Automatic Determination of Uncertainty versus Data Density

Brandon Wilde and Clayton V. Deutsch

It is useful to know how various measures of uncertainty respond to changes in data density. Calculating the change in uncertainty for different data densities required the practitioner to run numerous separate modeling and post-processing programs. This process has been streamlined into one program. The uncertainty measures calculated by this program are discussed. The results are sensitive to a number of input parameters. These sensitivities are reviewed. A new method for non-parametrically fitting a monotonic increasing or decreasing function is presented.

Introduction

The exploration of a geological resource requires a decision to be made regarding adequate data density. This decision guides the sampling effort. Three strategies have traditionally been used to determine the adequate data density: 1) to minimize the sampling cost for a specified level of uncertainty; 2) to minimize uncertainty for a given sampling budget; or 3) to respond to demands on sampling made by the legal authorities (Back, 2007). The latter two strategies are relatively straightforward. For a given budget, the sampling pattern that minimizes uncertainty has been determined (see McBratney, 1981). Regulatory requirements often leave the practitioner with little choice regarding the field measurement design. The first of these three strategies is the most ambiguous. Minimizing sampling cost is straightforward, but determining a *measure* of uncertainty as well as an *adequate level* of uncertainty is not. These factors must be determined in order to make the correct decision regarding data density.

Numerous methods have been applied to determine an adequate data density, primarily in the area of groundwater sampling. One of the most popular methods has been to minimize the global estimation variance by optimizing the locations of the data (Carrera et.al, 1984; Fiering, 1965; Hughes & Lettenmaier, 1981; Loaiciga, 1989; Rouhani, 1985; and Virdee & Kottegoda, 1984). Criminisi et.al (1997) chose to minimize the coefficient of variation of the optimal cost distribution while Aspie & Barnes' (1990) objective was to minimize the expected cost of classification errors. Each of these methods determines an optimum configuration, but none examines the dependence of uncertainty on data density.

Uncertainty, of course, does not depend solely on data density. A number of other factors also influence uncertainty including local grade level, spatial variability of the grades, economic selection thresholds, scale, and uncertainty in the modeling parameters. We primarily consider the relationship between uncertainty and data density.

It is useful to define the geometric measures used to describe data positioning such as drillhole spacing, drillhole density, and closeness to the nearest drillhole. These measures are related. They are illustrated and defined mathematically in Figure 1. Data spacing is simply the distance between adjacent data locations. For data located on a regular grid, the spacing is measured in two perpendicular directions. Data density is the number of data within some nominal volume or area. It is common to consider the number of data per section (1 mile x 1 mile) or data per hectare (100m x 100m). These measures are related. 1 sample/hectare is equal to 259 samples per section. When data are irregular, it is common to used data density to calculate effective data spacing.

We examine how various measures of uncertainty are related to data density. The methodology for determining various measures of uncertainty for different data spacings is presented. The concept of uncertainty is discussed as it applies to varying data density. Various measures of uncertainty that are calculated for given data spacings are briefly discussed. The process for summarizing the results of uncertainty calculations for many data densities is explained. Results are given for a number of different input statistics.

Methodology

A methodology has been developed for determining uncertainty for different data densities. The process of determining uncertainty for a given data density is as follows.

1. Generate unconditional 'truth' realization

2. Sample truth realization at desired spacing with error
3. Use samples to generate conditional realizations
4. Calculate uncertainty measures from the realizations

This can be done for many 'truth' realizations for the given spacing. The details of each of these steps are given.

A realization of the truth is generated through unconditional sequential Gaussian simulation. The desired covariance structure of the truth is input through a variogram model. The unconditional simulation is performed in Gaussian units and then back transformed according to an input reference distribution. The high resolution point estimates are block averaged and these block values are saved as the truth.

From the back transformed point estimates, 'samples' are taken at a specified spacing. Sampling error is added to these samples to simulate true sampling procedures and their inherent error. The sampling error is proportional to the sample value. The user specifies the standard deviation of the error as a proportion of the true sample value. This is the value, *X*, in equation (1). The standard deviation is then multiplied by a random standard normal value (*Y* in equation (2)) and this value is added to the true value to get the sample value (equation (2)).

These samples are considered as conditioning data and are normal scored to have a mean of zero and a standard deviation of one. They are then used to generate conditional Gaussian realizations at the same resolution as the point estimates. The simulated values are back transformed and block averaged to the same resolution as the simulated truth. Uncertainty measures are then calculated from the block averaged realizations.

