Programs for Data Spacing, Uncertainty, and Classification

Brandon J. Wilde and Clayton V. Deutsch

Data spacing and data density can be calculated different ways. Some methods for calculating these values as well as a tool for doing so are presented. These methods are based on the number of samples within some representative volume. The resulting data spacing/density values can be used to classify resources/reserves based on geometric criteria. A tool for this is discussed. Determining uncertainty measures at different data spacings is also useful. A program for performing these calculations as well as a program for plotting the results is presented.

1 Introduction

A combination of geometric and probabilistic criteria has been suggested as a reasonable basis for resources and reserves classification (Deutsch *et al.,*2006). An approach where geometric criteria are backed up by probabilistic criteria is advised. Tools for calculating these values are useful as classification is required for public disclosure. The first tool, `dataspacing`, can be used for calculating local geometric measures of data spacing and data density. Different ways of determining spacing and density are discussed. The second tool, `classify`, performs classification based on one variable and specified thresholds. The third tool, `aduds`, performs the methodology described in paper 108 herein and is used to calculate a number of uncertainty measures for a spatial variable for different data spacings. The final tool presented is used to present the results from the previously mentioned program. A plot of uncertainty vs. data spacing is produced.

2 Program for Determining Data Spacing: `dataspacing`

`Dataspacing` is a program for calculating data spacing on a regular grid in two or three dimensions with irregularly spaced data. This program reads in a data file and determines local data spacing values on a regular grid using the methods described below. It also determines local data density. These values are written out to a grid file.

2.1 Data Spacing

Data spacing is the distance between adjacent data for a representative area. A densely sampled area has a small spacing relative to an area that is sparsely sampled. Data spacing at a location, $s(\mathbf{u})$, is determined by considering the number of nearby samples, $n_v(\mathbf{u})$, within some volume, $V(\mathbf{u})$. If $V(\mathbf{u})$ is two-dimensional, the square root of $V(\mathbf{u})$ divided by $n_v(\mathbf{u})$ gives data spacing as shown in Equation 1. When drillholes are vertical, the calculation of data spacing in three dimensions reduces to a two dimensional problem and Equation 1 can be used. When the drillholes are not all vertical, as shown in Figure 1 left, the data spacing calculation must consider a three-dimensional volume. The along-hole spacing, c, is included in the calculation thus defining the equivalent square drillhole spacing.

$$s(\mathbf{u}) = \left(\frac{V(\mathbf{u})}{n_V(\mathbf{u})} \right)^{1/2} \quad 1$$

$$s(\mathbf{u}) = \left(\frac{V(\mathbf{u})}{c \cdot n_V(\mathbf{u})} \right)^{1/2} \quad 2$$

To calculate data spacing at a location, either $V(\mathbf{u})$ or $n_V(\mathbf{u})$ are normally fixed. If $n_V(\mathbf{u})$ is fixed i.e. $n_V(\mathbf{u}) = n_V, \forall \mathbf{u} \in A$, the volume $V(\mathbf{u})$ required to encompass the n_V data is calculated and spacing is determined as defined previously. If $V(\mathbf{u})$ is fixed i.e. $V(\mathrm{u}) = V, \forall \mathrm{u} \in A$, the number of observations $n_V(\mathbf{u})$ that fall within V is determined and spacing is determined as defined previously. The choice of n_V or V affects the results: too small of a volume or too few samples leads to noisy results; too large of a volume or too many samples leads to over smoothing. The units of spacing depend on the units of V. For example, if V is 1 mile x 1 mile (1 section) spacing has units of miles.

2.2 Data Density

Data density is the number of data observations per unit volume, commonly reported as number of data per section or hectare. Data density at a location, $d(\mathbf{u})$, is determined by considering the number of nearby samples, $n_v(\mathbf{u})$, within some volume, $V(\mathbf{u})$. Dividing the number of samples by their volume gives data density (Equation 3). If the data are arranged such that many observations fall within a small volume, data density is high. If a large volume contains few observations, data density is low.

$$d(\mathbf{u}) = \frac{n_V(\mathbf{u})}{V(\mathbf{u})} \quad 3$$

Data density at a location, $d(\mathbf{u})$, is determined in the same manner as data spacing by fixing either V or n_V and calculating the other parameter. The units of density are expressed as a quantity per volume such as wells per section. The units of density can be easily converted. There are ten thousand square meters in a hectare and 2,589,988.11 square meters in a section. To convert density simply multiply by the appropriate constant. For example, if density is given as 1 well per section, this can be converted to wells per section by multiplying by 2,589,988.11 divided by 10,000 to give 259 wells per hectare. When the data coordinates are in units of meters, density will have units of samples per square meter. This should be converted to more meaningful units.

