A Statistical Resampling Program for Correlated Data: `Spatial_Boostrap`

Clayton V. Deutsch

Department of Civil & Environmental Engineering
University of Alberta

Abstract

The bootstrap resampling procedure is widely used to quantify uncertainty in statistical parameters. The two most important assumptions are that (1) the initial data distribution is representative of the entire population and (2) the data are independent. This short note presents a bootstrap program that resamples with correlation, which relaxes the second assumption of independence. A LU simulation algorithm is used to simulate values under a multivariate Gaussian model. The program is extremely fast and suitable for virtually all data sets. The program parameters are documented thoroughly and a number of synthetic and real examples demonstrate the place of the program.

Introduction

Geostatistical realizations are being used increasingly for uncertainty quantification. Uncertainty in global statistical parameters such as the mean porosity within facies is an important component of uncertainty quantification. Simulated realizations that use fixed input parameters can significantly underestimate global uncertainty. This underestimation of uncertainty is especially significant when the reservoir is large relative to the variogram range; the global histogram is reproduced very closely by each realization. Uncertainty in the global mean and other parameters should be accounted for by (1) establishing their distributions of uncertainty, and (2) generating each geostatistical realization with different mean values and other input parameters that are sampled from their distributions of uncertainty. The bootstrap procedure is considered for the first step – establishing distributions of uncertainty in input statistics.

The bootstrap is a popular application of Monte Carlo simulation technique and was developed by Efron (1983). The bootstrap is a statistical resampling technique that permits the quantification of uncertainty in statistics by resampling from the original data, in other words, "pulling yourself up by your bootstraps." In the simplest case, consider n data values of a single variable (z_i, $i=1,...,n$) and a calculated statistic, say, the experimental mean m_z. The bootstrap can be used to calculate the uncertainty in the calculated statistic by the following simple procedure:

1. Assemble the representative distribution of the Z random variable using declustering and debiasing techniques if appropriate: $F_Z(z)$. This distribution could simply be the equal weighted histogram of the n data.
2. Draw n values from the representative distribution, that is, generate n uniformly distributed random numbers p_i, $i=1,...,n$ and read the corresponding quantiles: $z_{s,i} = F_Z^{-1}(p_i)$, $i=1,...,n$. The number of data drawn is typically equal to the number of data

available in the first place. The distribution of simulated values is not identical to the initial data distribution because they are drawn randomly and with replacement.

3. Calculate the statistic of interest from the resampled set of data, say, the experimental mean m_{sz}.
4. Return to step 2 and repeat many times, say, $L=1000$.
5. Assemble the distribution of uncertainty in the calculated statistic.

The result is trivial in the case of the mean since we know (1) that the average of independent identically distributed samples tends to a normal distribution, and (2) the variance of an average of independent random variables of variance σ^2 is σ^2/n, where n is the number of samples in the average. There is no need to bootstrap the mean. The bootstrap gets particularly interesting for more complex statistics such as the proportion above a cutoff, the average above a cutoff, the correlation between two variables, the regression relationship between multiple variables, and so on.

The bootstrap assumes that the data are independent one from another and representative of the underlying population. The independence assumption may be acceptable early in reservoir appraisal with widely spaced wells; however, highly correlated data (such as nearby measurements along a well) and close wells do not meet this assumption. The bootstrap procedure must be modified to account for the spatial correlation in the input data.

A number of people have proposed a spatial bootstrap over the years. Andy Solow proposed an approach in the early 1980s. André Journel proposed a method in a SCRF paper in the mid 1990s. Other workers in geostatistics have almost certainly hacked together some code for this over the years. This paper presents some clean GSLIB-like code based on an efficient matrix simulation approach.

Methodology

There is a need to account for the spatial correlation between the data when simulating n values from the distribution of n data. Any number of geostatistical simulation techniques could be used. The simulated values are unconditional and are only required at the data locations. The LU method was chosen because (1) it is simple and efficient for a large number of realizations, and (2) the number of data would typically not exceed 10000 or so; sequential methods are not warranted. A 3-D variogram model is required to establish the covariance values between each pair of data. Implicit assumptions to the LU method are multivariate Gaussianity and stationarity, which are ubiquitous in geostatistics.

We are to simulate n values from the deemed representative histogram $F_Z(z)$ following a known 3-D variogram model $\gamma(\mathbf{h})$, which is the variogram of the normal scores of the Z variable. The basic algorithm is to perform an LU decomposition of the n by n covariance matrix:

$$\mathbf{C} = \mathbf{L}\,\mathbf{U} \tag{1}$$

Where **C**, **L**, and **U** are n by n matrices. **C** is built directly from the variogram model. **L** and **U** are lower and upper triangular matrices calculated by a Cholesky LU decomposition; the code from Numerical Recipes is convenient. Unconditional Gaussian simulations are quickly calculated by a simple matrix multiplication:

$$\mathbf{y}^{(l)} = \mathbf{L}\,\mathbf{w}^{(l)},\ \ l=1,\ldots,L \tag{2}$$

where **w** is a *n* by 1 vector of independent Gaussian values and **y** is the resulting *n* by 1 vector of unconditionally simulated values with the correct covariance. The superscript $l=1,...,L$ denotes one of the large number, *L*, of realizations we are generating. These Gaussian values can be converted to probability values to draw from the representative distribution.