Uncertainty and its Measures

Uncertainty is exactly what the word implies, a lack of complete certainty, that is, the existence of more than one possibility where the *true* outcome/state/result/value is not known (Hubbard, 2007). As it applies to the geological sciences, uncertainty can be defined as the state of having limited knowledge where it is impossible to describe the existing state. It is easy to understand the relationship between data density and uncertainty: increased data density will lead to decreased uncertainty, that is, increased certainty regarding the existing state. Uncertainty is described and quantified through various measures. A measure of uncertainty is defined as a set of probabilities assigned to a set of possibilities (Hubbard, 2007).

There are numerous measures of uncertainty that could be compared for different data densities. The measures of uncertainty considered here include two measures of spread: standard deviation and the difference between the 90th and 10th percentiles; two relative measures of spread: coefficient of variation and the difference between the 90th and 10th percentiles divided by the 50th percentile; a measure of precision: the probability of a simulated value to fall within a given distance from the mean; and two classification measures: the probability of Type I and Type II errors.

Each of these measures is discussed individually. An example is shown for each measure.

Standard Deviation

At a location there are *L* simulated values where *L* is the number of realizations. The standard deviation of these *L* values is determined. The expected standard deviation is determined by considering all locations. This value is reported for the given data density. Standard deviation decreases as data density increases as illustrated in Figure 2a.

P90-P10

From the *L* simulated values at a location, the 10th and 90th percentiles can be determined. The difference in these values is reported for the given data density. The difference decreases as data density increases as shown in Figure 2b.

Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean. The mean and standard deviation of the *L* simulated values at a location are determined and their ratio reported for the given data density. This ratio decreases as data density increases as illustrated in Figure 2c.

(P90-P10)/P50

In addition to the 10th and 90th percentiles, the 50th percentile is determined from the *L* simulated values. The ratio of the difference between the 10th and 90th percentiles to the 50th percentile is reported for the given data density. This ratio decreases with increasing data density as shown in Figure 2d.

Precision

Precision is measured as the probability of an estimate to be within a given distance from the truth. For example, what is the probability of the estimate to be within ±15% of the truth. The truth is taken as the mean of the *L* simulated values and the probability is the number of simulated values that are within the given distance divided by *L*. This probability is reported for the given data density. As data density increases, so also does this probability as illustrated in Figure 2e.

Type I & Type II Errors

Type I and II errors are used to describe possible errors made in a decision making process. In this context, they are only useful with respect to some critical threshold, or cutoff grade. The type I error refers to classifying a block as ore when it is really waste (false positive) while the type II error refers to classifying a block as waste when it is really ore (false negative). Both of these errors decrease with increasing data density. The probability of misclassification is sensitive to the cutoff value. Overall, the probability of misclassification decreases as data density increases as shown in Figure 2f.

Program

The process of determining these measures of uncertainty for various data spacings has been developed within the program datspacsim which was built off of the code for the sgsim program. Various parameters have been added and removed from the sgsim parameter file (Figure 3). The input data file parameter and its associated parameters have been removed as no conditioning data is required. The output file is now the first parameter as seen on line 4. Line 5 specifies the output file for the monotonic non-parametric fitting described later. The file with the reference distribution is entered on line 6. The typical reference distribution parameters comprise lines 7-11. Line 12 specifies the resolution of the point estimates in the horizontal plane. Line 13 enters the size of the blocks into which the point estimates are averaged. The block size must be a multiple of the point spacing. Line 14 controls the number of cells in the model by determining the number of samples to consider in the x and y directions. On line 15, the parameter that specifies the number of data spacings, *n*, to consider is given. The next *n* lines are where the data spacings in the x and y directions are given. The vertical grid parameters are specified on line 19. Line 20 specifies how many reference models to generate per data spacing and line 21 establishes the number of realizations to generate per reference model. The percent sampling error is specified in Line 22. This is the value, *X*, used in equation (1). The standard deviation of the sampling error is proportional to the sampled true value. It is multiplied by a value drawn randomly from the standard normal distribution and this value is added to the sampled true value to give the sample value at that location as shown in equation (2). The critical threshold option and value are entered on Line 23. The desired confidence interval to report is given on line 24. The rest of the parameters are the same as for sgsim and are for entering the random number seed and variogram model (Deutsch & Journel, 1998).

$$\sigma_{error} = \frac{X}{100\%} \bullet \; z_{true} \tag{1}$$

$$z_{samp} = z_{true} + Y \bullet \sigma_{error} \tag{2}$$

Data density is the preferred measure for comparing to the various uncertainty measures. Unfortunately, data density is not as easy to visualize as data spacing. To account for the increased ease of thinking in terms of spacing, the user enters the desired x and y data spacing in the parameter file. The program then converts this measure into Effective Areal Data Density (EADD) to take advantage of the linear relationship between data density and data quantity.