2.3 V and n_V Calculations

Both data spacing and data density are calculated from V and n_V. One of these values is fixed allowing the other to be calculated from the nearby data. The program provides for V to be either circular or square, depending on the preference of the user. A square window works best when the data are regularly spaced while a round window works well for irregularly spaced data. Artifacts can occur otherwise.

If V is constant, n_V is location dependent. The size of V is specified by a in two dimensions (Figure 2) and by a and h in three dimensions (Figure 3). This 'moving window' volume is translated over the domain with an origin at $\mathbf{u}$, counting the number of data, $n_V(\mathbf{u})$, that fall within it to arrive at data spacing and data density at each location.

If n_V is constant, V is location dependent. In two dimensions, $V(\mathbf{u})$ is calculated by finding $a(\mathbf{u})$ such that exactly n_v data fall within V($\mathbf{u}$). $a(\mathbf{u})$ is twice the average distance to the n_V and n_V+1 samples as shown in Equation 4 where r_i is the distance to the i^{th} sample from $\mathbf{u}$. This is illustrated in Figure 4 for $n_V=4$. $V(\mathbf{u})$ for a square volume in two dimensions is calculated as shown in Equation 5 and for a circle volume in Equation 6.

$$a(\mathbf{u}) = r_{n_v} + r_{n_v+1} \quad 4$$

$$V(\mathbf{u}) = a^2(\mathbf{u}) \quad 5$$

$$V(\mathbf{u}) = \frac{\pi a^2(\mathbf{u})}{4} \quad 6$$

In three dimensions, $V(\mathbf{u})$ is also calculated by finding $a(\mathbf{u})$, but considering only those data that fall within $\pm h/2$ from $\mathbf{u}$ in the vertical direction. The volume is determined by applying Equation 5 or 6 as appropriate and multiplying by h. Samples are found – not drillholes.

3 Program for Classification: `classify`

Geometric criteria have been proposed as a basis for classifying resources and reserves (Deutsch *et al.,* 2006). This program performs this classification. Any criteria could be used for classification; this is a generic program which reads in a gridded file and n classification thresholds t_i, $i=1,...,n$. The values in the gridded file are then assigned a classification value c_j, $j=1,...,n+1$. No effort is made to interpret these classifications i.e. measured, indicated or inferred. This step is left to the practitioner and is dependent on the type of deposit under study and the associated degree of spatial heterogeneity.

This process is illustrated in Figure 5. n classification thresholds are applied to the random variable Z. The first classification, c_1, is applied to the values of Z less than t_1. The second classification, c_2,

is applied to values of *Z* greater than or equal to t_1 and less than t_2. This continues up to the n^{th} threshold where any values of Z greater than or equal to t_{n-1} and less than t_n are classified as *n* and any values of *Z* greater than or equal to t_n are classified as *n+1.*

4 Program for Determining Uncertainty at Various Data Spacings: aduds

A methodology using SGS to evaluate the relationship between uncertainty and data spacing has been proposed (see paper 108 herein). This program simulates realizations of the truth, samples the realizations at specified spacings, generates new realizations conditional to the samples, and calculates local measures of uncertainty. It then writes out the local uncertainty measures for each data spacing allowing the distribution of uncertainty for a given spacing to be compared to the distributions for other spacings.

