"Geostatistical estimation applied to highly skewed data", Joint Statistical Meetings,
Dallas, Texas, August 1999

Geostatistical estimation applied to highly skewed data

Dr. Isobel Clark, Geostokos Limited, Alloa, Scotland

It is common knowledge that ordinary geostatistical methods do not deal well with highly skewed sample data. In recent times "distribution free" methods have been advocated to avoid this problem --- the most popular in practice, at this time, being the multi-indicator methods. As with all techniques, these have their strengths and weaknesses. One of the most obvious of the drawbacks to a multi-indicator approach is the necessity to model many semi-variogram graphs and to carry out many simultaneous kriging or co-kriging estimations. Where the distribution of sample values is reasonably simple and stable, it would seem more practical to use the known features of the distribution and associated methodology.

In this paper we consider the simplest non-Normal case --- that of lognormal kriging. If the values within a deposit are known to be stationary and lognormal, then the logarithms of these values should be Normal. If all geostatistical analysis is carried out on the logarithms, there should be no problems with semi-variogram interpretation and modelling, or with kriging (whether simple or ordinary). Universal kriging may also be used if the residuals from the trend in the logarithms of the sample values can be assumed to be Normal. The only potential problem is in the "backtransformation" of the logarithmic estimates to the original sample value scale. There appears to be some disagreement in the general geostatistical literature as to how this backtransform should be carried out.

We illustrate this paper with a case study on a Wits type reef deposit which is a simulation based on a real set of data. Figure 1 shows the histogram of sample values taken at random within an area of the reef. It can be seen that the data is very well behaved when logarithms are taken.

Semi-variogram calculation and modelling was carried out on this data set. Cross validation confirmed the semi-variogram model. A kriging exercise was also undertaken to estimate average values within square panels over the study area. The results of this exercise are:

- an estimate for the average logarithmic value within the panel $\hat{T}^*$;
- a standard error for this estimate σ_{ok};
- the within panel variance for logarithmic variances $\sigma^2_{\theta/A}$ or $\bar{\gamma}(A,A)$.

Figure 2 shows the simplified histograms of original sample values and kriged panel values. Because this is a simulation, we are also able to show the corresponding histogram of the "true" panel values. It should be remembered that, in this context, original sample value means the logarithmic transform of the values. The kriged panel value is the optimal weighted average estimator produced by ordinary kriging on the logarithms. The "true" panel value is the average of the logarithms of all known values (from simulation) within the panels. It can easily be seen that the kriging estimator and the panel values are pretty close — even though the "true" values are produced from about 25 times as much information. All three distributions show the same mean, whilst both "true" panels and kriged estimators show the reduced standard deviation (spread) expected when considering panels rather than 'point' data. The difference in standard deviations between kriged estimate and true value is,

theoretically, calculated from the classic volume/variance effect. The difference in variances must be included in the backtransformation to 'raw' sample values. Figure 3 shows the effect of simply anti-logging the kriging estimate of the logarithmic mean --- in effect, assuming that the variances of estimator and actual value are equal. The variance of the "true" panel values is given by:

$$\sigma_A^2 = C_{tot} - \bar{\gamma}(A, A)$$

where C_{tot} represents the total sill or 'point' variance, σ^2. The variance of the estimators is:

$$\sigma_{T^*}^2 = C_{tot} - \bar{\gamma}(S, S) = C_{tot} - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(S_i, S_j)$$

that is, the total variance less the variance of values amongst the samples in the weighted average estimator. In practice, this can be calculated by using the equivalent expression:

$$\sigma_A^2 - \sigma_{T^*}^2 = \sigma_{ok}^2 - \lambda$$

where λ is the langrangian multiplier produced by the ordinary kriging process. If simple kriging is used, the longer form is more appropriate (obviously). However, adding one-half of this variance difference to the estimator before anti-logging does not result in sensible results. The backtransformed estimates are consistently lower than the "true" panel values. The larger the panel, the larger the difference between backtransform and true average (see Figure 4). The reason for this becomes clear if we simply plot histograms of the average logarithm in the panel and the logarithm of the "true" value, as in Figure 5. It can be seen from this that, when we consider panel averages, it is not only the variance which changes. The logarithmic mean also changes --- in order to preserve the overall average in the 'raw' values. To maintain unbiassedness in our backtransformed estimates, we must incorporate the change in the logarithmic average as well as the change in the variances. This may be expressed as follows:

$$\sigma^2 - \sigma_{T^*}^2 = \sigma_{ok}^2 - 2\lambda + \sigma_{0/A}^2 = \sigma_{ok}^2 - 2\lambda + \bar{\gamma}(A, A)$$

that is, the correct backtransformed estimator for a panel average is the anti-log (exponentiation) of:

$$T^* + \frac{1}{2}\sigma_{ok}^2 - \lambda + \frac{1}{2}\bar{\gamma}(A, A)$$

Figure 6 shows the comparison between kriged estimates backtransformed using this expression and "true" panel values.

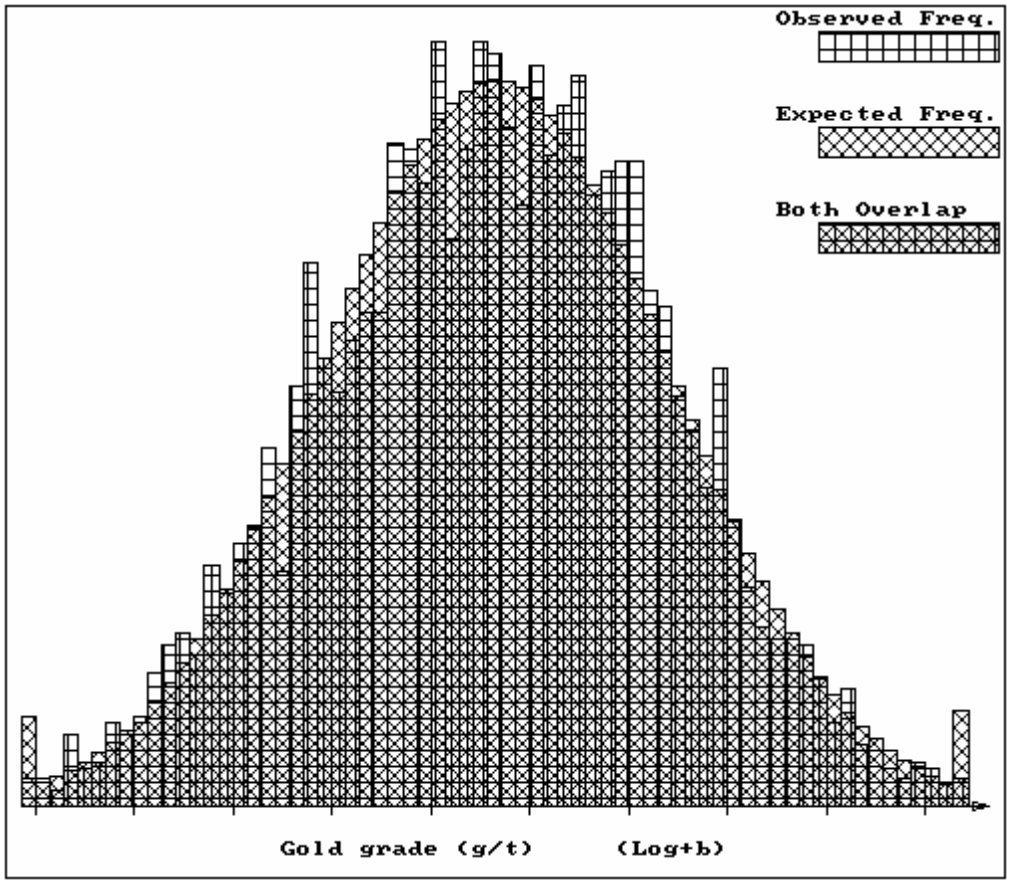

Figure 1: histogram of original sample values (logarithmic scale) and best fit Normal distribution