For each data spacing, a number of truth and estimates models are generated. The uncertainty measures are averaged and written out giving one value for each uncertainty measure. A distribution of uncertainty could be considered. The monotonic non-parametric fitting procedure described herein is employed to create a monotonic fit to the results. The results of this fitting are written out to the file specified in line 5. These measures can be plotted using scatplt2008.

Demonstration

The measures of uncertainty are sensitive to a number of input parameters including the reference distribution, variogram, scale, number of realizations, and the classification threshold. This section demonstrates how uncertainty responds to each of these parameters.

Reference Distribution

The spread of the histogram affects uncertainty. A distribution with more spread will cause more uncertainty at the same data density than a distribution with less spread. Consider the three distributions shown in Figure 4. Each has a mean of 1.0. The coefficient of variation varies from 0.5 to 2.0. Figure 5 shows how the uncertainty is related to the reference distribution. As the coefficient of variation increases, so does the uncertainty at a given data density. There is a very direct relationship. As the coefficient of variation doubles, so does the uncertainty.

Variogram

The continuity of the variogram affects uncertainty for a given data density. As one would expect, more continuity in the variogram decreases uncertainty. Three isotropic variograms with one structure and no nugget were considered. The variogram range varied from 50 to 200m as shown in Figure 6. Figure 7 demonstrates the relationship between uncertainty and the variogram range. As expected, the uncertainty decreases as the continuity increases. At very low densities (wide spacings) the uncertainty is the same; most unsampled locations are not near data. The uncertainty decreases rapidly for the more continuous variogram. The decrease in uncertainty is not so rapid for the less continuous variogram.

Number of Realizations

The user must determine how many realizations to use for determining the various measures of uncertainty. Increasing the number of realizations is computationally expensive. The results are relatively insensitive to the number of realizations used as shown in Figure 8. The bandwidth of uncertainty stabilizes with more realizations. For instance, at low data densities the uncertainty using 500 realizations in Figure 8 is slightly more stable than the uncertainty using 100 realizations. This is an expected result. More realizations will increase stability, but also increase computational expense; 50 realizations is recommended as a minimum.

Scale

Uncertainty depends on the size of the blocks into which the point estimates are averaged. Larger block sizes will smooth out the highs and lows reducing the uncertainty. This smoothing does not happen as much for smaller blocks. Increasing the block size decreases the uncertainty thereby increasing the precision of the estimates. For this comparison, point scale estimates were generated at 4m spacing which were then averaged into blocks of 4, 8, and 48m size (Figure 9). Figure 10 demonstrates how increased block size decreases uncertainty increasing precision. Precision is higher for the larger block sizes.

Classification Threshold

The probability of misclassification is highly dependent on the classification threshold. Consider the cross plot of true and estimated values in Figure 11. The blocks that have been classified as ore but are truly waste have been highlighted in red; the blocks that have been classified as waste but are truly ore have been highlighted in blue. This has been done for three cutoffs: 0.25, 1.0, and 2.0. The probability of misclassification is quite low for cutoffs of 0.25 and 2.0 and is higher for a cutoff of 1.0. Figure 12a demonstrates the relationship between the probability of misclassification and the classification threshold. Figure 12b compares this plot with the reference distribution used for this case. The probability of misclassification is proportional to the frequency of the classification threshold in the reference distribution. This is further illustrated in Figure 13 which shows the probability of Type I and Type II errors vs. data density for different cutoffs. Again, the probability of misclassification is low for cutoffs of 0.25 and 2.0 and high for a cutoff of 1.0. This is due to the increased frequency of the value 1.0 in the reference distribution.

