, in Earth Sciences concentrations of elements tend to show a lognormal behavior, that is the logarithm of the variable is normally distributed. Therefore, if the probability plot is built using a logarithmic scale in the *X* axis, we will be comparing the cumulative logarithm of the variable with the quantile of a Gaussian distribution.

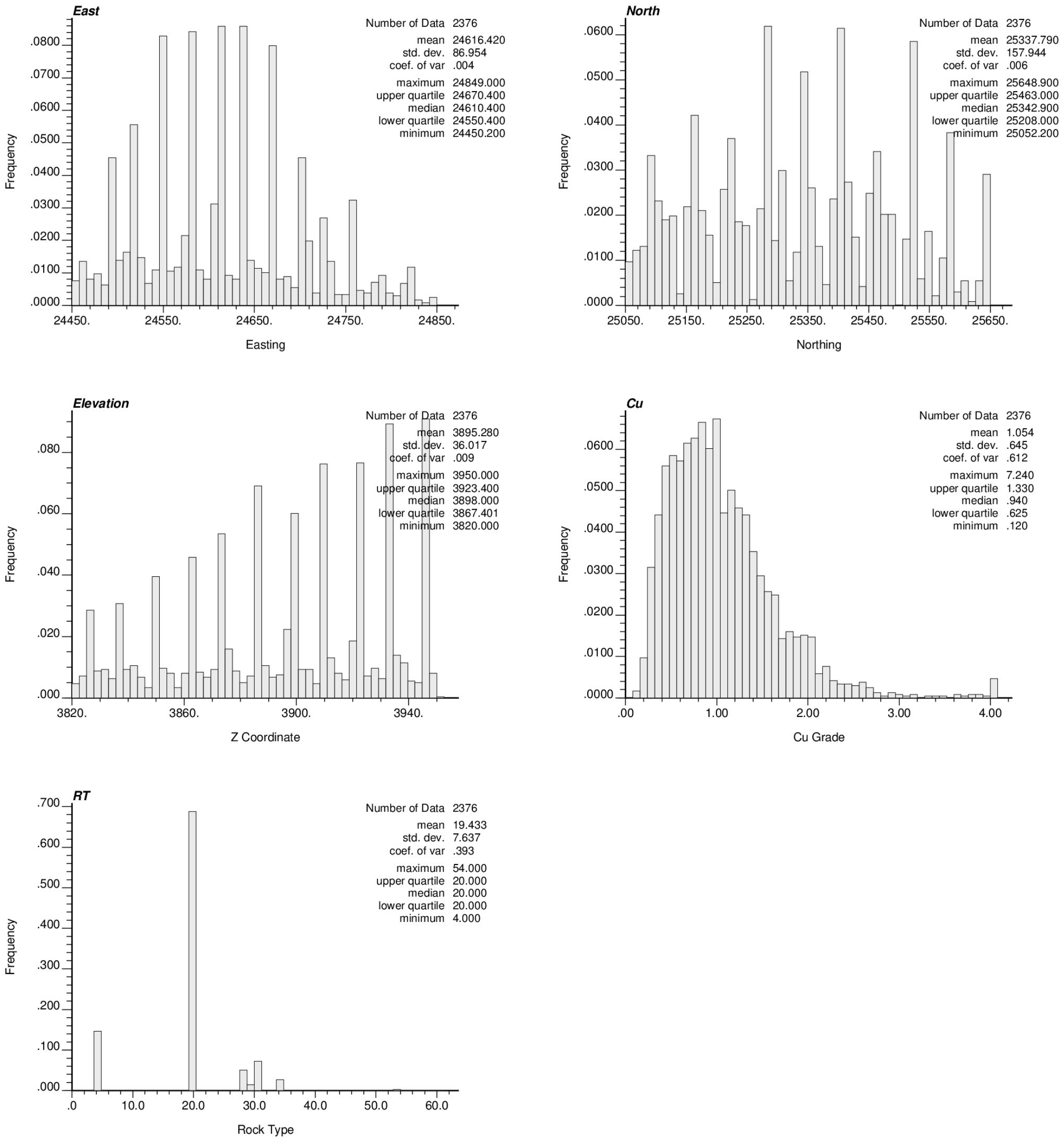


Figure 1: Histograms of all the variables in the drillhole dataset

But, since the logarithm of the variable in a lognormally distributed variable is normally distributed, it will show as a straight line. In summary, probability plots are a visual tool that can help understand whether the variables we are analyzing are normally or lognormally distributed.

