STATIONARITY

This is another fundamental assumption, presuming invariance through space of properties of the random function.

The most general case assumes that the joint probability distribution of any number of random variables is the same.

Most geostatistical formulations require independence of location only for moments and just up to order two.

BIAS

Bias in sampling denotes preference in taking the measurements.

Bias in an estimator implies that the calculations are preferentially loaded relative to the true value, either systematically too high or too low.

- When undetected or not compensated, bias induces erroneous results.
- The opposite of the quality of being biased is to be unbiased.

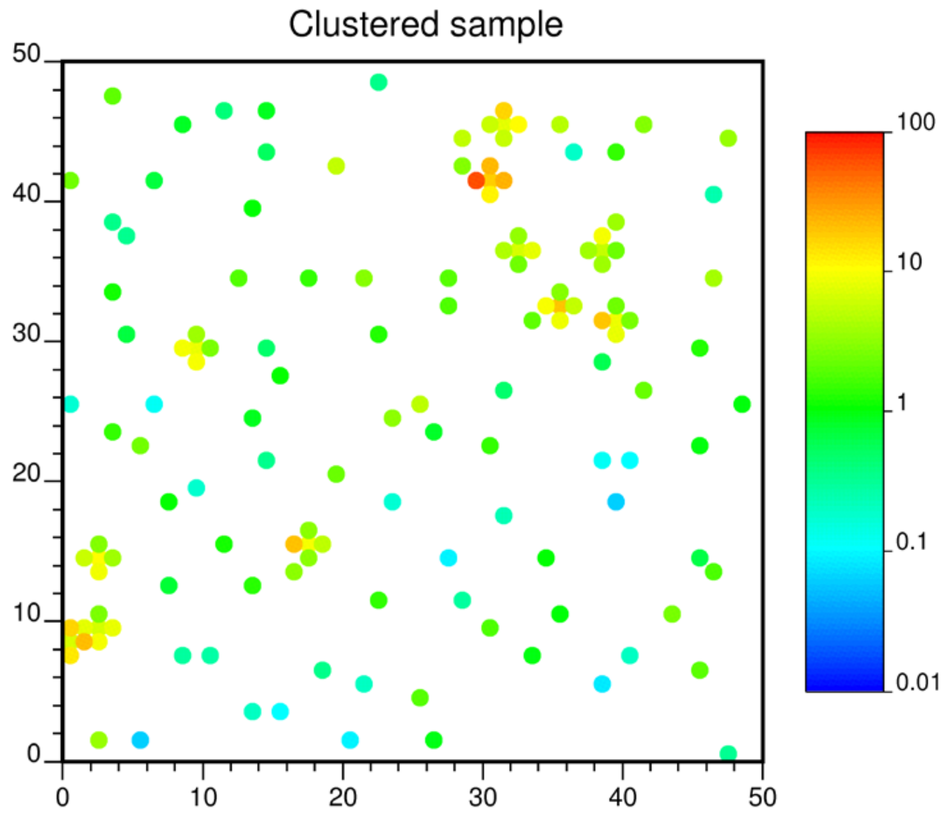
PREFERENTIAL SAMPLING

Although estimation and simulation methods are robust to clustered preferential sampling, parameters that need to be inferred from the same sample prior to estimation or simulation can be seriously distorted, especially for small samples.

The solution to preferential sampling is preparation of a compensated sample to eliminate the clustering, for which there are several methods.

