A Short Note on Aggregating Information from Multiple Sources

Julián M. Ortiz* and Clayton V. Deutsch**

*Department of Mining Engineering, University of Chile
**Centre for Computational Geostatistics, University of Alberta

This short note presents some preliminary results about a study regarding the effect of the choice of the model for combining information from multiple sources. A data set is split in two subsets and four subsets. The probability of not exceeding thresholds corresponding to the first quartile, the median and the third quartile, are calculated at validation points with each subset. Then, the two subsets are combined in different ways to evaluate the required probabilities. This is repeated in the second case, where four subsets are used.

Introduction

In many areas, particularly in petroleum geostatistics, several sources of information are available. Typically well logs, geologic trends and seismic data are considered for building geostatistical models. These sources are used to inform about a particular variable at every location of the simulation field. These sources are somehow dependent of each other since they are related to the same variable of interest; however, the dependency may not be known.

Several methodologies in geostatistics consider the integration of these sources of information. Many are based on some assumption of independence. The goal of this paper is to describe the typical methodologies considered to integrate information in a simulation context and to discuss when the modeler should be concerned with the dependency between multiple variables.

Models of dependency

Several models of dependency exist that allow integration of multiple sources of information. The exact expression when integrating N variables for a conditional probability is:

$$P(A \mid B_1,...,B_N) = \frac{P(A, B_1,...,B_N)}{P(B_1,...,B_N)}$$

where A is the variable of interest and $B_1,\ldots, B_N$ are N secondary variables informing the primary variable A. Bayes law gives the exact expression for the conditional probability:

$$P(A \mid B_1,...,B_N) = \frac{P(A)\cdot P(B_1 \mid A)\cdot P(B_2 \mid A, B_1)\cdot P(B_3 \mid A, B_1, B_2)\cdot ...\cdot P(B_N \mid A, B_1,..., B_{N-1})}{P(B_1,...,B_N)}$$

This exact expression requires knowing the joint probabilities. These joint distributions are often unknown and some simplification is required. The easiest assumption is to consider that all sources are independent. This leads to the following solution:

$$\frac{P(A \mid B_1,...,B_N)}{P(A)} = \frac{P(A \mid B_1)}{P(A)} \cdot \frac{P(A \mid B_2)}{P(A)} \cdot \ldots \cdot \frac{P(A \mid B_N)}{P(A)} \cdot$$

This solution is often unrealistic and inconsistent results can be obtained, that is, they may easily fall outside the [0,1] interval.

A second approximation is the assumption of conditionally independence with respect to the variable of interest, that is:

$$\begin{aligned} &P(B_2 \mid A, B_1) = P(B_2 \mid A) \\ &P(B_3 \mid A, B_1, B_2) = P(B_3 \mid A) \\ &\vdots \\ &P(B_N \mid A, B_1,..., B_{N-1}) = P(B_N \mid A) \end{aligned}$$

This allows rewriting the expression for the conditional distribution as:

$$P(A \mid B_1,..., B_N) = \frac{P(A) \cdot P(B_1 \mid A) \cdot P(B_2 \mid A) \cdot P(B_3 \mid A) \cdot ... \cdot P(B_N \mid A)}{P(B_1,..., B_N)}$$

Since the joint distribution for the secondary variables is also unknown, this approach can be further simplified by assuming that the conditional probabilities of the complements can be written similarly:

$$P(\overline{A} \mid B_1,..., B_N) = \frac{P(\overline{A}) \cdot P(B_1 \mid \overline{A}) \cdot P(B_2 \mid \overline{A}) \cdot P(B_3 \mid \overline{A}) \cdot ... \cdot P(B_N \mid \overline{A})}{P(B_1,..., B_N)}$$

Taking the ratio between these two expressions gives the permanence of ratios model of dependency (Journel, 2002; Krishnan, 2004), which entails conditional independence:

$$\frac{P(\overline{A} \mid B_1,..., B_N)}{P(A \mid B_1,..., B_N)} = \frac{P(\overline{A}) \cdot P(B_1 \mid \overline{A}) \cdot P(B_2 \mid \overline{A}) \cdot P(B_3 \mid \overline{A}) \cdot ... \cdot P(B_N \mid \overline{A})}{P(A) \cdot P(B_1 \mid A) \cdot P(B_2 \mid A) \cdot P(B_3 \mid A) \cdot ... \cdot P(B_N \mid A)}$$

