Debiasing for Improved Inference of the One Point Statistic

Michael J. Pyrcz[1] and Clayton V. Deutsch[1]

ABSTRACT

Strategic project decisions are based on the distributions global variables, for example, total mineable resource, or recoverable oil volume. These global variables distributions are very sensitive to the input one point statistic, that is, histogram and rock type proportions. Representivity in the one point statistic retains significance in all spatial models.

Spatial sampling bias and nonrepresentative sampling complicate this process of building representative one point statistics. This work outlines the cause of bias, the inability of standard declustering to correct for this bias and two methods for correcting bias in the one point statistics: "trend modeling for debiasing" and "debiasing by qualitative data". An example is presented of each technique based on a poly metallic data set.

INTRODUCTION

Great computational effort is exerted to build realistic geostatistical simulation models. The goodness of these models is judged by their ability to reproduce input one-point statistics and continuity structures. Geostatistical techniques slavishly reproduce input lithofacies proportions and the histogram of petrophysical properties (Gringarten et al., pg. 1, 2000). There is no intrinsic declustering or debiasing within geostatistical simulation algorithms. Geostatistical simulation always weights the input distribution. Gaussian simulation in particular ensures that the input distribution is approximately reproduced. Bias in the input distributions must be removed as a separate step before simulation.

There is a great emphasis on reproduction of the variogram in geostatistics. The continuity modeled by the variogram is important in calculating most response variables. For example, the connectivity of grades has a significant impact on recoverable reserves. Also, the continuity of the grades has a dramatic affect on the homogenization requirements of plant feed.

The importance of representative input distributions must be evaluated with respect to the sensitivity of the response variable to bias in the input statistics. Simulated models are only an intermediate result. Management decisions focus on the results after the application of a transfer function. For example, a numerical model of rock types, mineral grade and density in itself is of little use to management. The key variables such as grade-tonnage curves are found by applying optimized pit design to the constituent numerical models. In this case, the optimized pit design is the transfer function and the grade tonnage curves are response variables. There is a unique response variable for each simulated realization. In most settings the input distribution has a first order affect on the response variables.

Declustering methods, such as cell declustering, are commonly applied. Declustering is ineffective in cases with spatial bias. We believe that debiasing tools such as "trend modeling for debiasing" and "debiasing by qualitative data" should be brought into common practice for the purpose of improving the inference of the one point statistic. We review the sources of spatial bias and the limitations of declustering. Debiasing methodologies are then introduced and an example of each is presented.

SPATIAL SAMPLING BIAS

The term spatial sampling bias refers to the biased selection of sampling locations. For example, this would include the preferential sampling of a subset of the area of interest with high grades. The bias referred to here is not the same as sampling bias in the context of sample handling procedures.

---

[1] University of Alberta, Edmonton, Alberta, CANADA

It is natural that spatial data are collected in a biased manner. Preferential sampling in interesting areas is intentional and facilitated by geologic intuition, analogue data and previous samples. This practice of collecting biased spatial samples is encouraged by technical and economical constraints, such as future production goals, accessibility and the costs of laboratory work.

The cost of uncertainty is not the same everywhere in the area of interest. For example, the cost of uncertainty within a high-grade region is much higher than the cost of uncertainty within clearly waste material. Good delineation and a high level of certainty within the high grade materials allows for accurate reserves estimation and optimum mine planning.

Future production goals may also encourage biased spatial sampling. It is common to start mining in high-grade regions. In this case it is desirable to delineate and characterize the high-grade regions.

Practical issues of accessibility can also cause biased spatial sampling. For example, the drilling depth or available drilling stations may constrain selection. In the presence of a vertical trend, limited depth of drilling may result in a subset of the underlying distribution not being sampled. There are many possible scenarios under which accessibility would be a concern (see Figure 1).