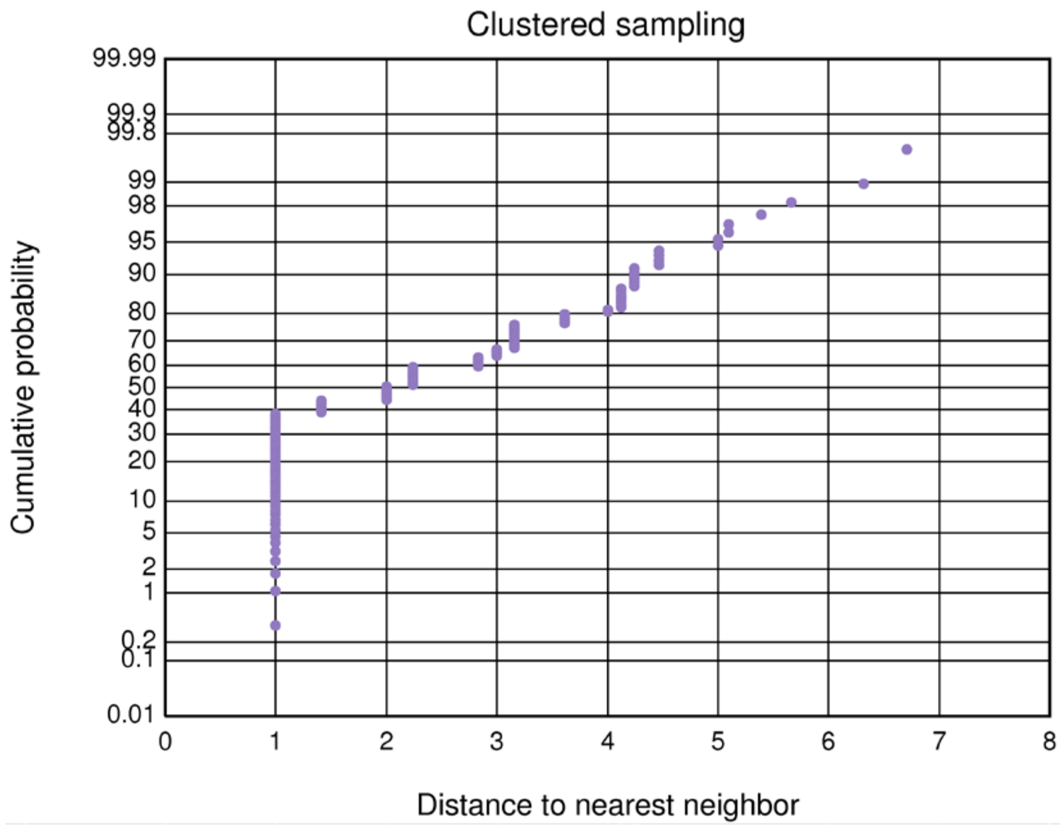
Declustering is important for the inference of global parameters, such as any of those associated with the histogram.