The parameters are similar to those for the sgsim program. Two files are output: one contains the expected local uncertainty for each data spacing while the other has the local uncertainty value for all locations. A point-scale grid is specified as well as a block size and the data spacings of interest. The block size and data spacings must be multiples of the point-scale grid spacing. The number of truth realizations, *L*, as well as the number of conditional realizations for each truth realization, *K*, is specified (see paper 108 for details). The methodology considers a heteroscedastic sampling error which requires the specification of the desired sampling error to consider. Two parameters relating to the local uncertainty must be specified. The first is a threshold which is necessary to consider misclassification errors. The other is a measure of +/- uncertainty as a percentage. For example, if this parameter is set at 15%, the program calculates the probability for the simulated block values to be within ±15% at each location. Finally, a random number seed and variogram model are specified.

$$nxb = \frac{nx}{bxsize/pxsize}$$
$$nyb = \frac{ny}{bysize/pysize} \qquad 7$$

The output file with the local uncertainty measures contains *nxb*nyb*L* local measures where *L* is the number of truth realizations and *nxb* and *nyb* are the number of blocks in the *x* and *y* directions as defined in Equation 7 where *nx* is the number of point-scale estimates, *bxsize* is the size of the block, and *pxsize* is the point-scale estimate spacing in the x-direction. The same applies to the y-direction. For example, if there are 100 point-scale estimates spaced every 1m in the x-direction and the block size is 5m, there are 20 blocks in the x-direction.

5 Program for Plotting Uncertainty Distributions for Different Data Spacings: undistplt

The aduds program previously discussed outputs local uncertainty values for different data spacings. It is useful to visualize the distributions of uncertainty for different spacings on the same plot, particularly when the analysis has been performed for various input parameters (e.g. different variogram models). This permits the observation of the relationship between uncertainty and data spacing as well as showing the influence of variations in the input parameters. An example of the output from this program is shown in Figure 6. The uncertainty measure is standard deviation. The black distributions represent one set of input parameters while the red distributions represent another.

The parameters follow typical GSLIB format (Deutsch and Journel, 1998). The number of data spacings as well as the number of distributions per data spacing are required as input to the program. After specifying the plot limits (setting the max < min will determine these automatically), two parameters controlling the size and position of the line histograms are specified. The first parameter controls the width of the line histogram (*w* in Figure 7) while the other controls the offset (*o* in Figure 7). These are useful for modifying the plot when different numbers of distributions are plotted and when the line histogram is omitted, which is the next parameter in the file.

6 Conclusions

Data spacing and data density can be used as criteria for classifying resource/reserves. Calculating these geometric measures is straightforward, a tool for doing so was presented. A choice can be made to classify resources/reserves based on data spacing or data density thresholds. A program for applying the classification based on thresholds was presented. It can be useful to validate the choice of threshold by evaluating uncertainty at various data spacings. A program for evaluating uncertainty at different data spacings was reviewed. It produces distributions of uncertainty for each data spacing. A plotting program for comparing the distributions is useful and was presented herein.

References

Deutsch CV and Journel AG, 1998. GSLIB: Geostatistical Software Library and User's Guide, Oxford University Press. New York. 369p.

Deutsch CV, Leuangthong O, and Ortiz JC, 2006. A case for geometric criteria in resources and reserves classification. Centre for Computational Geostatistics Annual Report Eight. University of Alberta. 22p.

Figures

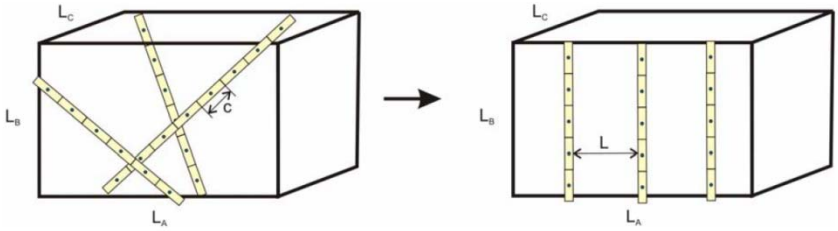


Figure 1: Illustration of a 3-D volume containing *n* samples with along-hole spacing of *c* (from Deutsch *et al.,* 2006).

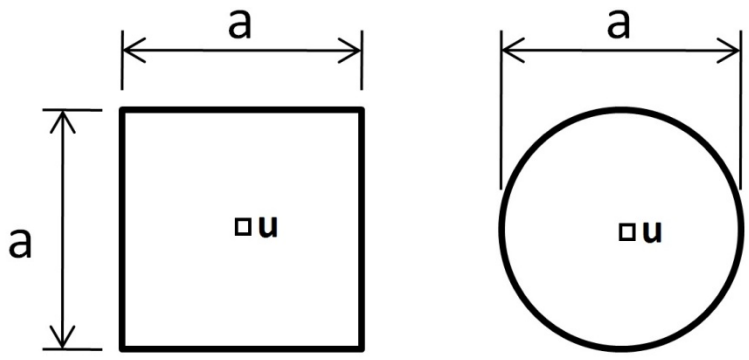


Figure 2: The size of a fixed volume, *V*, is specified by *a* in two dimensions**.** Note that the central location (**u**) is scanned over the entire domain.