Monotonic Non-Parametric Curve Fitting

The uncertainty measures calculated are not always monotonic increasing or decreasing as they should be. For this reason, it is useful to fit a line of best fit to the results. It was determined that a non-parametric fit gives the best results. In order to ensure that the results are either monotonically increasing or decreasing, a fitting is proposed. The fitting is non-parametric. This discussion will assume we are fitting a decreasing measure. Increasing measures can be fit using the same method by making all the values of the dependent variable negative. This makes the measure decreasing.

The monotonic non-parametric fitting (MNPF) proceeds as follows:

1. Find the 'bottom' boundary starting with the first two points, search forward for a pair of points which are increasing, that is, where the second point in the pair is greater than the first. Where this occurs, set the bottom value for the second point equal to the first. Do this for all points. This process creates the red line shown in Figure 14b.
2. Find the 'top' boundary starting with the last two points and working backward, following the same procedure given above to get the green line shown in Figure 14b. Note that the top and bottom boundaries are equal between the first four points.
3. Remove plateaus by searching forward again, this time looking for points where the top and bottom boundary are not the same. The first time this occurs, set the new bottom boundary as 75% of the way between the current bottom and top boundary. If the top and bottom boundaries for the next point are also not equal, set the new bottom boundary for this point at 25% of the distance between the current bottom and top. This procedure creates the orange line shown in Figure 14c.
4. Continue to remove plateaus by searching backward, again looking for points where the top and bottom boundaries are different. When this is first encountered, set the new top at 25% of the distance between the current bottom and top. At the next point, set the new top at 75% of the distance between the current bottom and top. This procedure produces the blue line shown in Figure 14c.
5. Repeat steps 3 and 4 until the top and bottom boundaries are equal at every point. The end result is the black line shown in Figure 14d.

This procedure produces a series of line segments which will always be decreasing. This means that any uncertainty measures interpolated from the data will not inappropriately increase for increased data density. The procedure described above has been implemented in the subroutine mnpf.

Conclusion

Data density is an important consideration for natural resource developers. Practitioners are constantly faced with the question of how much data is enough. It is useful to examine how various measures of uncertainty vary with data density in determining an appropriate data density. No unified tool for determining uncertainty measures for various data densities was previously available. A tool for calculating various measures of uncertainty at different data densities has been presented. The uncertainty results are sensitive to a number of input parameters. The response of the uncertainty measures to changes in the input parameters was demonstrated. These examples have shown that the relationship between uncertainty and data density depends mainly on the variogram and histogram.

References

Aspie, D, & Barnes, R.J. 1990. *Infill-sampling design and the cost of classification errors.* Mathematical Geology. V.22. No.8.

Back, P.E. 2007. *A model for estimating the value of sampling programs and the optimal number of samples for contaminated soil.* Environmental Geology. V.52. No.3.

Carrera, J., Usunoff, E., & Szidarovszky, F. 1984. *A method for optimal observation network design for groundwater management.* Journal of Hydrology. 73.

Criminisi, A., Tucciarelli, T., & Karatzas, G.P. 1997. *A methodology to determine optimal transmissivity measurement locations in groundwater quality management models with scarce field information.* Water Resources Research. V.33. No.6.

Deutsch, C.V, & Journel, A.G. 1998. *GSLIB: Geostatistical software library and user's guide.* Oxford University Press. 2nd Ed.

Fiering, M.B. 1965. *An optimization scheme for gaging.* Water Resources Research. V.1. No.4.

Hubbard, D., 2007. *How to Measure Anything: Finding the Value of Intangibles in Business.* John Wiley & Sons.

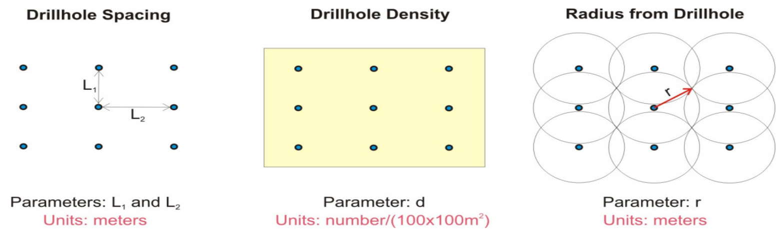
Hughes, J.P., & Lettenmaier, D.P. 1981. *Data requirements for kriging: estimation and network design.* Water Resources Research. V.17. No.6.

Loaiciga, H.A. 1989. *An optimization approach for groundwater quality monitoring and network design.* Water Resources Research. V.25. No.8.