A step forward is to assume a model of redundancy such as the τ model (Journel, 2002). This model adds a parameter to compute the conditional probabilities in order to add some redundancy.

$$\frac{\frac{P(\overline{A} \mid B_1,..., B_N)}{P(A \mid B_1,..., B_N)}}{\frac{P(\overline{A})}{P(A)}} = \left(\frac{\frac{P(\overline{A} \mid B_2)}{P(A \mid B_2)}}{\frac{P(\overline{A}, B_1)}{P(A, B_1)}} \right)^{\tau_1} \cdots \left(\frac{\frac{P(\overline{A} \mid B_N)}{P(A \mid B_N)}}{\frac{P(\overline{A}, B_1)}{P(A, B_1)}} \right)^{\tau_{N-1}}$$

The parameters must be estimated or can be determined by cross-validation or jack-knife.

Other methods could be considered for integration of the information, although they are not considered in this note: Bayesian updating under a multivariate Gaussian assumption or with a non-parametric approach. This approach could be combined with transformations such as ACE for making the relationships linear, and a dimension reduction technique such as PCA to determine independent factors.

As more and more variables are available, the degree of redundancy they have with respect to the variable of interest becomes more difficult to characterize. At some point, a data-based calibration is required to avoid the bias in the estimation of conditional distributions if an assumption of independence (or conditional independence) is used.

Case study

A data set containing 923 data in two dimensions is used to illustrate the effect of the integration approach. From the data, 190 are taken as validation locations, that is, these are the locations where the probability of exceeding a given threshold will be estimated.

The remaining data are first divided into two subsets (**Figure 1**). These subsets have been taken randomly, therefore, only small fluctuations can be seen in their statistics (**Figure 2**).