When looking at the probability plot of copper grades in **Figure 1**, it can be seen that the *X* axis has been modified to reflect a logarithmic scale. The cumulative distribution in the plot, shows a fairly linear behavior. This is an indication that the copper grade distribution is close to a lognormal distribution.

Probability plots are also good to see categorical variables, such as the rock type.we can clearly distinguish between the different categories and have a first approximate estimate of the proportions. Looking at the bottom plot in **Figure 2**, we can see that rock type category 4 has a frequency of about 15%, category 20 about 70%, while the others are rather small.

Maps and cross sections are shown in **Figure 3** and plan views over 12m slices are presented in **Figure 4**.

From **Figure 3**, we can see that many drillholes are vertical, but some are inclined. There is a fairly regular grid at the center of the deposit, that seems to be like a triangular grid. Grades have lower grades toward the boundaries of the plane, with two areas of higher grades, slightly towards the north and south of the center of the domain. No particular trend is seen vertically. The bench by bench plan views in **Figure 4** confirm some of these findings.

The plots for the rock type are not very useful, since the color scale is not appropriate.

We complete this preliminary exploratory data analysis by looking at the grade distribution per rock type. **Figures 5** and **6** display histograms and probability plots for each

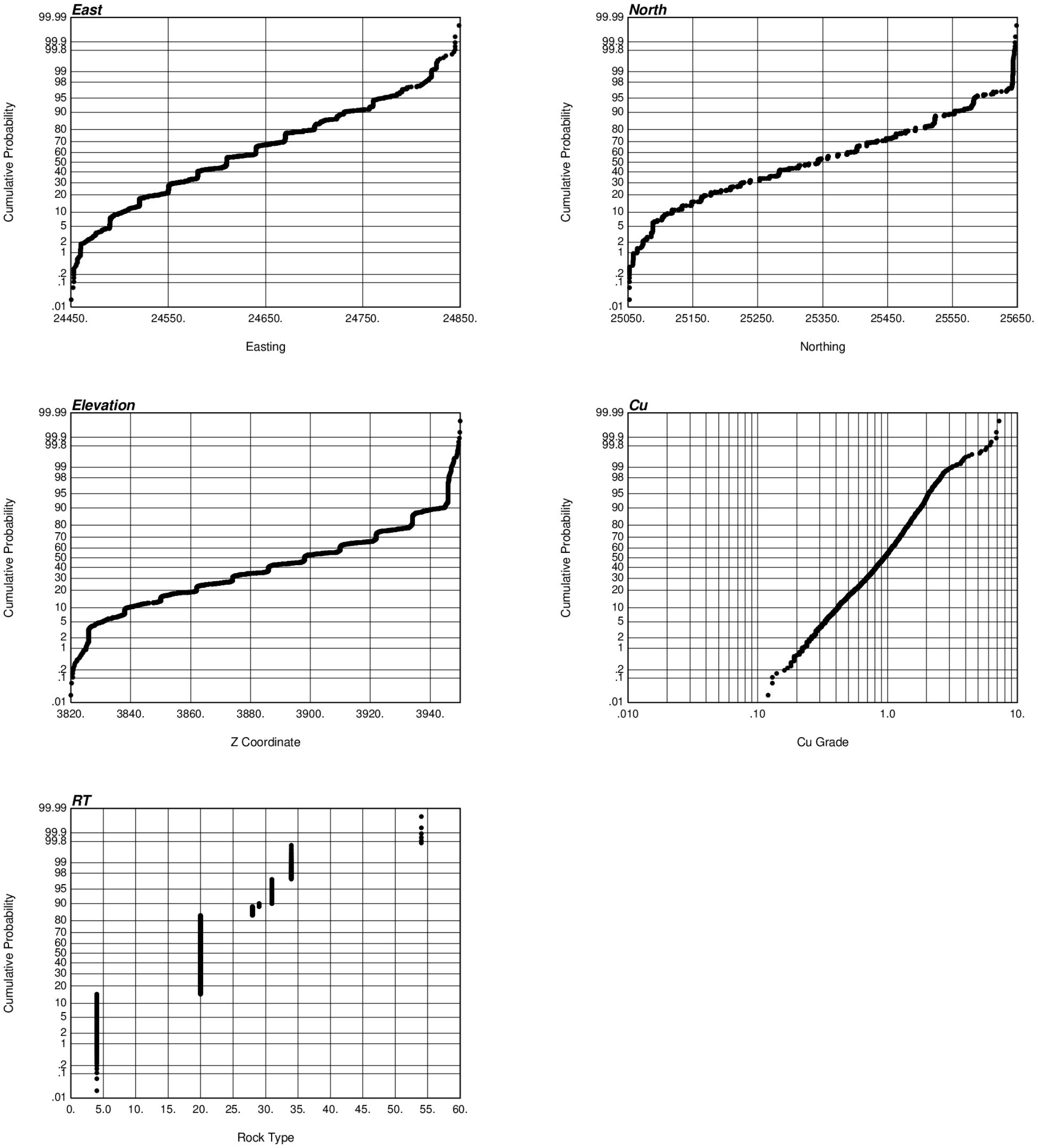
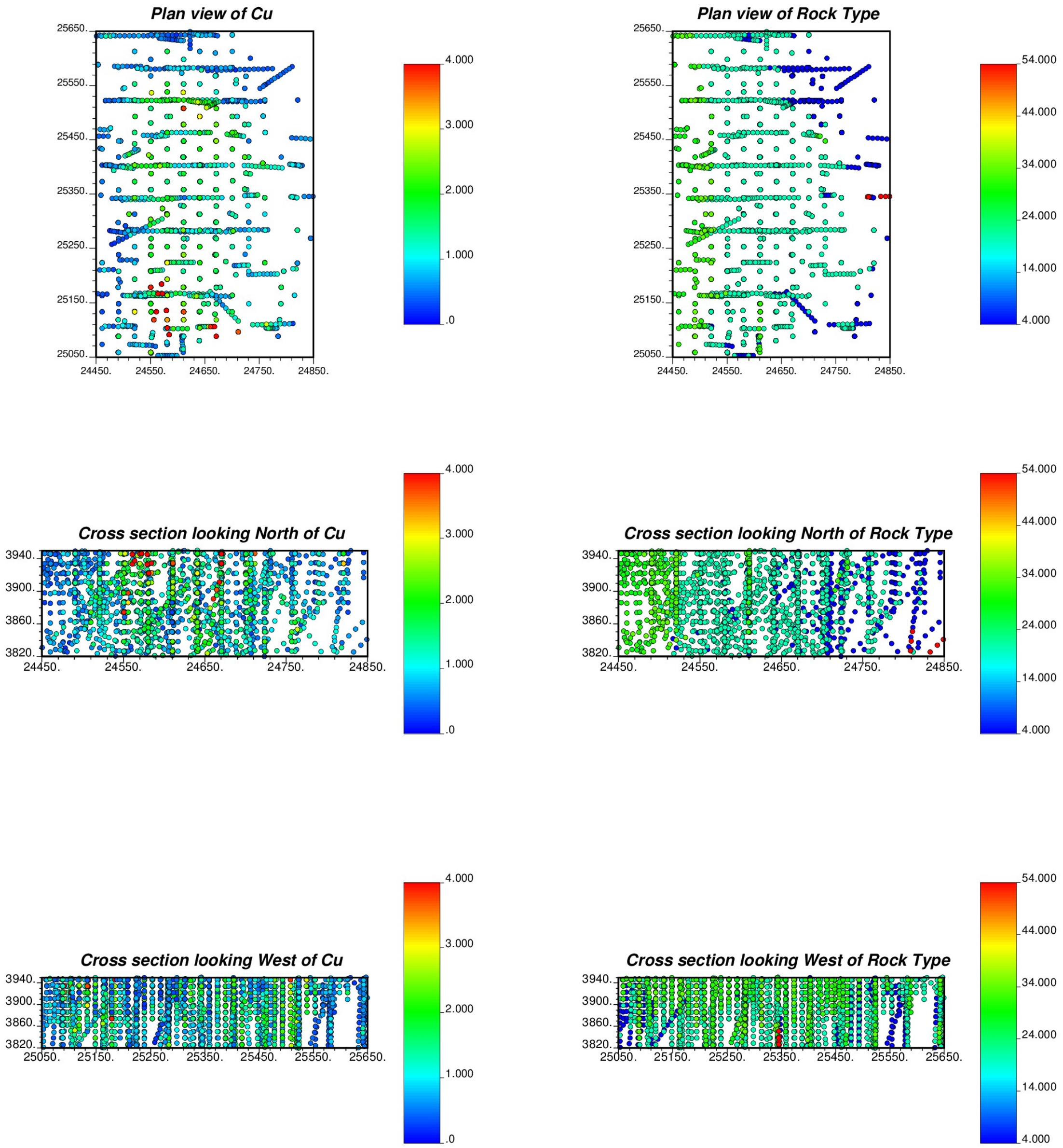