EXAMPLE OF SAMPLE WITH PREFERENTIAL CLUSTERING



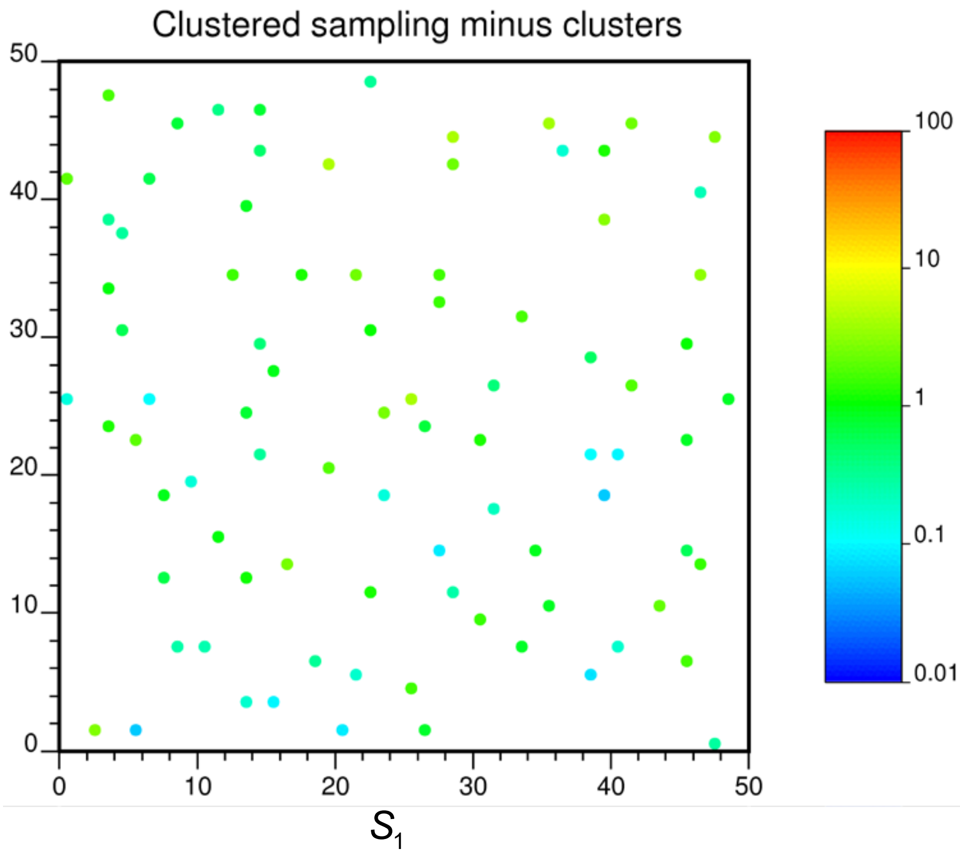
USGS

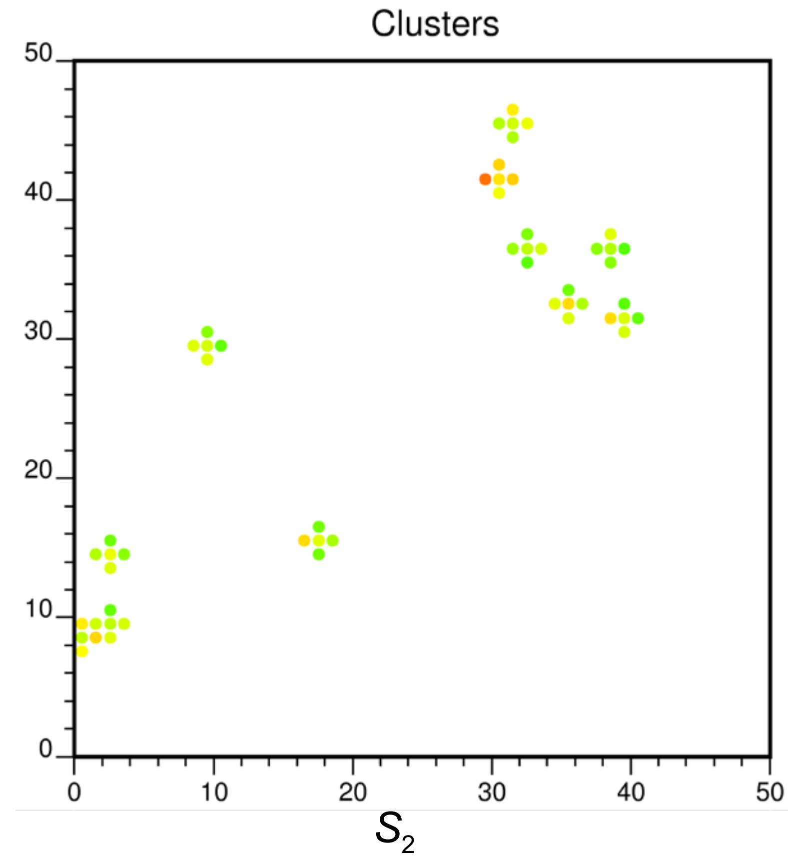
DECLUSTERING (1)



Detect presence of clusters by preparing a cumulative distribution of distance to nearest neighbor.

DECLUSTERING (2)