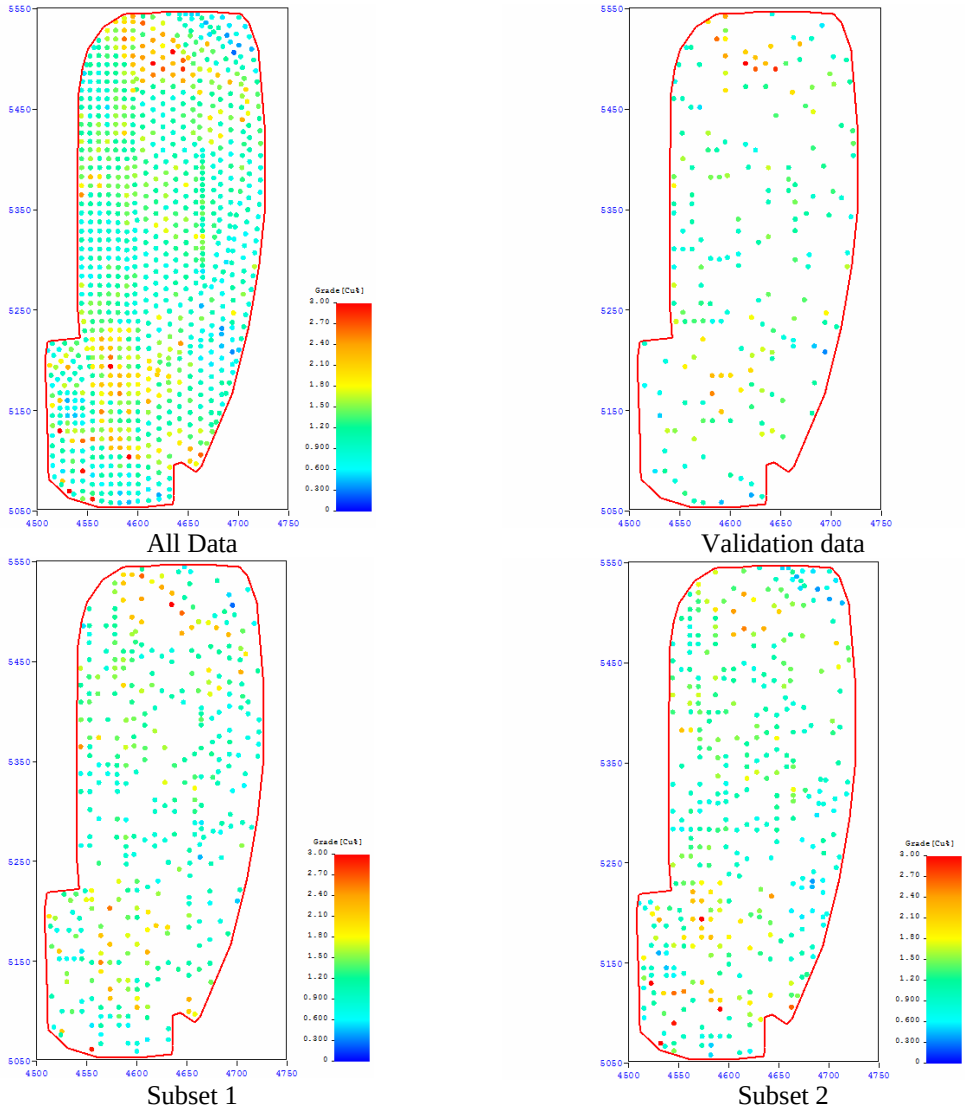

Figure 1: Location maps of the available data (top left), validation set (top right), and two subsets (bottom).

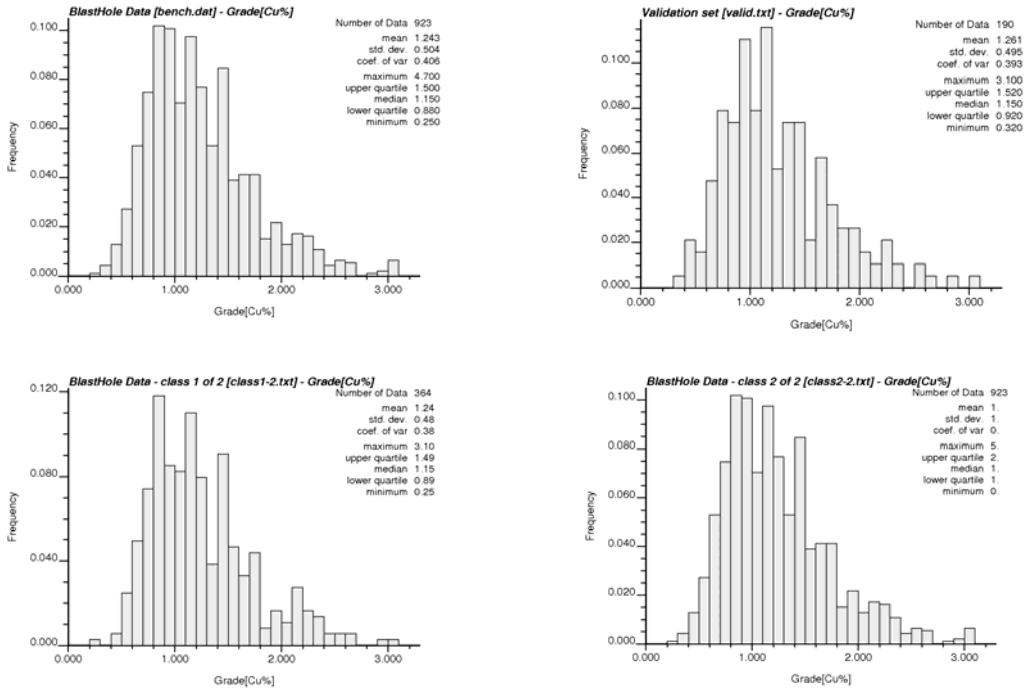

Figure 2: Histograms and basic statistics of the data available, validation set, and two subsets.

The probability of not exceeding the first, second (median) and third quartiles of the data distribution are calculated. From these estimated probabilities, two performance measures are calculated:

1. The mean squared error calculated as the sum for the three estimated probabilities of the differences between the estimated probability computed and the true probability, which is known since the actual values are known. The latter can be either 0 or 1.
2. The maximum difference for the three points computed of the cumulative distribution function, with respect to the true value at each threshold.

The estimated probabilities from the two subsets are then combined using different methods, namely:

1. Full independence assumption
2. Conditional independence assumption
3. Permanence of ratios assumption with parameter t=0.1
4. Permanence of ratios assumption with parameter t=0.5
5. Permanence of ratios assumption with parameter t=2.0

Results are compared with the case when all the data are used to estimate the probabilities of not exceeding the three thresholds, with indicator kriging.

MSE	*Average*	*Min*	*Max*
All data	0.1472	0.0025	0.8271
Subset 1	0.1526	0.0089	0.6444
Subset 2	0.1558	0.0021	0.8140
FI	0.2255	0.0010	1.8539
CI	0.1718	0.0003	0.7431
PR 0.1	0.1513	0.0094	0.6483
PR 0.5	0.1582	0.0021	0.6999
PR 2.0	0.1927	0.0000	0.7563

Table 1: Performance comparison of the integration methods for the case of two subsets, using mean squared errors.