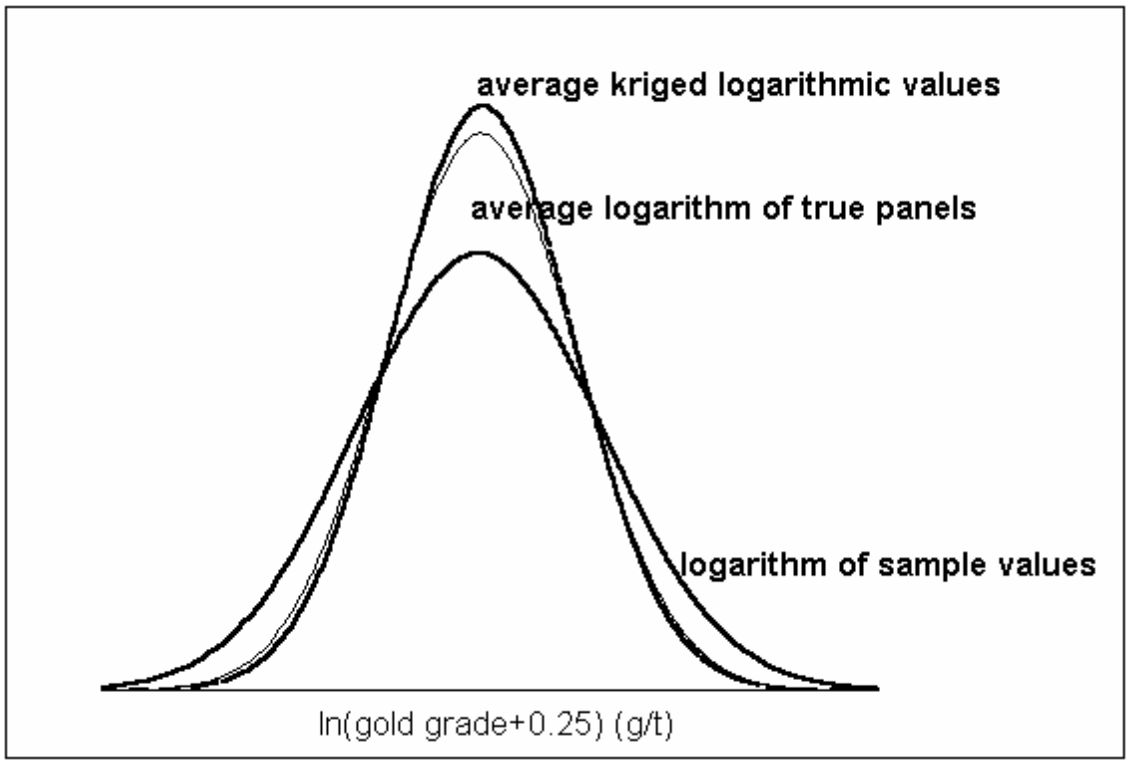

Figure 2: comparison of sample, true panel and kriged Values

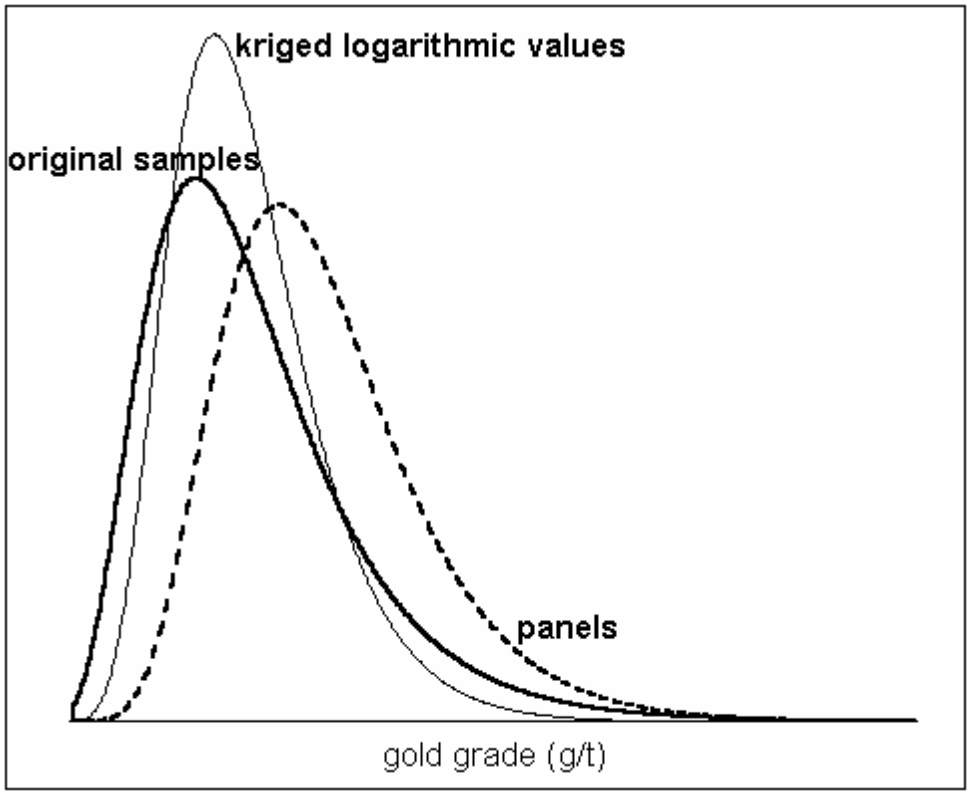

Figure 3: comparison of