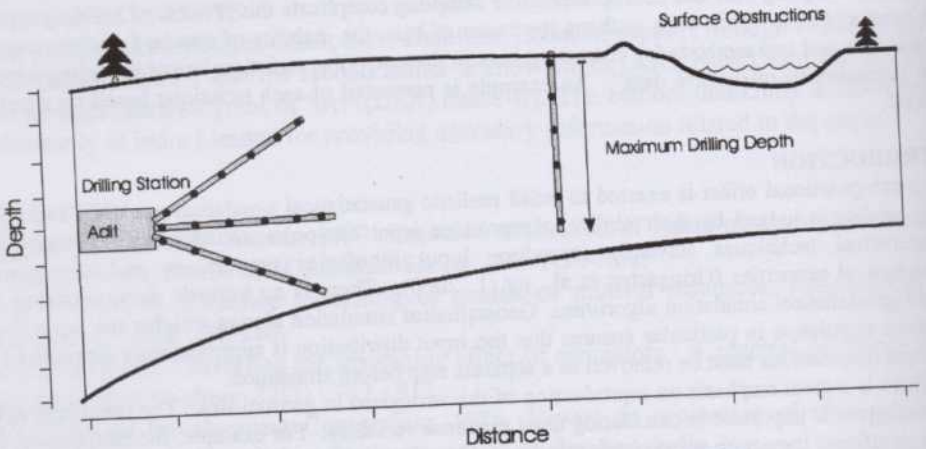

Figure 1 Some examples of accessibility constraints illustrated on a cross section.

Spatial sampling bias may also be introduced at the assaying stage. For example, when removing sections of core for the purpose of permeability measurement, it is unlikely that a section of shale would be subjected to expensive testing. Likewise, barren rock may not be sent for assays.

STATISTICAL ASSUMPTIONS

Conventional statistics do not provide reasonable solutions to the problem of constructing representative distributions. A simple random sample from the population of interest would be unbiased, but inappropriate in most cases. A sample is said to be unbiased when each unit of the population has the same probability of being sampled. In conventional statistics this is accomplished by avoiding preferential sampling or opportunity sampling. As explained above, there are many reasons that geologic samples are collected in a biased manner.

Regular or random stratified sampling may be able to provide a good approximation of a representative distribution. Sampling on a regular grid is rarely practical for the same accessibility and economic reasons stated above. Regular sampling grids may be applied in preliminary resource investigation. Nonsystematic infill drilling often augments these sampling campaigns. One approach would be to omit the clustered infill samples for the purpose of building distributions. While this would more closely agree with conventional statistical theory, throwing away expensive information is not very satisfying (Isaaks and Srivastava, pg. 237-238, 1997).

DECLUSTERING

Declustering is well documented and widely applied (Deutsch, pg. 53-62, 2001; Isaaks and Srivastava, pg. 237 – 248, 1997; Goovaerts, pg. 77-82, 1997). There are various types of declustering methods, such as

cell, polygonal and kriging weight declustering. These methods rely on the weighting of the sample data, in order to account for spatial representivity. Polygonal declustering is shown in Figure 2. Figure 3 shows the weighted histogram. Note that weighting does not change the values: only the influence of each sample is changed.

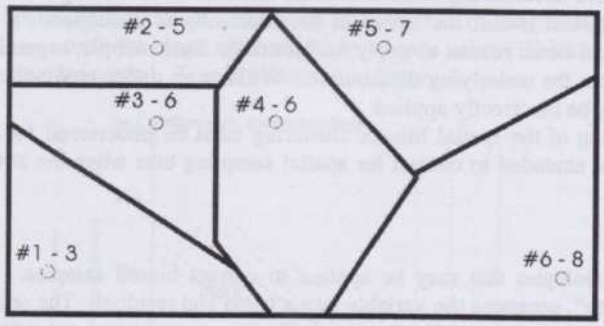

Figure 2 Sample data with the associated voronoi polygons, the declustering weight for each sample is the area normalized by the sum of all the areas.

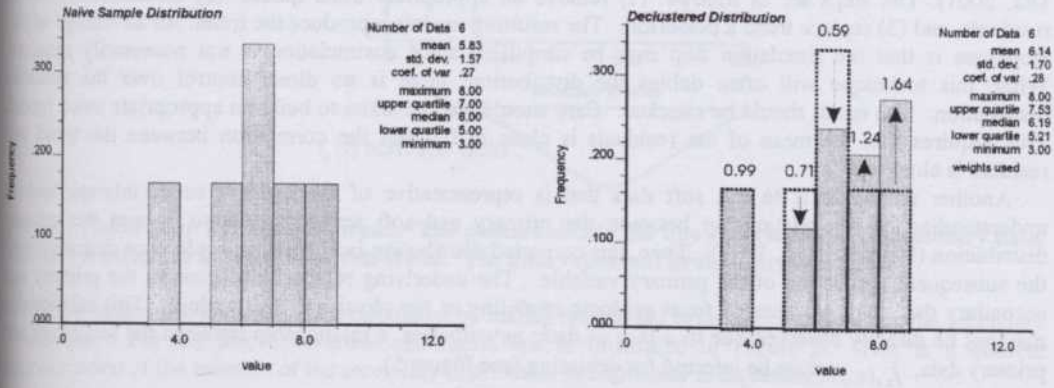

Figure 3 The influence of weighting on a distribution. On the right the naïve distribution (dotted line) is superimposed on the declustered distribution with the weights indicated.