From **Table 1**, it can be seen that, on average the MSE is reduced as all the data are used in a consistent manner. In this case, the first row of the table shows the results of performing indicator kriging using all the data that are not in the validation set. Since the estimated probabilities are calculated from a kriging system that accounts for the dependence and redundancy of the data, the MSE value is smaller as when the probabilities are estimated from each subset independently (second and third rows).

The use of a model of full independence clearly goes against the nature of these data. The consequence is that probabilities are easily estimated outside the allowed range of [0,1]. The MSE is therefore very high. This model should be immediately discarded as some correlation is always present when different variables are used to estimate the same attribute. If they are related to the attribute of interest, some relation between them must exist.

The conditional independence assumption provides in this case results that are worse than considering each subset separately. This means that disregarding the correlation between the information (or its redundancy) may play against the goal of integrating additional information.

The permanence of ratio models with different degrees of dependence show that the correlation can be calibrated and, if properly accounted for, results are improved. In this application, a t parameter of 0.1 provide better results than considering only one subset at a time, without integrating the information of the second subset. Jack-knife appears as a good approach to calibrate this parameter to account for dependency between variables, when their joint distribution is not known.

Table 2 shows the summary of the performance when measured with the maximum difference between the estimated probability value at a given threshold and the true probability (indicator value) at that same threshold.

D max	*Average*	*Min*	*Max*
All data	0.5332	0.0862	1.0000
Subset 1	0.5505	0.1599	0.9771
Subset 2	0.5465	0.0521	0.9700
FI	0.6094	0.0442	2.0367
CI	0.5986	0.0216	0.9988
PR 0.1	0.5580	0.1369	0.9781
PR 0.5	0.5819	0.0580	0.9915
PR 2.0	0.6139	0.0031	1.0000

Table 2: Performance comparison of the integration methods for the case of two subsets, using the maximum difference in probability between the estimated and true probability for the three indicators used.

From this analysis, it can be seen that, although the use of the permanence of ratios model with a parameter t=0.1 performs best than all other methods, however, it did not outperformed the result from using a single subset to perform the estimation of the probabilities.

The application is repeated considering now four subsets (**Figure 3**). Once again, these four subsets represent adequately the full data set, as seen in their histograms (**Figure 4**).

The same models for integrating the redundancy are used. The permanence of ratios is applied with a constant t parameter to integrate successively the four sources of information, that is, t is applied three times to account for the redundancy.

Results are summarized in **Table 3** for the MSE and **Table 4** for the maximum difference in probability.

MSE	*Average*	*Min*	*Max*
All data	0.1472	0.0025	0.8271
Subset 1	0.1593	0.0207	0.6437
Subset 2	0.1654	0.0140	0.6073
Subset 3	0.1664	0.0021	0.7492
Subset 4	0.1651	0.0099	0.7358
FI	0.6569	0.0005	7.5154
CI	0.1961	0.0000	0.7816
PR 0.1	0.1574	0.0095	0.6296
PR 0.5	0.1837	0.0002	0.6660
PR 2.0	0.2170	0.0000	0.6724

Table 3: Performance comparison of the integration methods for the case of four subsets, using mean squared errors.

D max	*Average*	*Min*	*Max*
All data	0.5332	0.0862	1.0000
Subset 1	0.5612	0.1850	0.9280
Subset 2	0.5715	0.1867	0.9604
Subset 3	0.5680	0.0663	0.9703
Subset 4	0.5612	0.1344	0.9425
FI	0.9484	0.0372	4.2231
CI	0.6287	0.0013	1.0000
PR 0.1	0.5807	0.1535	0.9617
PR 0.5	0.6250	0.0197	0.9991
PR 2.0	0.6385	0.0000	1.0000

Table 4: Performance comparison of the integration methods for the case of four subsets, using the maximum difference in probability between the estimated and true probability for the three indicators used.

Results are similar in this case than for the case of two subsets.

Discussion

The problem of integrating information from multiple sources is complicated, particularly when these sources are very different. Many techniques are available for integrating information. The main difficulty of this process is to understand the relationship between variables, particularly their redundancy and correlation.

