

Display of Cross Validation / Jackknife Results

Clayton V. Deutsch

Numerical models may be used for important decisions. They should be checked to the maximum extent possible. Cross validation and the jackknife are applicable when there are sufficient data. They resample from the available data or consider new data not used in modeling. They are well established in geostatistics, but there is a variation in how results are presented. This short note presents some suggestions and software for analysis of cross validation results.

Introduction

Careful checking of geostatistical models is best practice. Sometimes there are so few data that there is unavoidable reliance on a conceptual model; it may not be possible to perform many quantitative checks. At a minimum, the data and chosen statistical parameters must be reproduced and used correctly by the estimated model or simulated realizations.

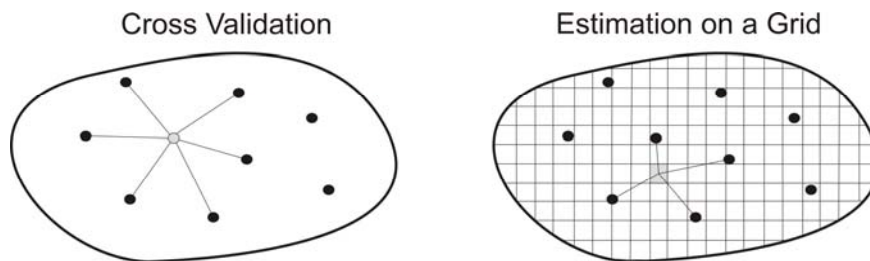
Perhaps the best check is to compare the predicted values to true values not used in the geostatistical model. Cross validation mimics this by leaving the data out one at a time. Sometimes set(s) of data are left out from the very beginning or new data become available providing a more flexible checking – sometimes called the jackknife (although the names are used interchangeably). Ideally, the checking would consider completely new data or data held back from the beginning, but that may be impossible or too time consuming. In any case, the pairs of true values and geostatistical predictions are compared with the goal to improve prediction and fairly assess the quality of the geostatistical model. This paper is concerned with tools to view and compare true values and predictions.

Cross validation and the jackknife generate the same type of results: pairs of true values and predictions. The variable under consideration could be continuous or categorical. The geostatistical prediction could be a single estimate or a distribution of uncertainty. Different tools are required to evaluate the results in estimation mode and probabilistic mode. This paper presents some useful tools.

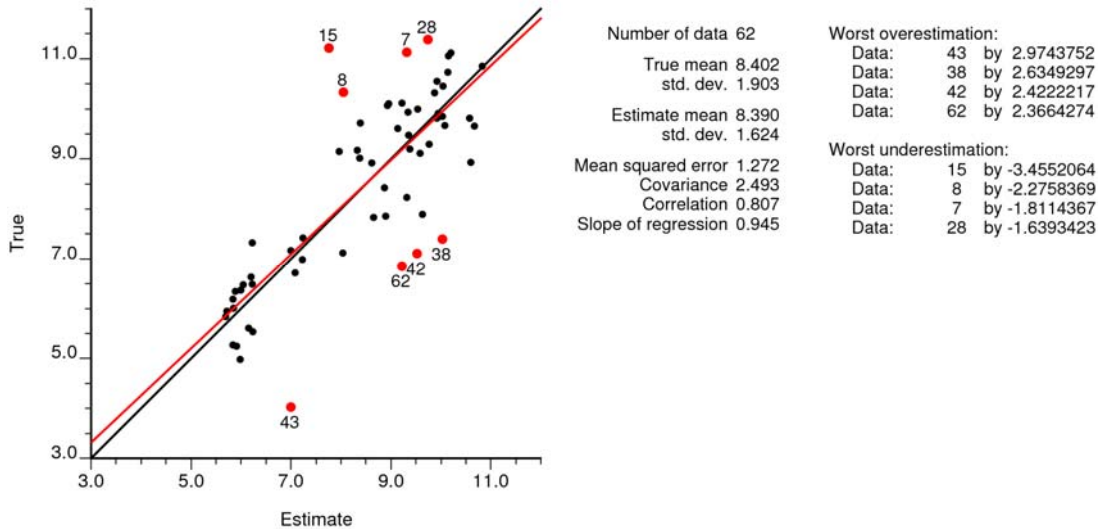
The ultimate goodness of a numerical model is impossible to assess. The joint distribution at all locations is important; checking one location at a time with cross validation or the jackknife does not provide much information on the joint distribution. In special cases with many data, it may be possible to check higher order distributions, but there is only one true distribution that is rarely sampled completely.