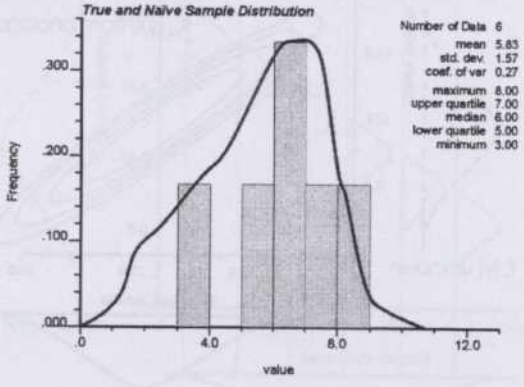

Figure 4 An example underlying distribution (bold line) and the sample distribution (histogram). The entire range of the true distribution has not been sampled.

There are two important assumptions in all declustering techniques: (1) the entire range of the true distribution has been sampled, and (2) the nature of the clustering is understood. Declustering may not perform well without these assumptions. The first assumption is required since the weighting only adjusts the influence of each sample on the distribution and does not change the actual sample value. Figure 4 shows an example where declustering could not work; there are no low samples to give more weight to.

The second assumption is that the nature of the clustering is understood. If the data have no spatial correlation, there would be no reason to apply declustering. Each sample, regardless of location, would be a random drawing from the underlying distribution. Without an understanding of the spatial nature of the data, declustering may be incorrectly applied.

Some understanding of the spatial bias or clustering must be understood for any numerical modeling. Special effort must be extended to correct for spatial sampling bias when the entire range of variability is not sampled.

DEBIASING

There are two methodologies that may be applied to correct biased samples. The first method, "trend modeling for debiasing", separates the variable into a trend and residual. The second approach, "debiasing by qualitative data", corrects the distribution with a representative secondary data distribution and calibration relationship to the primary variable of interest.

In the presence of a clear and persistent trend, trend modeling may be applied to ensure that the correct distribution is reproduced. Trend modeling is well established (Goovaerts, pg. 126, 1997; Deutsch, pg. 182, 2001). The steps are as follows: (1) remove an appropriate trend model, (2) stochastically model residuals, and (3) replace trend a posteriori. The resulting models reproduce the trend. An advantage of this technique is that the simulation step may be simplified since cosimulation is not necessarily required. While this technique will often debias the distribution, there is no direct control over the resulting distribution. The result should be checked. Care should also be taken to build an appropriate trend model. This requires that the mean of the residuals is close to 0.0 and the correlation between the trend and residual is close to 0.0.

Another technique is to use soft data that is representative of the entire area of interest, and an understanding of the relationship between the primary and soft secondary datato correct the primary distribution (Deutsch et al., 1999). Then, this corrected distribution is applied as a reference distribution to the subsequent simulation of the primary variable. The underlying relationship between the primary and secondary data may be assessed from geologic modeling or the physics of the setting. This relationship may not be directly observed due to a lack of data; nevertheless, a relationship between the secondary and primary data, $\hat{f}_{x,y}(x,y)$ must be inferred for debiasing (see Figure 5).

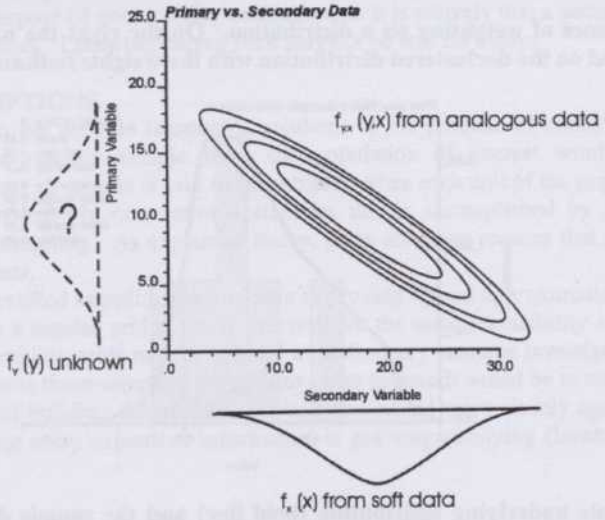

Figure 5 The calibration bivariate distribution, $\hat{f}_{x,y}(x,y)$, and known marginal distribution of the soft data variable, $f_x(x)$.