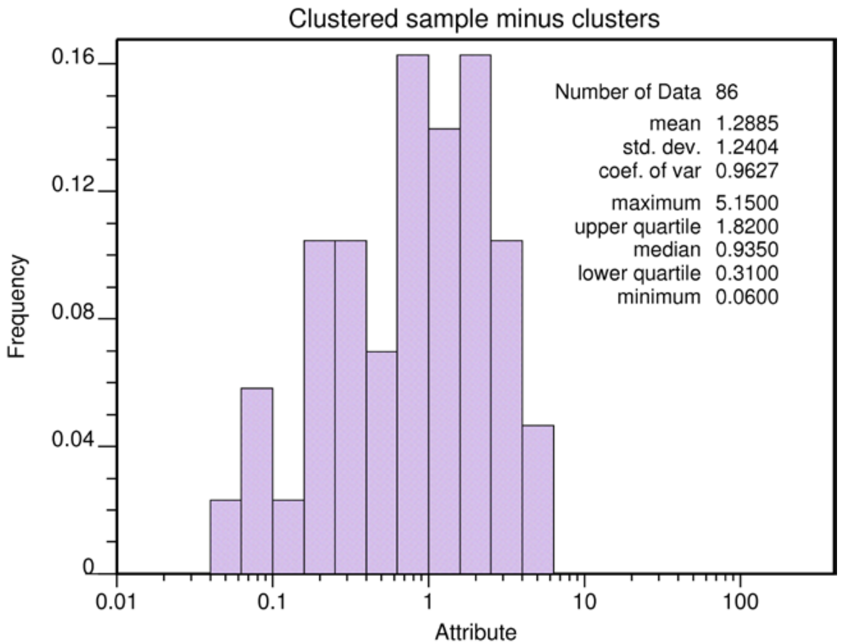


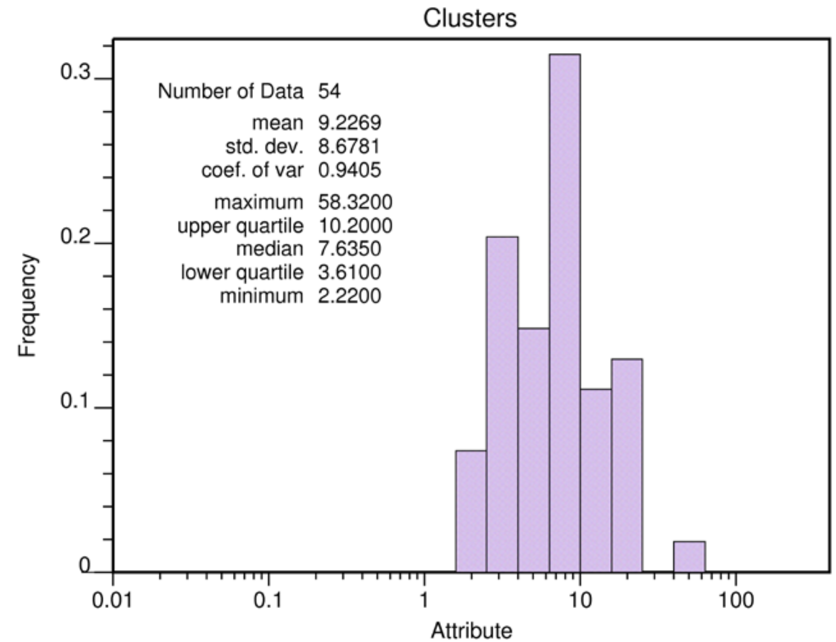


Decompose the clustering.

DECLUSTERING (3)

Histograms of the attribute by distance class

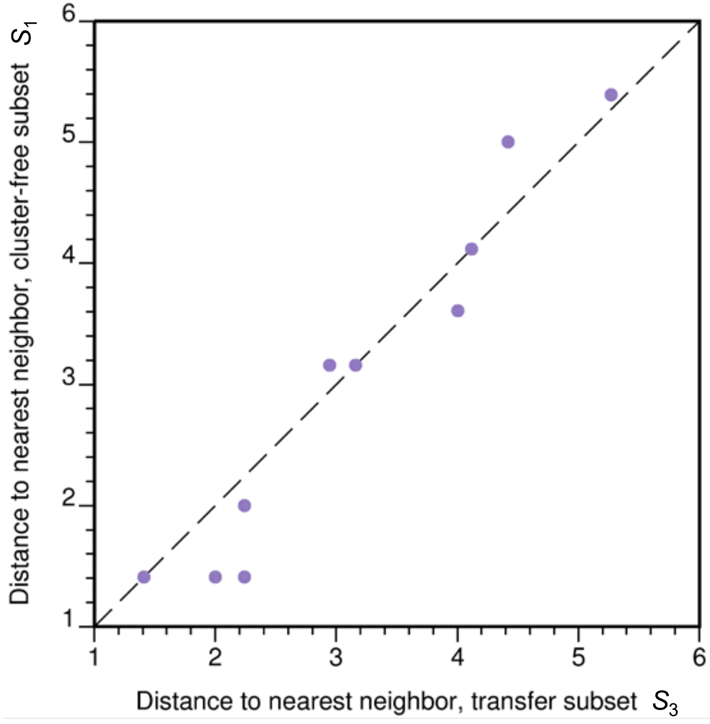




Preferential sampling shows as poor overlapping between the two histograms.