Estimation

Estimation remains an important task in geostatistics. The goal is to calculate the best estimate at unsampled locations. Implementing cross validation or the jackknife provides pairs of true values and estimates: $(z_i, z_i^*, i=1, \dots, n)$. The comparison of these values often leads to a pessimistic assessment of model performance. In general, the *real* unsampled locations have the benefit of all data and the data spacing for estimation on a grid will be less. Consider the sketch below. The distance to the data is quite large in cross validation; consider leaving one data out at a time. When estimating on a grid, some of the locations are right by the data locations and even in between the data, the data are closer than in cross validation. The disparity between data spacing in checking and real estimation may be even larger in the jackknife. Careful evaluation of the results is still relevant; blunders and problem data could be detected and the overall quality of prediction is assessed.



Let's consider continuous variables first. The most informative display is a cross plot between the pairs of estimates and true values. The `scatxval` program creates this plot. Any standard cross plotting software could be used, but there are some important considerations in making a good display: (1) the estimate should be the independent variable (X or abscissa axis) because we will have access to the estimates and not the truth, (2) the units of both axes should be the same and the 1:1 line should be shown, (3) the regression of the truth on the estimates should be shown to provide an indication of conditional bias, (4) relevant summary statistics should be shown, and (5) the worst cases of over- and under-estimation should be identified with the data ID so that they can be checked. Following is an example.



Checking summary statistics alone would not be as informative as checking the plot. We are looking for patterns or particular values that are anomalous. Some summary statistics are of interest. The means of the true and estimated values tells us of any systematic bias; they should be the same. The standard deviations of the true and estimated values tells us of the smoothing effect. The mean squared error (MSE) is a common summary of prediction performance; the MSE should be small. The covariance between the estimate and truth is another useful measure; it should be as high as possible for a good estimate. The correlation coefficient is also useful, but it is sensitive to smoothing and a low variance of the estimates. A smooth estimator will have a higher correlation with the truth, but that is not necessarily a good thing. The slope of the regression line should be close to one. A slope less than one would indicate conditional bias. This may be acceptable for interim estimates where smoothing is an even larger problem. The practitioner will have to assess these statistics in the context of their problem.

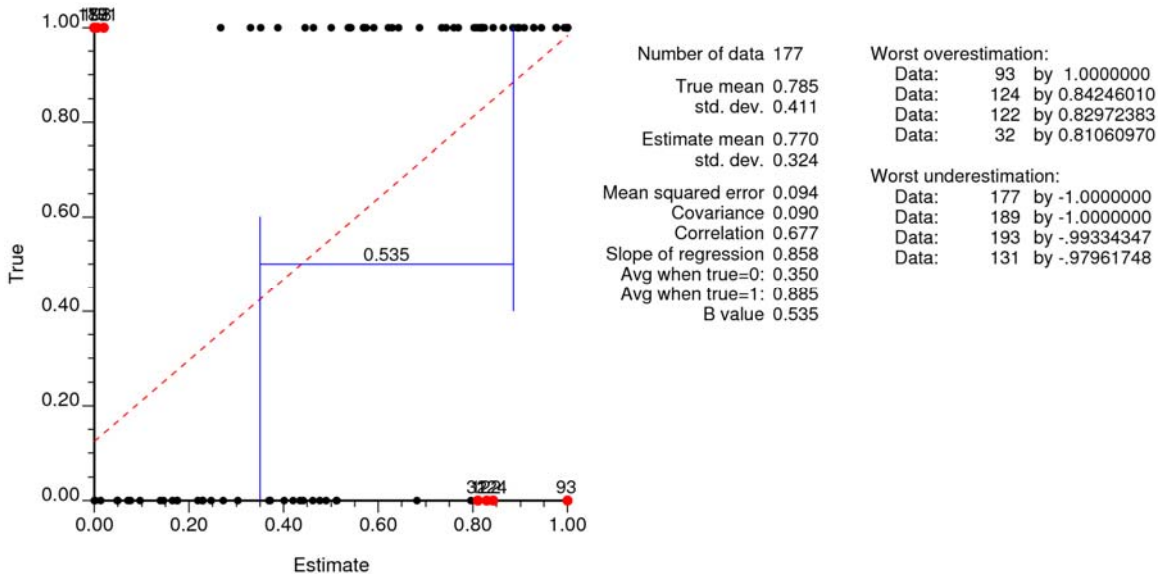
The program `scatxval` creates plots like that shown above. This is a GSLIB-like program. An example parameter file is given below. Notice that there are relatively few parameters. The pairs of true and estimates values must be in one data file. If there is no ID column, then the data line number will be used. The minimum and maximum are hardcoded to 0 to 1 for categorical variables. The number of worst estimates to show should not be set too big or the plot will get messy.

```

1           Parameters for SCATXVAL
2           *****
3
4  START OF PARAMETERS:
5  kriging.dat           -file with cross validation results
6  4  5  0              - columns for true, estimate, ID
7  -1.0  1.0e21        - trimming limits
8  scatxval.ps         -file for Postscript output
9  0                   -0=continuous, 1=categorical
10 3.0 12.0  0         -minimum, maximum, (0=arith, 1=log)
11 4                   -number of worst estimates to show