The construction of the bivariate calibration is the difficult component of debiasing. There are a variety of techniques for building this distribution. For example, the program SDDECLUS by Deutsch relies on the user submitting data pairs which describe the bivariate relationship. This approach allows for the greatest flexibility, since there is no constraint on the form of the bivariate calibration. For each paired primary data a weight is assigned based on the secondary distribution

Another method is to calculate a series of conditional distributions of the primary given the secondary secondary data, $\mathbf{f_{primary|secondary}}$, over the range of observed secondary value. This can be extrapolated over the range of all secondary data by a trend. This is illustrated in Figure 7.

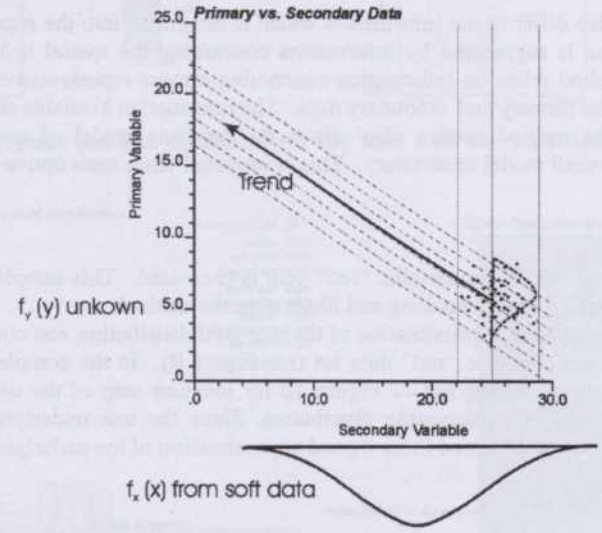

Figure 7 Calibration by bivariate trend. The points indicate the known primary and secondary data. The arrow indicates a linear bivariate trend. The lines represent probability contours.

The primary distribution is then calculated by scaling the binned bivariate calibration by the secondary distribution. For the above bivariate calibration this is illustrated in Figure 8. This is a discrete approximation of the solution of the secondary distribution as expressed in Equation 1.

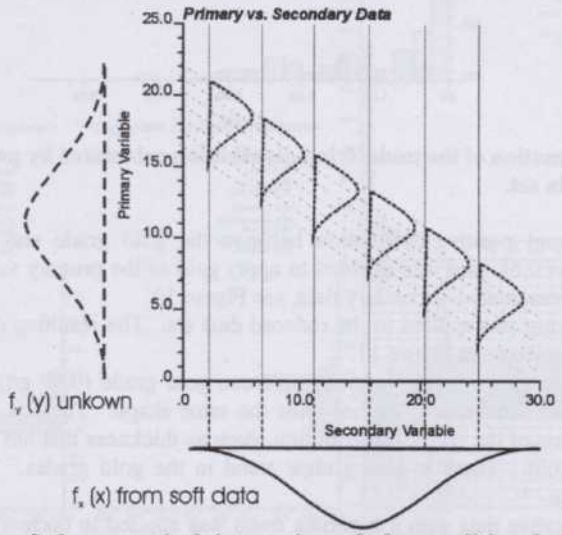

Figure 8 An illustration of the numerical integration of the conditional distribution along the previously indicated linear bivariate trend.

$$f_y(y) = \int_x f_{y|x}(y \mid x) \cdot f_x(x) dx \quad (1)$$

The trend method indirectly corrects the global distribution. This leads to models with precise trend reproduction and indirect control over the distribution. The qualitative method focuses on directly correcting the global distribution and retaining consistency by applying the secondary data as collocated data in the simulation. The result is direct control over the reproduced distribution and indirect control over trend reproduction.

The two techniques, also differ in the information which is integrated into the numerical model. In the first method the simulation is augmented by information concerning the spatial behavior of the primary variable. The second method relies on information concerning a more representative secondary data and the relationship between the primary and secondary data. The information available may limit the ability to apply either method. The method chosen also affects the resulting model of uncertainty. Each will potentially decrease the overall model uncertainty. This is expected since each option involves information to the numerical model.

EXAMPLE

A realistic data set based on a 2D poly metallic "red" vein is presented. This sample set was gathered by drilling. Some data were removed for checking and illustrating the method.

For the sake of comparison, an approximation of the true gold distribution was constructed by applying polygonal declustering to the complete "red" data set (see Figure 9). In the complete data set the entire area of interest is well delineated (see left of Figure 10 for location map of the complete data set) and polygonal declustering results in a reasonable distribution. Since the true underlying distribution is not available, this distribution will be assumed to be a good approximation of the underlying distribution.