McBratney, A.B., Webster, R., & Burgess, T.M. 1981. *The design of optimal sampling schemes for local estimation and mapping of regionalized variables – I.* Computers and Geosciences. V.7. No.4.

Rouhani, S. 1985. *Variance reduction analysis.* Water Resources Research. V. 21. No.6.

Virdee, T.S., & Kottegoda, N.T. 1984. *A brief review of kriging and its application to optimal interpolation and observation well selection.* Hydrological Sciences. 29 4



$$d = \frac{10000}{L_1 \cdot L_2}$$

$$\left(\frac{L_1 + L_2}{2}\right) = \sqrt{\frac{10000}{d}}$$

$$L_1^2 + L_2^2 = (2r)^2$$

$$r = \sqrt{\frac{L_1^2 + L_2^2}{4}}$$

Figure 1: Illustration of the geometric measures drillhole spacing, drillhole density, and distance to nearest drillhole.

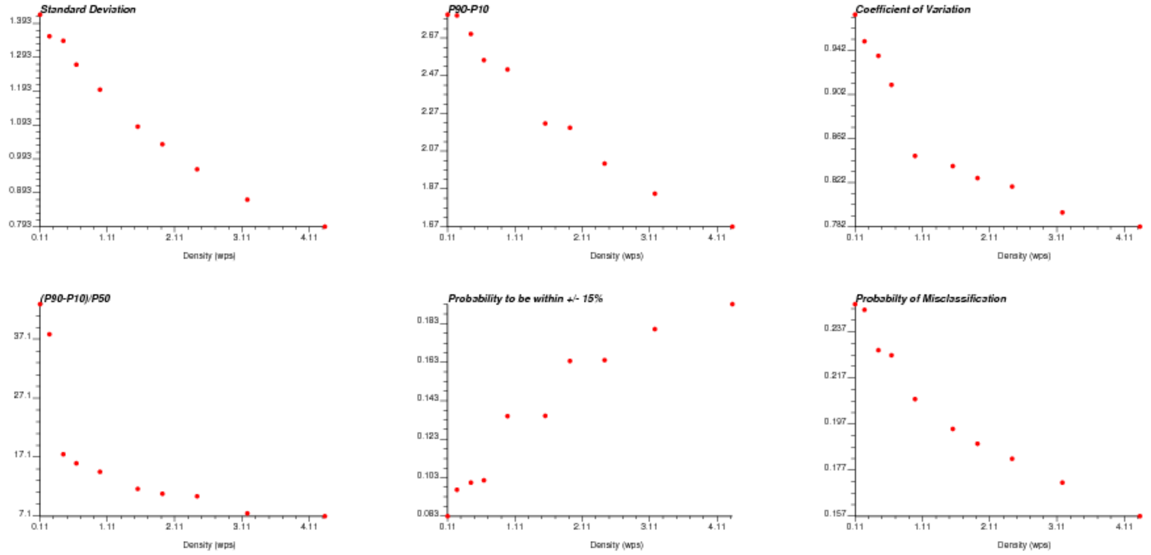


Figure 2: Behaviours of the uncertainty measures with respect to data density: a) standard deviation; b) P90-P10; c) coefficient of variation; d) (P90-P10)/P50; d) probability to be in 15% of the truth; f) probability of misclassification.

```
                        Parameters for DATSPACSIM
                        *************************
        START OF PARAMETERS:
        dataspacing.out                 -uncertainty measures output
        dataspacing.fit                 -mnp fitting output file
        red.dat                         -file with reference distribution
        5    0                          -  columns for vr and wt
          -1.0e21  1.0e21               -  trimming limits
        0.0   5.0                       -  zmin, zmax(tail extrapolation)
        1      0.0                      -  lower tail option, parameter
        1      5.0                      -  upper tail option, parameter
        4     4                         -point estimate spacing in x, y
        12  12                          -block size (multiple of point estimate spacing)
        21  21                          -# of samples in x, y
        3                               -number of data spacings to consider
        5     5                         -spacing in x, y (multiple of point estimate spacing)
        7     7                         -spacing in x, y (multiple of point estimate spacing)
        10  10                          -spacing in x, y (multiple of point estimate spacing)
        1     1.0                       -vertical grid: nz, zsiz
        10                              -number of reference models (per spacing)
        100                             -number of realizations per reference model
        15                              -% sampling error st.dev.
        1   1.5                         -critical threshold: option, value(eg. cutoff grade)
        15                              -confidence interval to report (%)
        69069                           -random number seed
        1     0.0                       -nst, nugget effect
        1   1.0   0.0      0.0      0.0 -it,cc,ang1,ang2,ang3
                   10.0     10.0    10.0 -a_hmax, a_hmin, a_vert
```