```

Categorical variables are different; they are considered in an indicator formalism. The true values are only 0 or 1. The estimates are continuous values between 0 and 1. The plot is somewhat different. Some considerations include: (1) the 1:1 line is not relevant, (2) the regression line is interesting, but not as relevant as with continuous variables, and (3) the average of the estimated probabilities conditional to the true values are important – see the blue lines. The difference between those averages – the “B” value reminiscent of the the “B” calibration parameter in Markov-Bayes – is an important measure of the quality of the estimates. A large value is good – that means the presence/absence of the category is predicted correctly.



There are other displays that could be useful including a histogram of errors and a posting of the errors on maps and sections, but the cross plots created by scatxval should prove useful.

Probabilistic

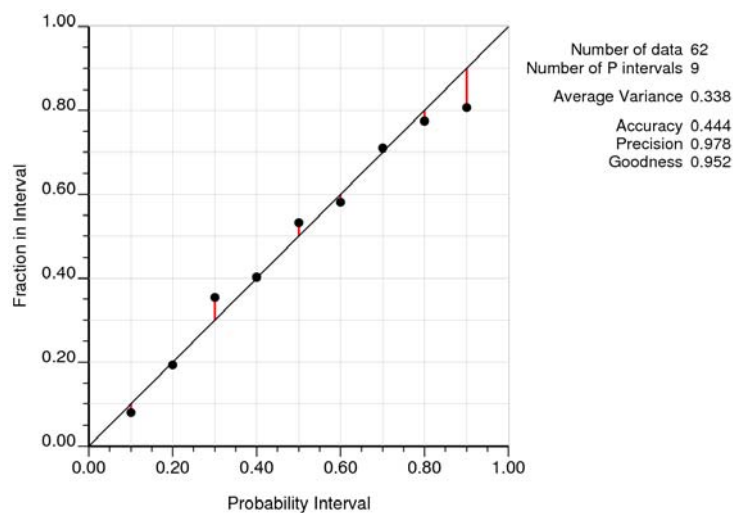
Checking the actual fraction of true values that fall within specified probability intervals is useful. The distributions of uncertainty are specified by a conditional mean and variance in Gaussian units.

The program `accplt_ns` was prepared to create an accuracy plot when the true values and conditional distributions are in Gaussian (or normal score units). There are few parameters (see below). The true values and conditional distributions must be in one file. The conditional distributions are specified by a mean and variance coming from the normal equations or simple kriging. Often, these results are from `kt3d`. The probability increment is the bin width for the probability values. Often, 5% or 10% intervals are reasonable. There are two output files: a PostScript plot and a text file that contains the quantile values of the true values.

```
START OF PARAMETERS:
kt3d.out          -file with true values and cond. dists.
4 5 6            - columns for true, mean, variance
0.10            -probability increment (0.05 or 0.1)
accplt_ns.ps     -file for accuracy plot
accplt_ns.dat    -file for text output
```

An example plot is shown below. The axes are always scaled between 0 and 1. A grid of light gray lines shows the probability intervals. The red lines and bullets show the deviations of the actual probabilities from the target probability interval. The number of data entering the calculation and the number of probability intervals are noted to the right. The average variance is shown as a summary of the

local uncertainty. The accuracy, precision and goodness statistics are not particularly helpful but they are reported.



Conclusion

Cross validation and the jackknife provide useful checks on the geostatistical modeling methodology. They do not provide unequivocal proof that the model is good or fit for a particular purpose; nevertheless, they identify blunders and problem data. At times, these results can be used to compare different algorithms and implementation choices.