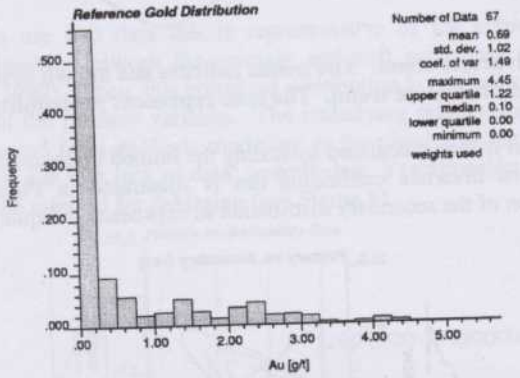

Figure 9 – An approximation of the underlying distribution calculated by polygonal declustering of the complete "red" data set.

There is a significant positive correlation between the gold grade and the thickness of the vein (correlation coefficient = 0.6), so it was decided to apply gold as the primary variable and a smooth kriged thickness map as the representative secondary data, see Figure 10.

Polygonal declustering was applied to the reduced data set. The resulting declustered distribution and the voronoi polygons are shown in Figure 11.

There is a great difference between the underlying mean gold grade (0.69 g/t) and the declustered mean gold grade (1.25 g/t) and distributions do not have the same shape. There is additional information that could aid in the inference of the correct distribution, such as thickness that has a significant correlation to the primary variable, gold. There is also a clear trend in the gold grades. This analogue information improves the distribution.

Debiasing by qualitative data with a bivariate trend was applied to correct the gold distribution. The results are shown in Figure 12.

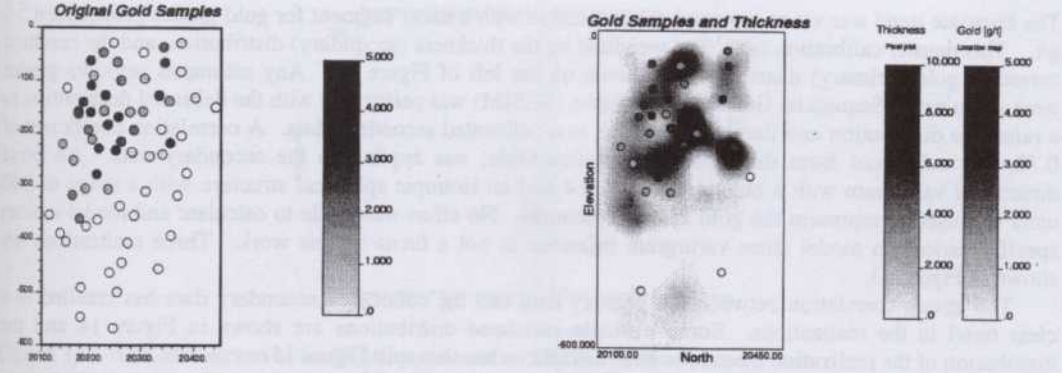

Figure 10 – The original red.dat database (on the left) and the modified data base with kriged thickness map.

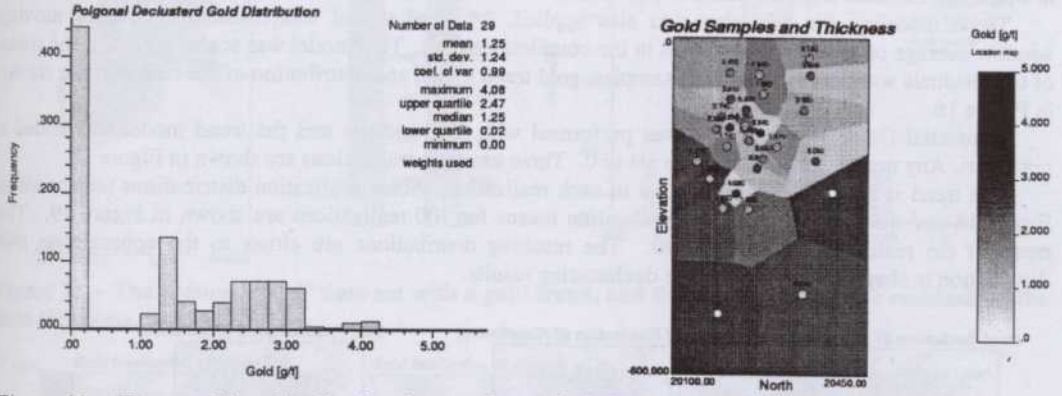

Figure 11 – The resulting distribution from polygonal declustering of the modified red.dat data set and a location map of the data set, with the associated voursior polygons.