sample, true panel and anti-logged kriging estimates

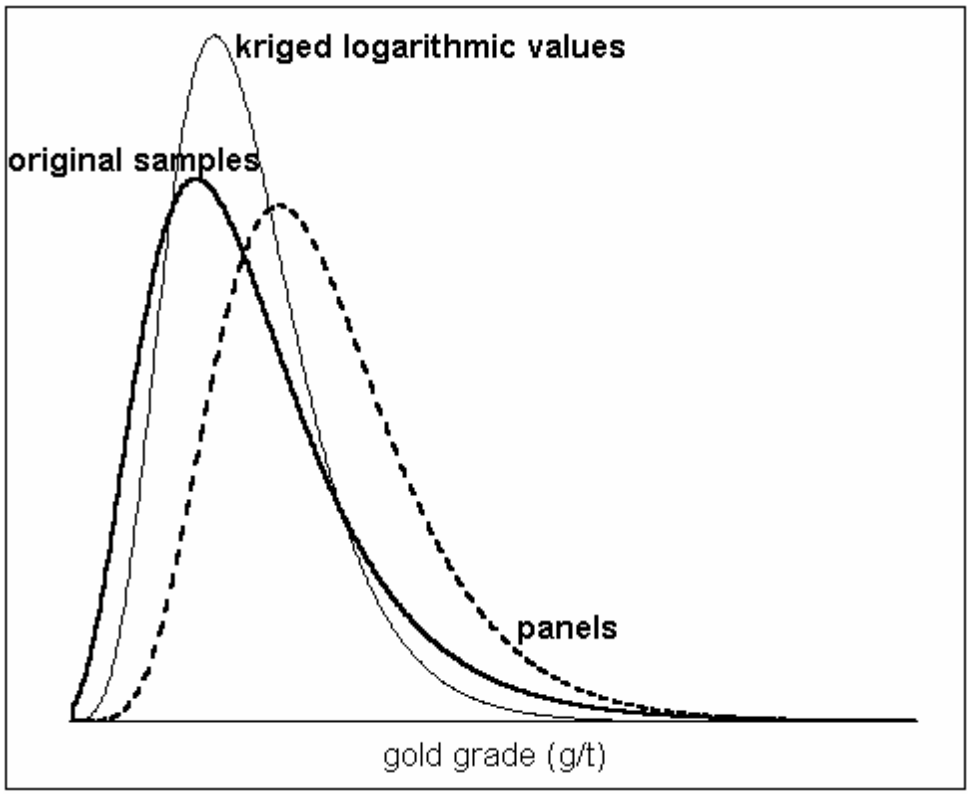

Figure 4: average of true panel values (logged) versus average of logarithms in panel