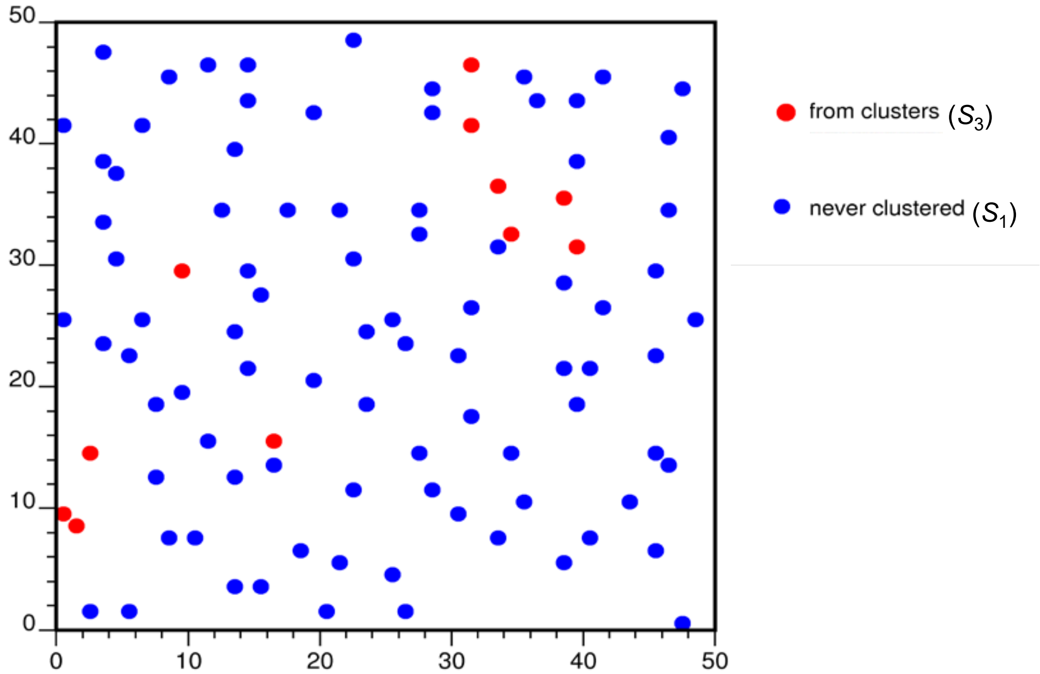
DECLUSTERING (4)

One possibility is to obtain the declustered subset (S_4) by expanding the subset without clusters (S_1) by transferring a few observations (S_3) from the clustered subset (S_2).



- Transfer observations from (S_2) by decreasing distance to nearest neighbor in S_4.
- Stop transferring points when the distribution of distances for the original subset S_1 is about the same as that for the transferred observations (S_3) within S_4.

POSTING OF A DECLUSTERED SAMPLE (S_4)



Best results are obtained by randomly repeating the drawing from S_2 to generate multiple declustered sets. The final answer is obtained by taking median values from the cumulative distributions of all declustered sets.