Figure 3: Parameters for DATSPACSIM Program

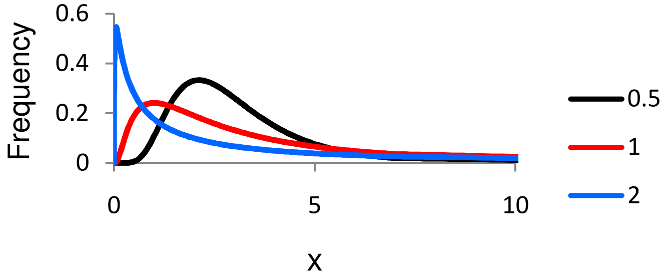


Figure 4: Lognormal reference distributions with mean of 1.0 and standard deviation ranging from 0.5 to 2.0.

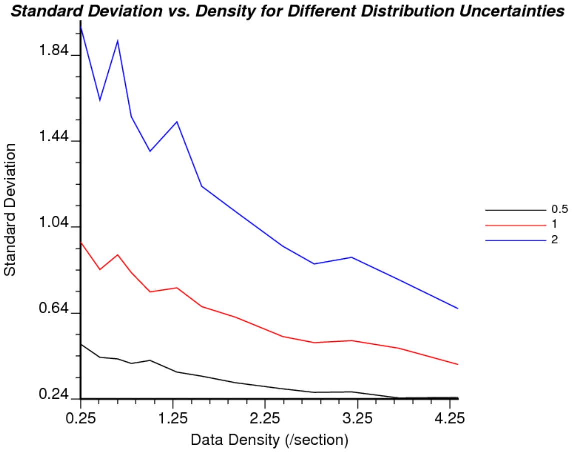


Figure 5: Uncertainty results vs. data density for the reference distributions shown in Figure 4.

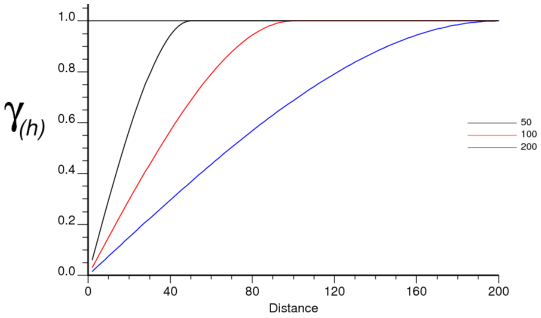


Figure 6: Input isotropic variograms with one spherical structure, no nugget, and ranges varying from 50 to 200m.

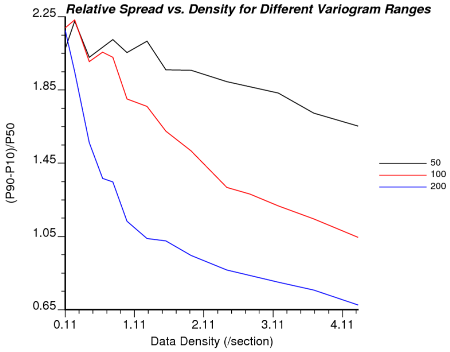


Figure 7: Relative spread vs. data density for the variograms shown in Figure 6.

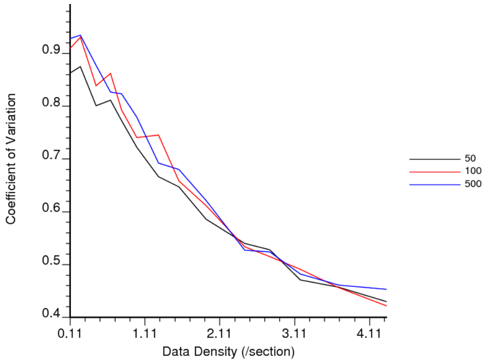


Figure 8: Uncertainty results for different numbers of realizations.