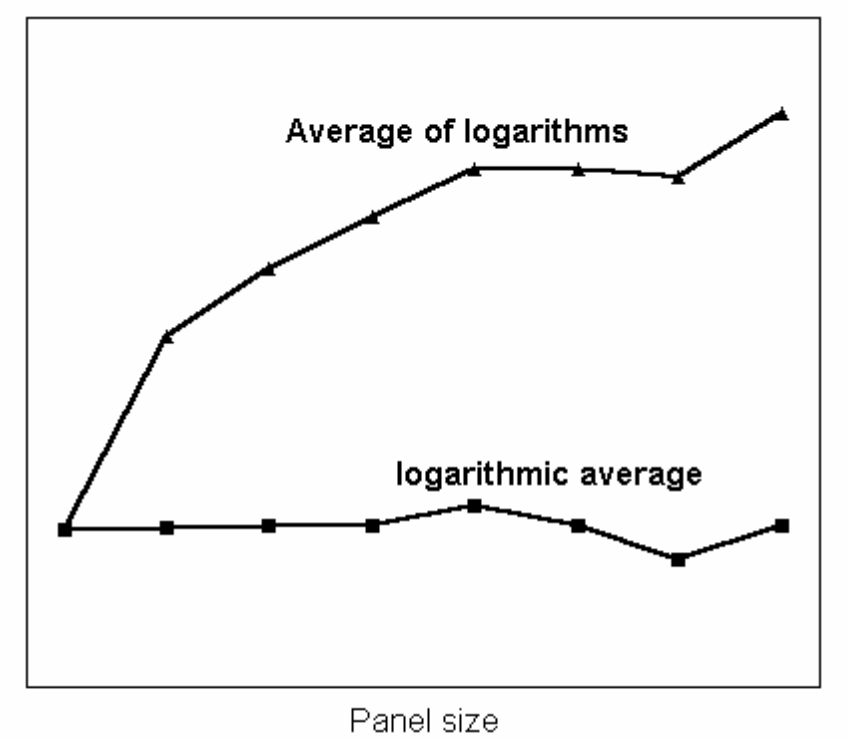

Figure 5: histogram of true panel averages versus histogram of average logarithm in panel

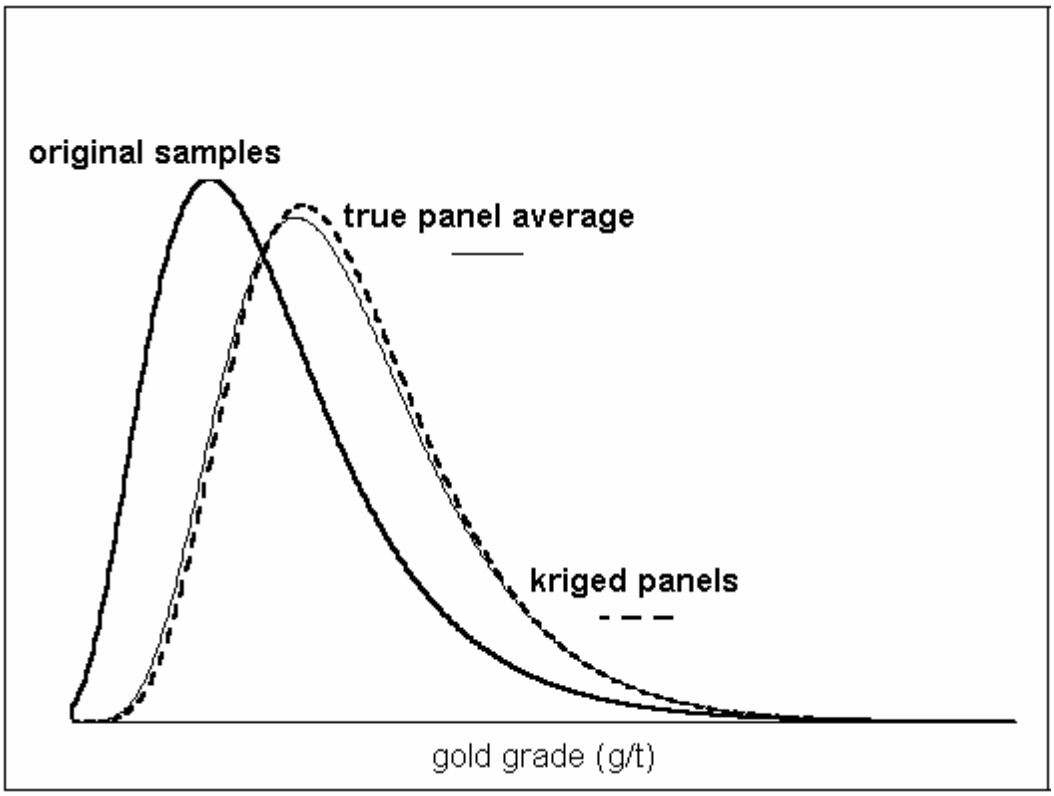

Figure 6: true panel values and backtransformed kriged panel values