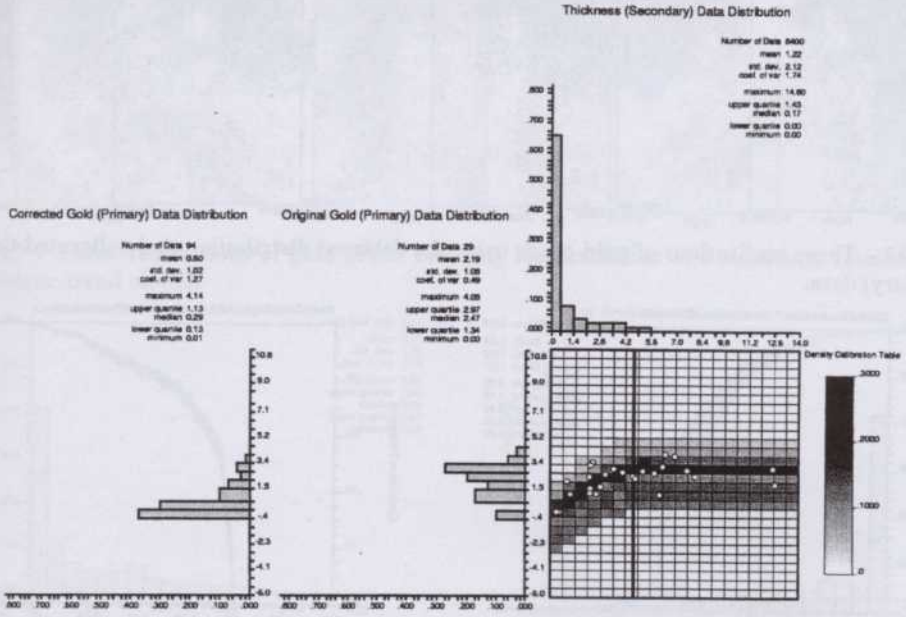

Figure 12 – The density calibration table with collocated thickness data, the thickness distribution, original gold distribution and the corrected gold distribution

The bivariate trend was set as a second order function with a linear segment for gold grades greater than 5.0 g/t. This density calibration table was weighted by the thickness (secondary) distribution, and the resulting corrected gold (primary) distribution is shown on the left of Figure 12. Any estimated negative grades were set to zero. Sequential Gaussian simulation (SGSIM) was performed with the debiased distribution as a reference distribution and the thickness map as a collocated secondary data. A correlation coefficient of 0.72, was calculated from the density calibration table, was applied to the secondary data. An omni directional variogram with a nugget effect of 0.4 and an isotropic spherical structure with a range of 140 units was used to represent the gold spatial continuity. No effort was made to calculate and model a more specific variogram model since variogram inference is not a focus of this work. Three realizations are shown in Figure 13.

The strong correlation between the primary data and the collocated secondary data has resulted in a clear trend in the realizations. Some example simulated distributions are shown in Figure 14 and the distribution of the realization means for 100 realizations are shown in Figure 15.

The average of the realization means is 0.84, which is higher than the average of the reference distribution (see Figure 9). Nevertheless, the resulting distribution is closer to the reference true distribution in shape and statistics than the declustering results.

Trend modeling for debiasing was also applied. A trend model was constructed from a moving window average of all the gold samples in the complete data set. This model was scaled such that the mean of the residuals was near 0. The gold samples, gold trend model and distribution of the residuals are shown in Figure 16.

Sequential Gaussian simulation was performed with the residuals and the trend model was added a posteriori. Any negative estimates were set to 0. Three example realizations are shown in Figure 17.

The trend is consistently reproduced in each realization. Some realization distributions are shown in Figure 18 and the distribution of the realization means for 100 realizations are shown in Figure 19. The mean of the realization means is 0.90. The resulting distributions are closer to the approximate true distribution in shape and mean than the declustering results.