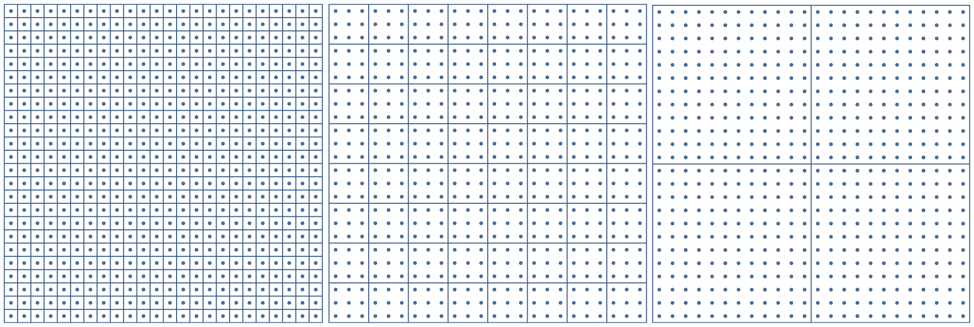


Figure 9: Different block sizes used for upscaling the point estimates: 4m, 12m, and 48m.

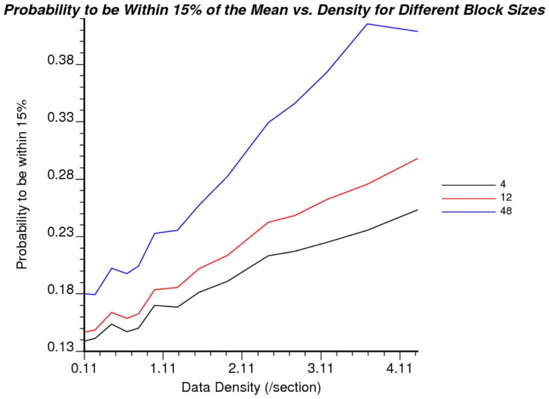


Figure 10: Precision vs. data density for block sizes of 4, 12, and 48m.

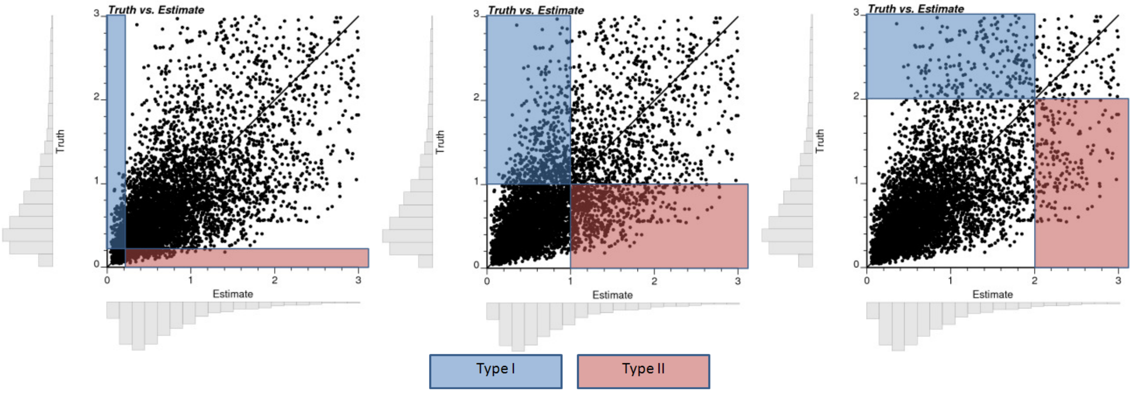


Figure 11: Type I and Type II errors for cutoff values of 0.25, 1, and 2 respectively.

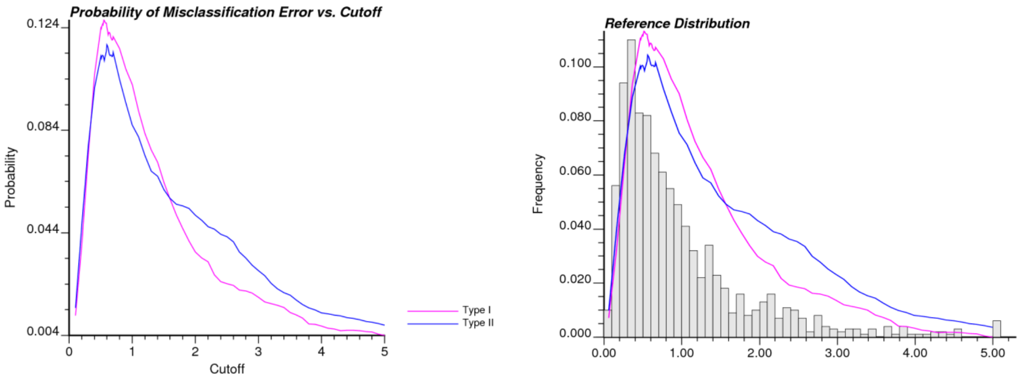


Figure 12: Relationship between misclassification error, cutoff, and the reference distribution.