Figure 2: Probability plots of all the variables in the drillhole dataset


Plan view of Cu

Plan view of Rock Type

Cross section looking North of Cu

Cross section looking North of Rock Type

Cross section looking West of Cu

Cross section looking West of Rock Type

Figure 3: Location maps projecting the data into the three main planes

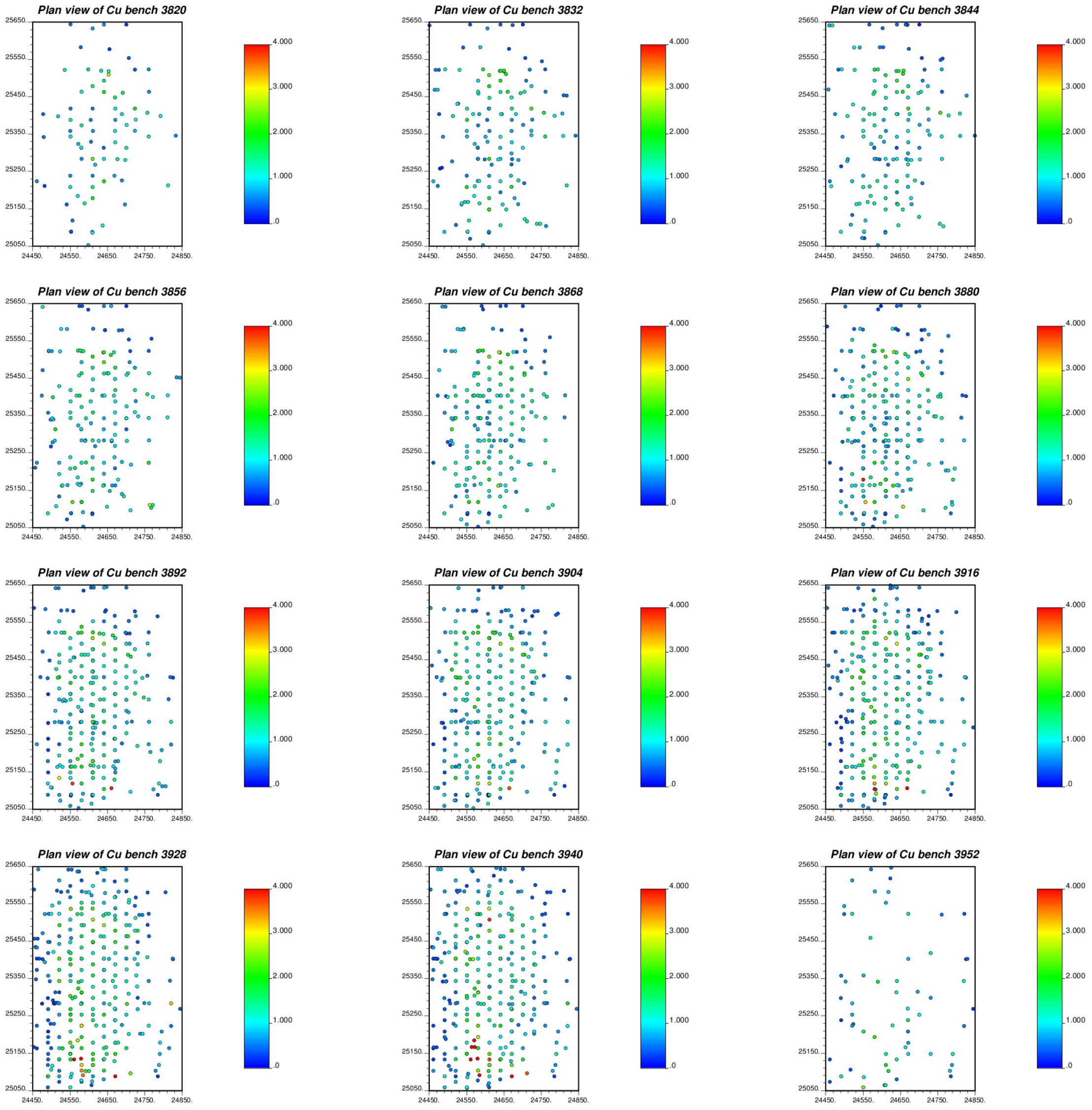


Figure 4: Plan views of copper grades by bench. Every slice is 12m

rock type. A total of 7 categories exist. Some of them have very few samples. Grade