Not accounting for the redundancy will lead to believing that, as more sources of information are available, more is known about the variable of interest. This simple study showed that full independence is a poor assumption. This was expected, since all subsets correspond to the same variable, hence the redundancy of information is very high. This is supported by the fact that a low t value in the permanence of ratio model, generated the better results. This can be interpreted in words as "give little weight to subsets 2, 3, and 4, as most of the information was given by subset 1".

In this case study, uncertainty due to the inference of modeling parameters has not been accounted for. The same parameters (variogram model, threshold and probabilities) have been used in all cases, hence the importance of the amount of data for estimating key parameters of first order of importance for the modeling process, has not been included.

Other techniques of integration of information should be investigated in further studies, as well as the use of sources of information of different nature. We are exploring different methods to aggregate information data sources and when it is necessary to seek complex and problem-specific models of redundancy.

References

Not all these have been cited in the text, but they are relevant for the topic of the paper.

Efron, B, 1982. The jackknife, the bootstrap, and other resampling plans. Philadelphia, Society for Industrial and Applied Mathematics. Pp. 100.

Journel, A. G., 2002. Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. Mathematical Geology, 34(5):573–596.

Caumon, G. and Journel, A. G., 2004. Early uncertainty assessment: application to a hydrocarbon reservoir appraisal. In O. Leuangthong and C. Deutsch (Eds), Proc. Seventh International Geostatistics Congress, Banff , Springer, p. 551-558.

Doyen, P., Guidish, T., de Buyl, M., 1989. Monte Carlo simulation of lithology from seismic data in a channel sand reservoir. SPE 19588. Society of Petroleum Engineers (SPE), Richardson, TX, USA

Doyen, P. M., den Boer, L. D., and Pillet, W. R., 1996. Seismic porosity mapping in the Ekofisk field using a new form of collocated cokriging. Society of Petroleum Engineers. SPE 36498.

Neufeld, C., and Deutsch, C. V., 2004. Incorporating Secondary Data in the Prediction of Reservoir Properties Using Bayesian Updating. In Centre for Computational Geostatistics, Report 6.

Zanon, S., and Deutsch, C. V., 2003. Predicting Reservoir Performance with Multiple Geological, Geophysical, and Engineering Variables: Bayesian Updating Under a Multivariate Gaussian Model. In Centre for Computational Geostatistics, Report 5.