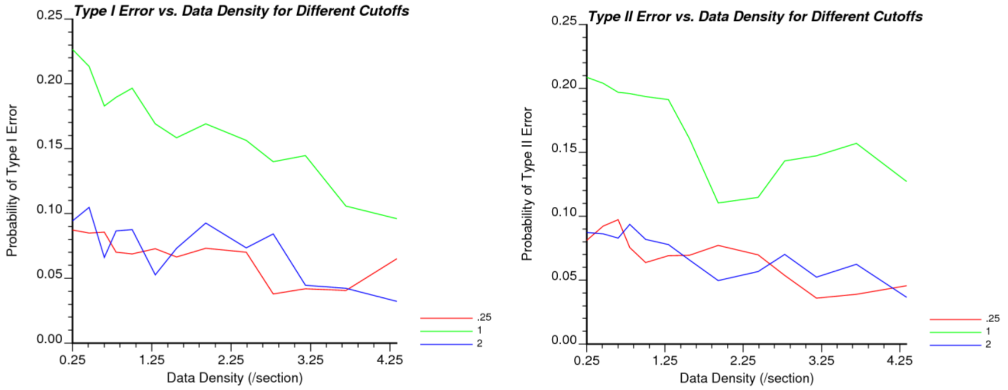


Figure 13: Probability of type I and type II errors vs. data density for different cutoffs.

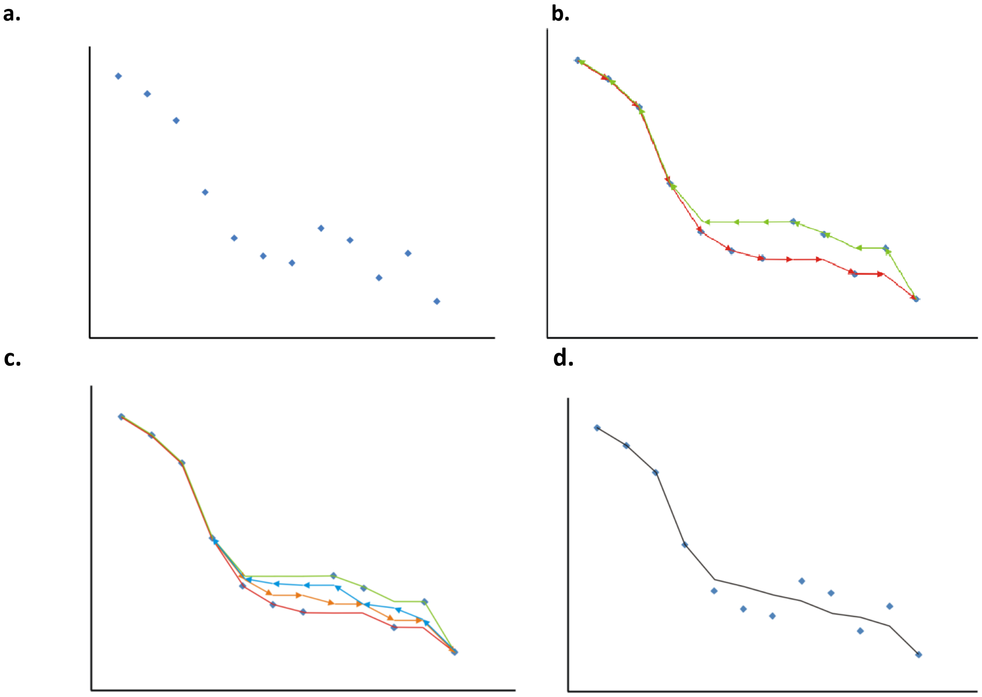


Figure 14: Illustration of the monotonic non-parametric fitting process: a) points to be motonically fit; b) determination of the top and bottom boundaries; c) plateau removal; d) final fit.