$$\mathbf{p}^{(l)} = G(\mathbf{y}^{(l)}),\ l=1,...,L \tag{3}$$

Where $G^{-1}(\cdot)$ is the inverse of the standard normal distribution and **p** is an *n* by 1 vector of probability values [0,1]. The drawn z-values are calculated as:

$$\mathbf{z}^{(l)} = F_Z^{-1}(\mathbf{p}^{(l)}),\ l=1,...,L \tag{4}$$

Once the LU decomposition is performed (required only once), the generation of the simulated realizations can be calculated simply as: $\mathbf{z}^{(l)} = F_Z^{-1}(G(L\ \mathbf{w}^{(l)}))$, $l=1,...,L$.

This procedure is straightforward to code and can be implemented very simply. Different statistics can be calculated from each set of simulated values, e.g., the mean, the proportion above some critical cutoff, the mean above cutoff, and so on. The distribution of these statistics can be used for later geostatistical modeling or directly for uncertainty calculation.

Correlation in the *n* values will lead to greater uncertainty than the assumption of independence. There are effectively less independent data when the data are correlated. In fact, we can calculate the effective number of data as:

$$n_{eff} = \frac{\sigma_Z^2}{\sigma_{\bar{Z}}^2} \tag{5}$$

Where σ_Z^2 is the variance of the data values and $\sigma_{\bar{Z}}^2$ is the variance of the average values. The program that implements the simulation procedure will be described next; followed by some examples.

Program

The `Spatial_Bootstrap` program follows standard GSLIB conventions. Most of the functions are available in GSLIB. Two source code files are required: `Spatial_Bootstrap.for` and `Spatial_Bootstrap_Subs.for`; the subroutines have been collected to facilitate compilation if the compiled GSLIB library is not available. The parameters for the program:

```
Line   START OF PARAMETERS:
 1     example/cluster.dat                -file with reference distribution
 2     3   0                              -   columns for value and weight
 3     -1.0e21    1.0e21                  -   trimming limits
 4     example/cluster.dat                -file with locations to resample
 5     1   2   0                          -   columns for X, Y, Z locations
 6     1                                  -save realizations? (0=no, 1=yes)
 7     example/SB-realizations.out        -file for all simulated realizations
 8     example/Spatial_Bootstrap.out      -file for bootstrapped average
 9     0.0                                -threshold (average above is reported)
10     100                                -number of realizations
11     112063                             -random number seed
12     1    0.2                           -nst, nugget effect
13     1    0.8  0.0   0.0   0.0          -it,cc,ang1,ang2,ang3
14               10.0  10.0  10.0         -a_hmax, a_hmin, a_vert
```

The reference distribution (**Lines 1**, **2**, and **3**) specifies the distribution of data for the bootstrap. The reference distribution could the data at the locations being bootstrapped. The locations are specified on **Lines 4** and **5**. The bootstrapped realizations can be saved (optionally) to a data file (**Lines 6** and **7**). The bootstrapped average, the proportion above the specified cutoff, and the average above cutoff are saved in the second output file – **Line 8**. The cutoff is specified in **Line 9**. This cutoff can be used to define the threshold between net and non-net. The number of realizations is specified in **Line 10**. The random number seed is in **Line 11**. **Lines 12** and greater specify a 3-D variogram model in standard GSLIB conventions.

The program writes two output files: the simulated values and the bootstrap statistics. The statistics of the input data and the resampled average are written the screen. The effective number of data is also written to the screen. The LU simulation approach is very fast when there are 100s of data. The program will take minutes of CPU time (versus seconds) when there are 1000s of data. All memory allocation is dynamic so, theoretically, any number of data could be considered; however, the data should be subset or blocked-up to a larger scale if more than 5000 data are considered. The CPU time will be large and numerical precision problems can become important.

Examples

Figure 1 shows the results using `cluster.dat` from GSLIB. The location map and histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right. A single structure spherical variogram with a range of 10 distance units was used for the spatial bootstrap. The variance assuming independence has gone down almost exactly by 1/140 (44.92 to 0.3109); however, the variance in the case of the spatial bootstrap decreased to 0.9888, which is more than three times higher than when independence is assumed. The effective number of data is calculated as 45. We see how the spatial bootstrap accounts for correlation in the data.

The second example is with two wells through a reservoir that is about 100m thick; there are 212 data spaced about 1m. The two wells are 600m apart. The two wells are essentially uncorrelated, but the data within each well are highly correlated. Figure 2 shows the profile of porosity in the two wells and a histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right. The vertical variogram is fitted and is shown at the bottom. There are certainly not 212 independent data; in fact, the effective number of data is calculated as 14.