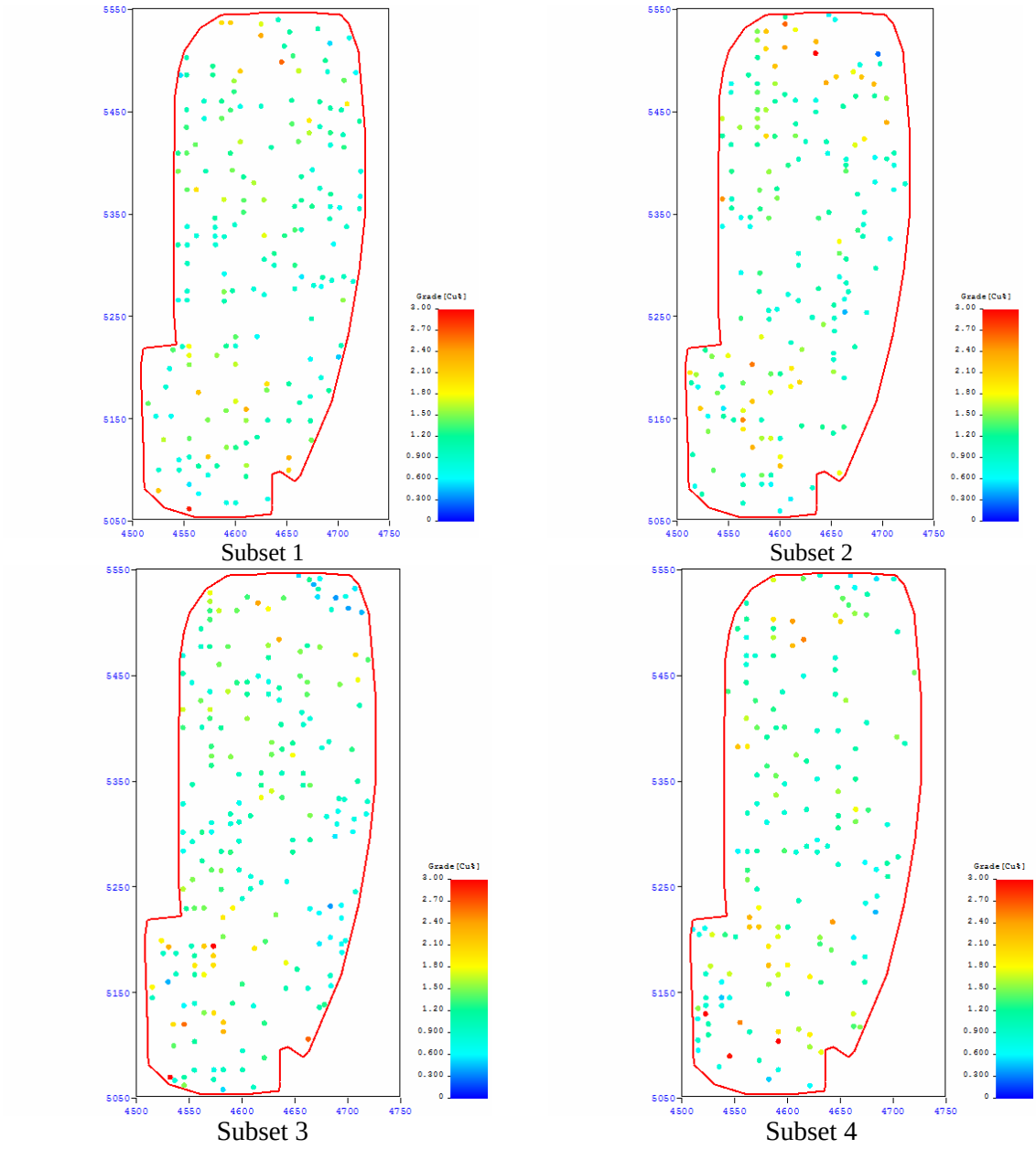

Figure 3: Location maps of the four subsets used in the second analysis.

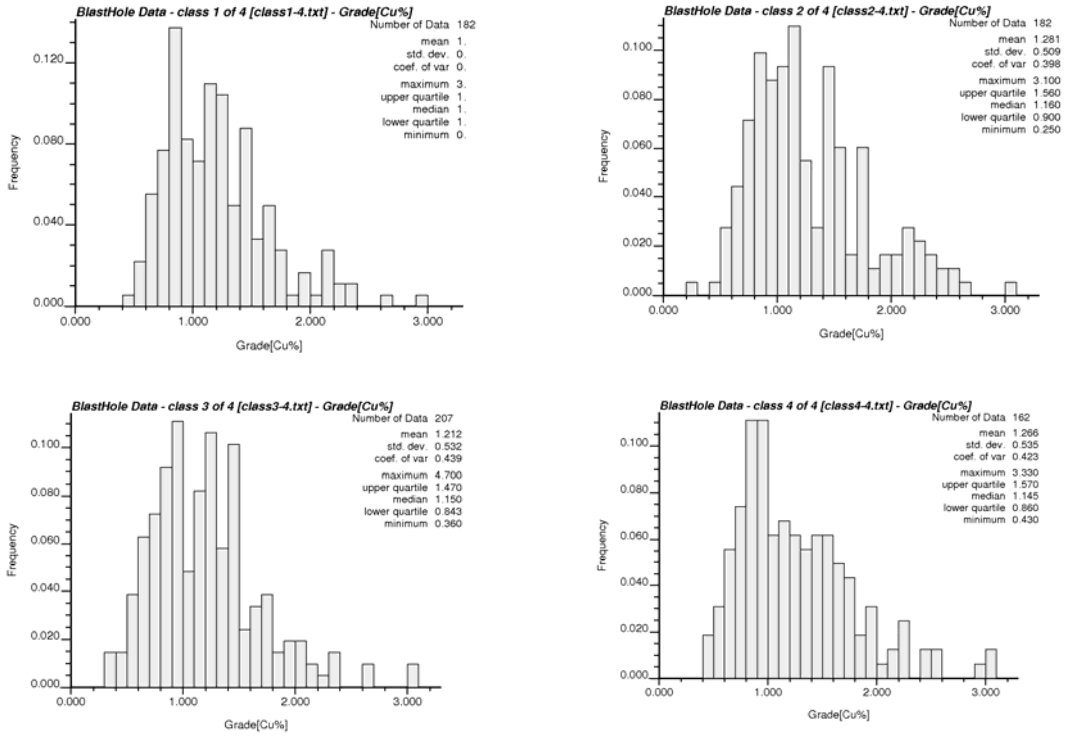

Figure 4: Histograms and basic statistics of the four subsets.