distributions are fairly consistent in terms of the shape. Units 28, 29 and 54 have the lowest grades. Unit 20 has the highest average grade, followed by units 31 and 34.

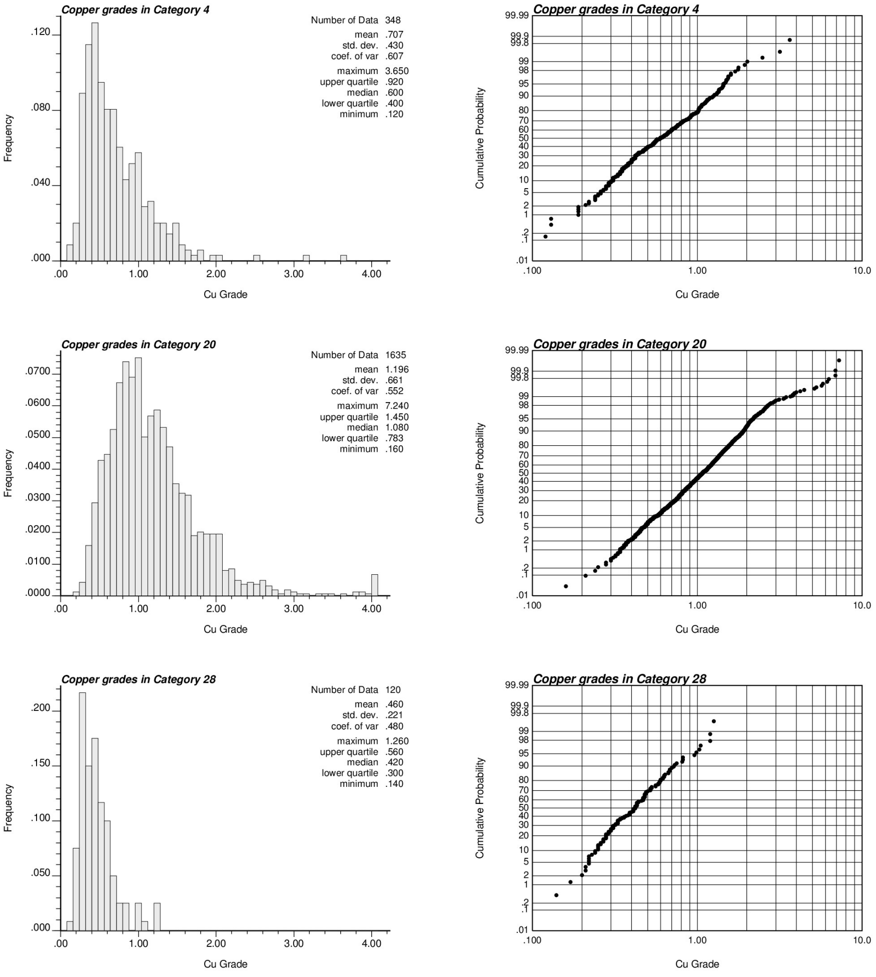


Figure 5: Histograms and log probability plots for copper grade per rock type (categories 4, 20, and 28)

Many more analyses can (and should) be done to fully understand the data. This example shows only some of the basic plots required to start gaining an understanding of the data, and their statistical and spatial distribution.