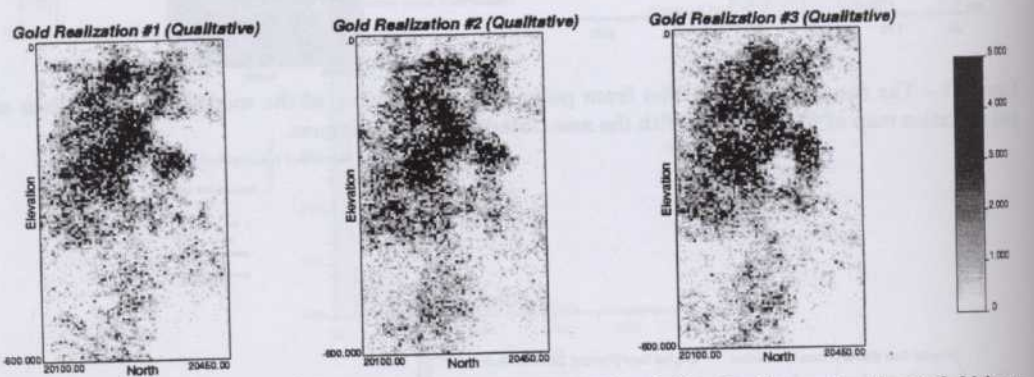

Figure 13 – Three realizations of gold grade using the debiased distribution and collocated thickness (secondary) data.

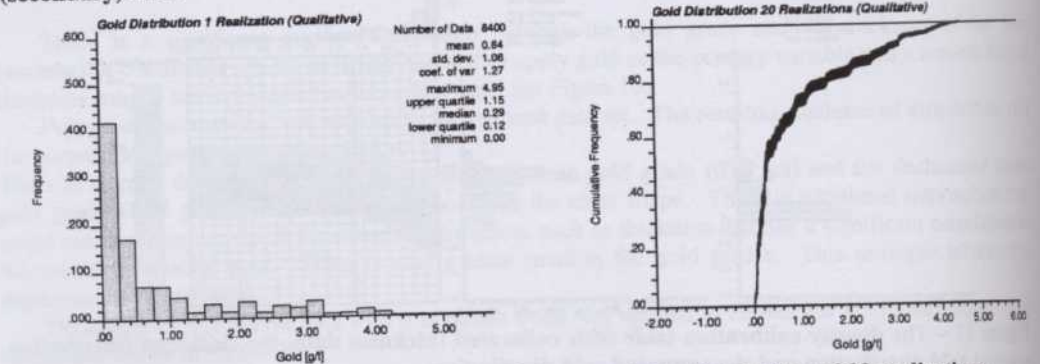

Figure 14 – The histogram of one realization and the cumulative distribution of 20 realizations.

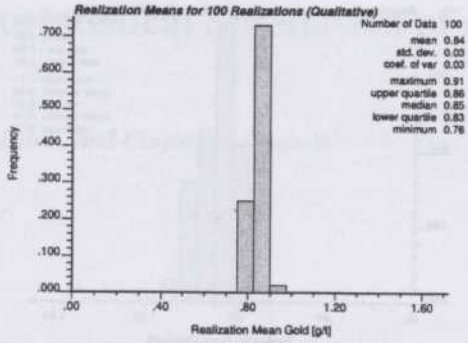

Figure 15 – The histogram of realization means for 100 realizations.

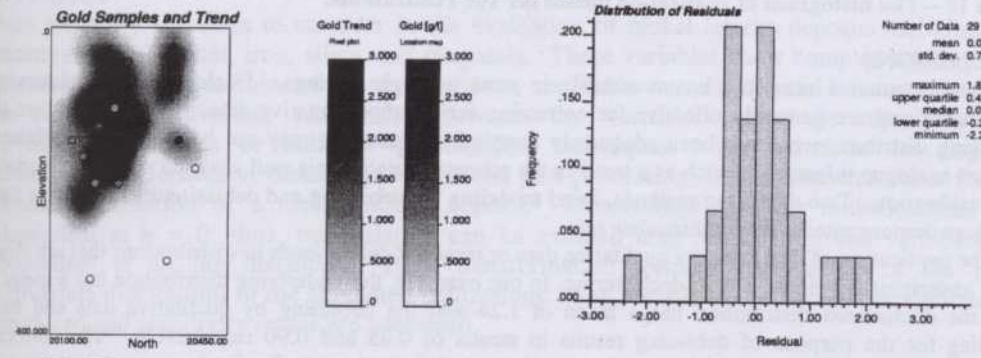

Figure 16 – The reduced "red" data set with a gold trend, and the distribution of the residuals at the data locations.