USGS

6. TRANSFORMATIONS

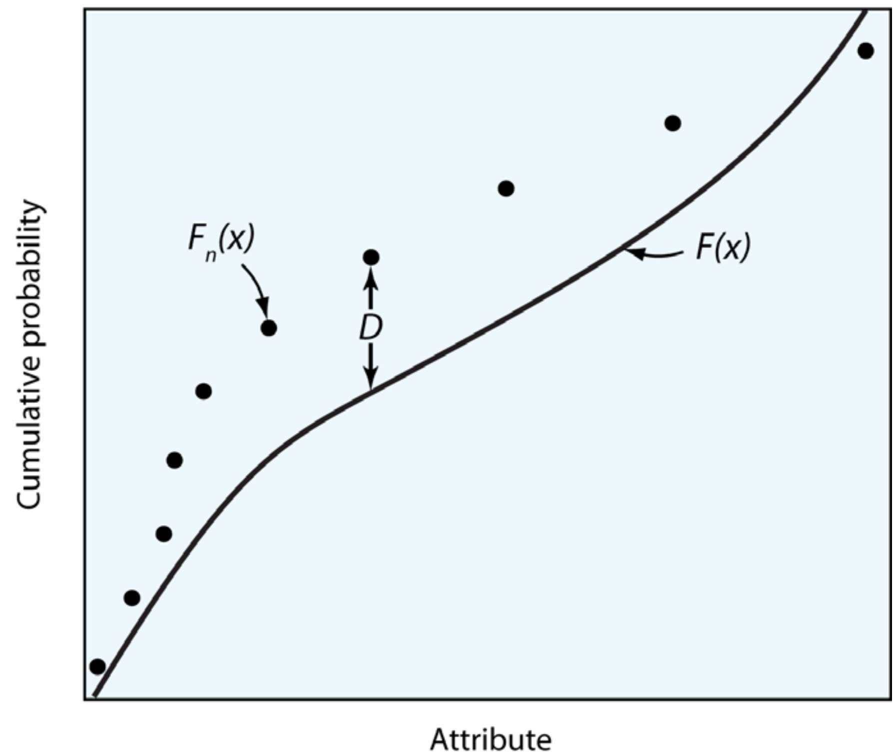
COMPARING DISTRIBUTIONS

The distribution followed by some data and the one that may be required by a method may be two different distributions.

If a method requires a particular distribution, such distribution becomes mandatory.

The simplest way to compare a sample, $F_n(x)$, to an hypothesized distribution, $F(x)$, is to superimpose their cumulative distributions.

By transforming the data, one can force the data to follow almost any other distribution.



KOLMOGOROV-SMIRNOV TEST

The maximum discrepancy, *D*, is the measure of similarity in the Kolmogorov-Smirnov test.

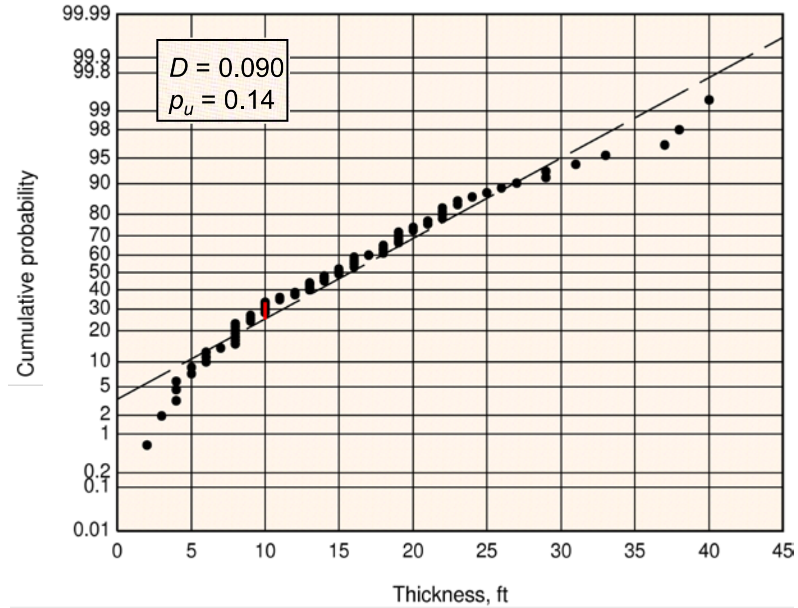
The significance of the discrepancy depends on the sample size and on whether the parameters of the hypothesized distribution are known. A statistic called *p*-value, commonly read from tables, is used to decide the significance of the discrepancy. Customarily, if the *p*-value is above 0.1 or 0.05, the equality of the distributions is accepted.

Standard tables available to run the significance of the discrepancy are not valid in geostatistics because the tables do not take into account spatial dependence.

It has been found, however, that the standard tables may be a limiting case for the situation considering spatial correlation. In that case, each time two distributions pass the Kolmogorov-Smirnov test without considering spatial correlation, they also pass it considering that correlation. The same is not true in case of rejection.

USGS

WEST LYONS FIELD, KANSAS



Remember that the cumulative probability of any normal distribution follows a straight line in normal probability scale.