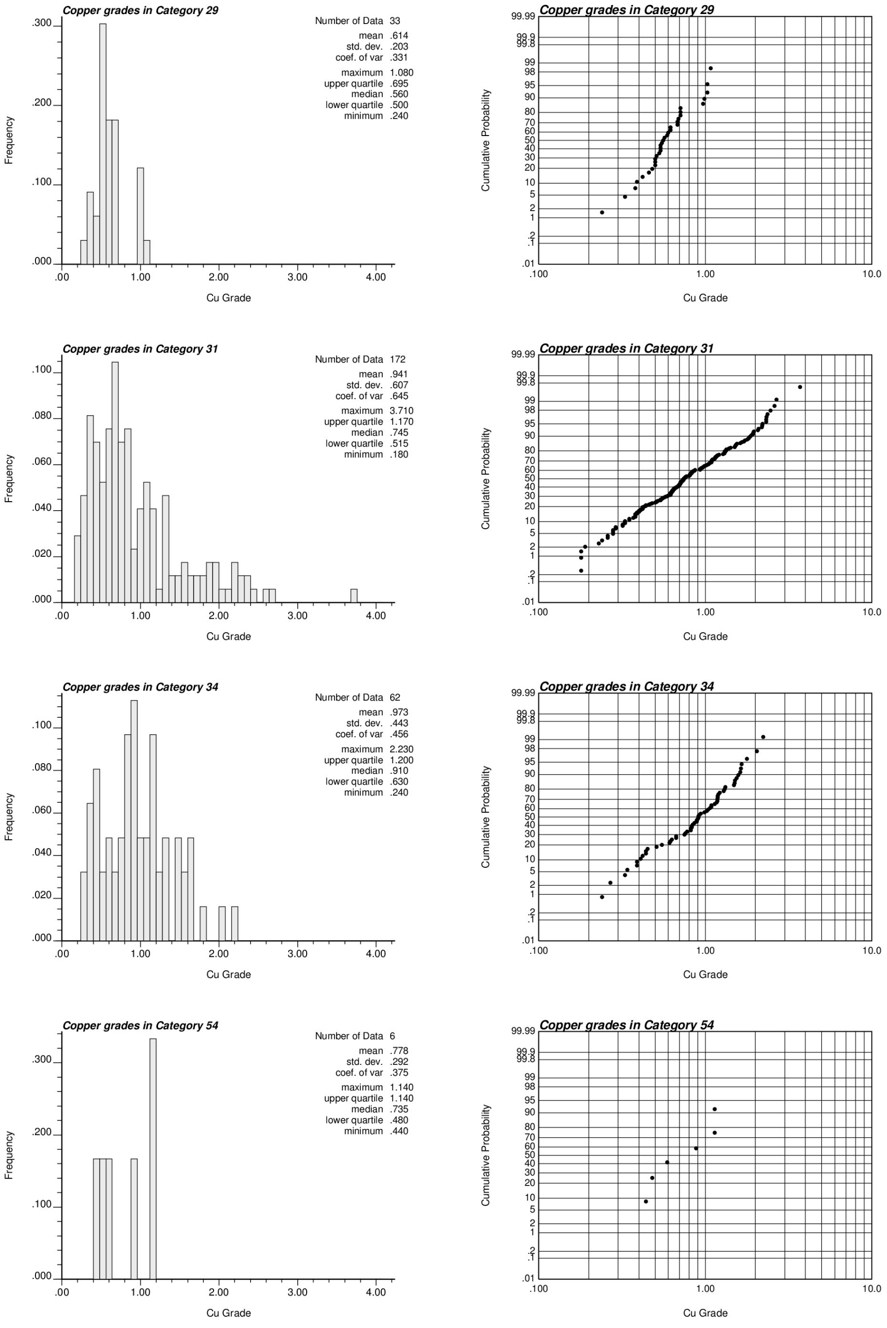


Figure 6: Histograms and log probability plots for copper grade per rock type (categories 29, 31, 34 and 54)