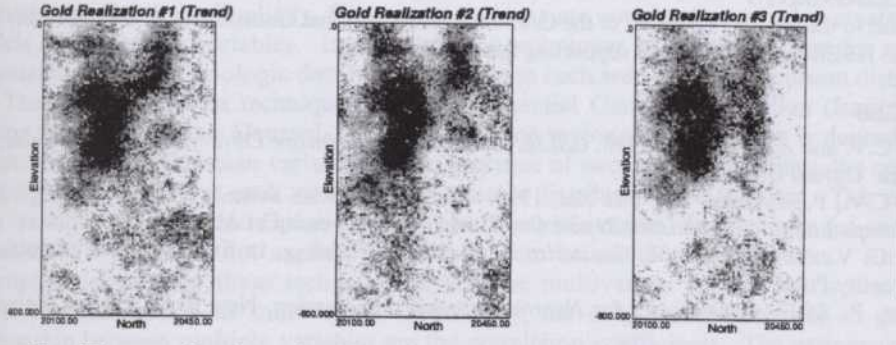

Figure 17 – Three realizations of gold grade resulting from addition of a stochastic residual and a deterministic trend model.

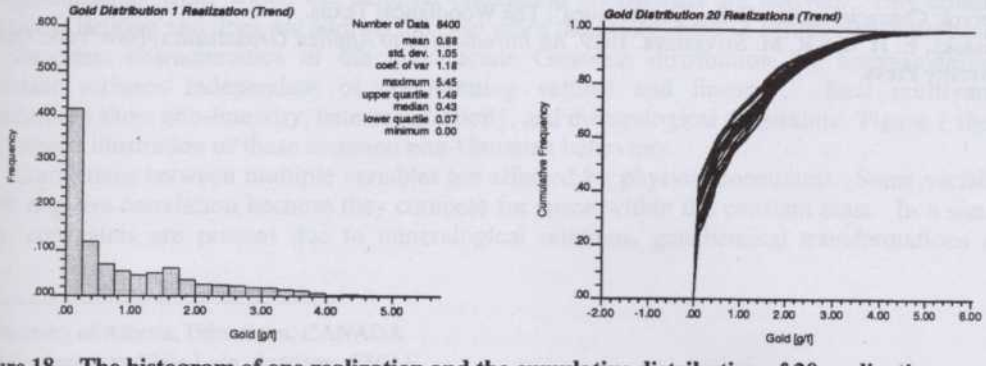

Figure 18 – The histogram of one realization and the cumulative distribution of 20 realizations.

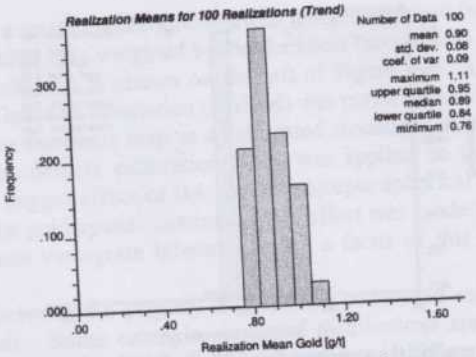

Figure 19 – The histogram of realization means for 100 realizations.

CONCLUSIONS

Nonrepresentative sampling is unavoidable in most geologic settings. Declustering techniques are widely used and are generally effective for correcting for nonrepresentative data. In settings where the underlying distribution has not been adequately sampled, declustering may not be adequate. Debiasing relies on analogue information such as a trend in the primary variable or a well sampled secondary variable and a calibration. Two debiasing methods, trend modeling for debiasing and debiasing by qualitative data, have been demonstrated with a mining data set.

The application of debiasing by qualitative data or trend modeling leads to distributions that are closer to the underlying distribution than declustering. In our example, the underlying distribution has a mean of 0.69, the declustered distribution has a mean of 1.24 and the debiasing by qualitative data and trend modeling for the purpose of debiasing results in means of 0.85 and 0.90 respectively. The corrected distribution shape and other statistics are closer to the underlying reference distribution.

ACKNOWLEDGMENTS

We are grateful to the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta and to NSERC and ICORE for supporting this research.

REFERENCES

Deutsch, C.V. and A.G. Journel. 1998. *GSLIB: Geostatistical Software Library: and User's Guide, 2nd Ed.* New York: Oxford University Press.

Deutsch, C.V., P. Frykman, and Y.L. Xie, 1999. Declustering with Seismic or "soft" Geologic Data, *Centre for Computational Geostatistics Report One 1998/1999*, University of Alberta.

Deutsch, C. V., November 2001. *Geostatistical Reservoir Modeling*, in final stages of production at Oxford University Press,

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

Gringarten, E., P. Frykman, and C.V. Deutsch. December 3-6, 2000. *Determination of Reliable Histogram and Variogram Parameters for Geostatistical Modeling*, AAPG Hedberg Symposium, "Applied Reservoir Characterization Using Geostatistics", The Woodlands, Texas.

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