At 10 ft, the thickness distribution of Lyons field reaches a maximum discrepancy of 0.09 against a normal distribution with the same mean and standard deviation as the sample. The *p*-value without considering spatial correlation, p_u, is 0.14.

Because the p_u-value is already above the threshold of 0.05–0.1, it is not necessary to go into the complications of considering spatial correlation. Without further ado, it can be accepted that the thickness is normally distributed.

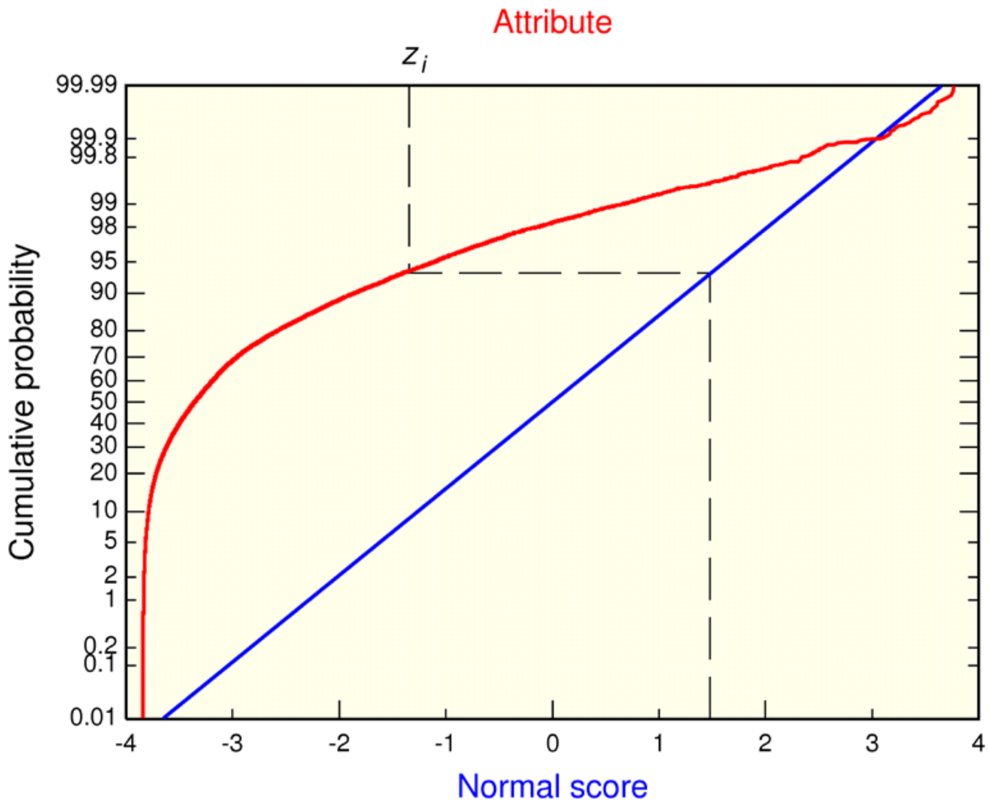
NORMAL SCORES

In geostatistics, as in classical statistics, the most versatile distribution is the normal distribution.

If the sample strongly deviates from normality, it may be convenient or required to transform the data to follow a normal distribution, so as to properly apply a specific method.

The transformation is fairly straightforward. It is performed by assigning to the cumulative probability of every observation the value of the standard normal distribution for the same cumulative probability.

NORMAL SCORES



For the value z_i of the attribute, the normal score is 1.45.

The transformation also can be used backwards to bring calculated values in the normal score space to the original attribute space.

LOGNORMAL TRANSFORMATION

- This is a special case of normal score transformation. The transformed value is simply the logarithm of the original observation.
- If the original distribution is exactly lognormal, the distribution of the transformed values will be exactly normal. Because this is never the case, the distribution for a transformed variable is only approximately normal.
- Most minerals and chemical elements have distributions of concentration close to lognormal, which explains the great attention paid to this distribution since the early days of geostatistics.