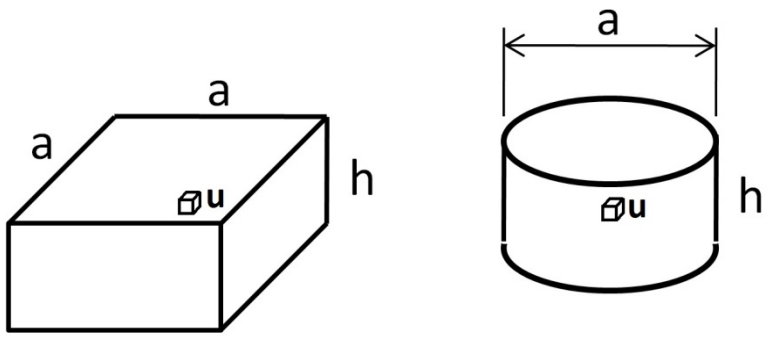
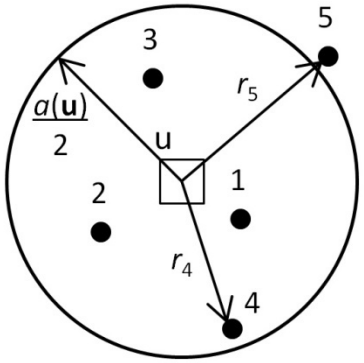


Figure 3: The size of a fixed volume, *V*, is specified by *a* and *h* in three dimensions.



$$\frac{a(\mathbf{u})}{2} = \frac{r_4 + r_5}{2}$$

$$a(\mathbf{u}) = r_4 + r_5$$

Figure 4: Example of determining $a(\mathbf{u})$ for n_v=4.

Figure 5: Illustration of classifications, c_i, relative to classification thresholds, t_j.

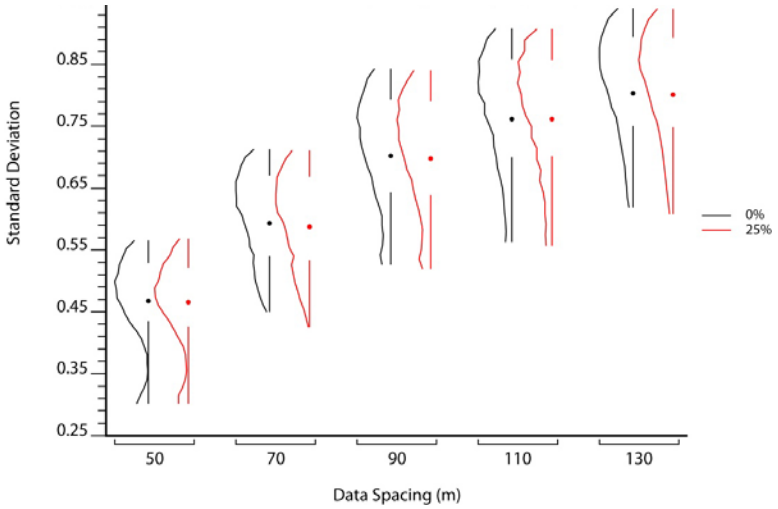


Figure 6: Example of plot output by `undistplt`.

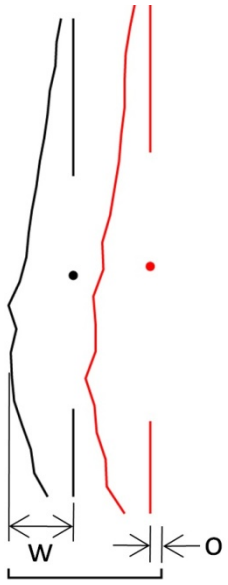


Figure 7: Illustration of the width, *w*, and offset, *o*, parameters.