INDICATORS

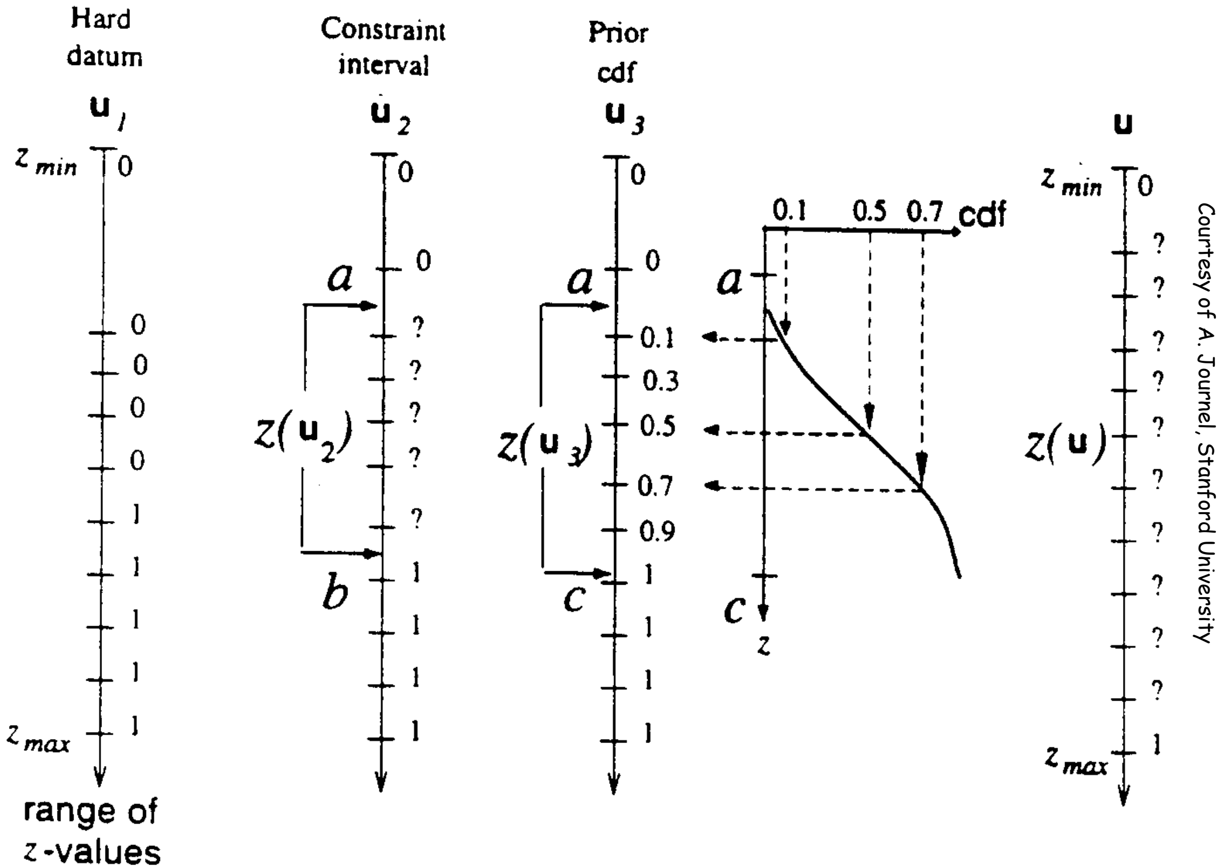
Given a continuous random function $Z(\mathbf{s})$, where **s** denotes geographical location, its indicator $I(\mathbf{s},u)$ is the binary transformation:

$$I(\mathbf{s},u)=\begin{cases}0, \text{ if } Z(\mathbf{s})>u\\ 1, \text{ if } Z(\mathbf{s})\leq u\end{cases}$$

The potential of indicators lies in the possibility to handle imprecise data and the fact that the expected value of an indicator is equal to the cumulative probability for the threshold value:

$$\begin{aligned}E\left[I(\mathbf{s},u\mid(n))\right]&=\text{Prob}\left[Z(\mathbf{s})\leq u\mid(n)\right]\\&=F(\mathbf{s},u\mid(n))\end{aligned}$$

INDICATOR CODING FOR DIFFERENT TYPES OF CONTINUOUS DATA



USGS