The third and final example is based on the well-used Amoco data. There are 62 wells through a reservoir that is about 70 feet thick. There are 3303 data. There are certainly not 3303 independent data, but there are probably more than 62 given the thickness of the reservoir relative to the vertical variogram range. Figure 3 shows a horizontal slice through the data and a histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right. The 3-D variogram points and model are shown on Figure 4. There are certainly not 3303 independent data; `Spatial_Bootstrap` calculated that there were 72, which is slightly more than the 62 wells. There is some horizontal correlation between the wells, but the vertical thickness of the reservoir implies more than 1 independent data per well.

Application Notes

The bootstrap is typically applied within deemed homogeneous geological rock types to establish uncertainty in parameters such as the mean, variance, and distribution shape. The uncertainty in the rock types and their proportions must be considered separately.

It would be straightforward to simulate multiple correlated values by drawing *sets* of values with correlation. The correlated probability values are used to pick up data *indices* and not *z*-values from the representative histogram $F_Z(z)$.

Conclusions

The bootstrap is a useful statistical tool to establish uncertainty in input parameters based on resampling the available data. The conventional bootstrap assumes data independence, which is unrealistic. The `Spatial_Bootstrap` program is a useful tool for the geostatistician performing uncertainty analysis. There remain strong assumptions of stationarity and representative input data distributions.

References

- Deutsch, C.V. 2002, *Geostatistical Reservoir Modeling*. Oxford University Press, New York.
- Deutsch, C.V. and Journel, A.G., 1998. *GSLIB: Geostatistical Software Library: and User's Guide*. Oxford University Press, New York, 2nd Ed.
- Efron, B., 1982, *The Jackknife, the Bootstrap, and other Resampling Plans*, Society for Industrial and Applied Math, Philadelphia,
- Efron, B., and R. J. Tibshirani, 1993, *An Introduction to the Bootstrap,* Chapman & Hall, New York.

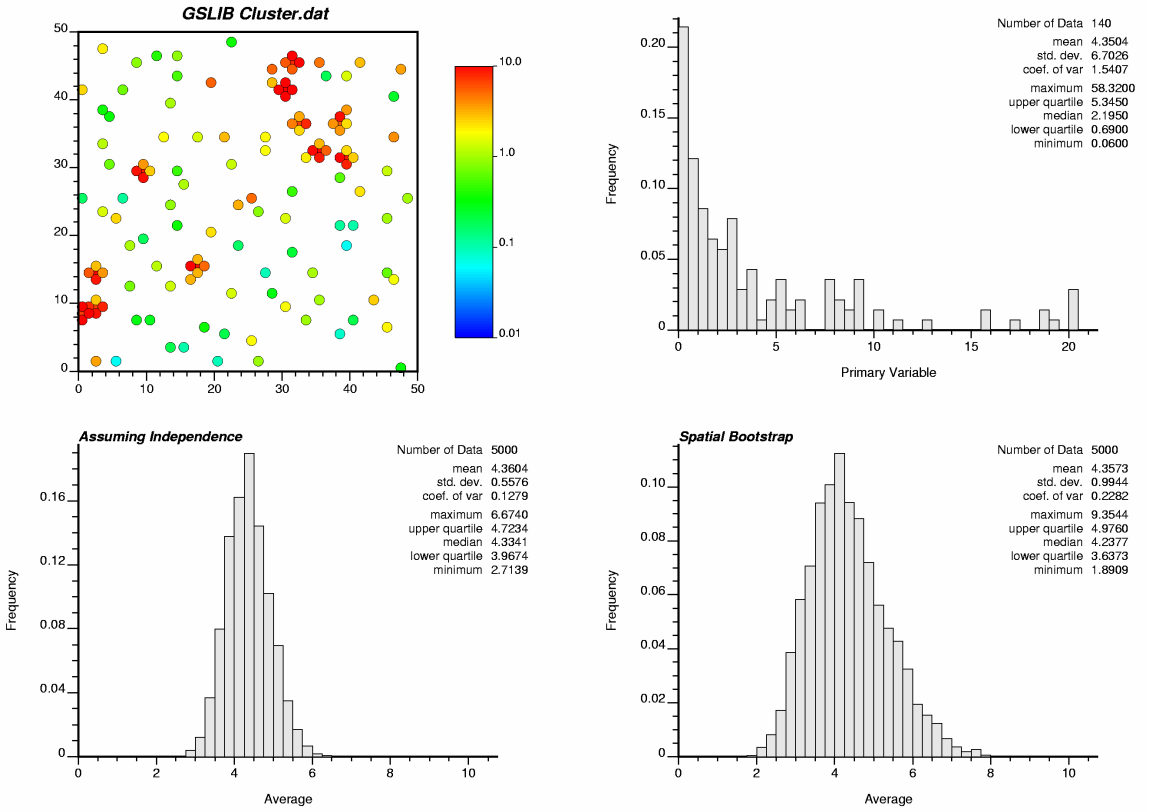

Figure 1: example with cluster.dat from GSLIB. The location map and histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right.

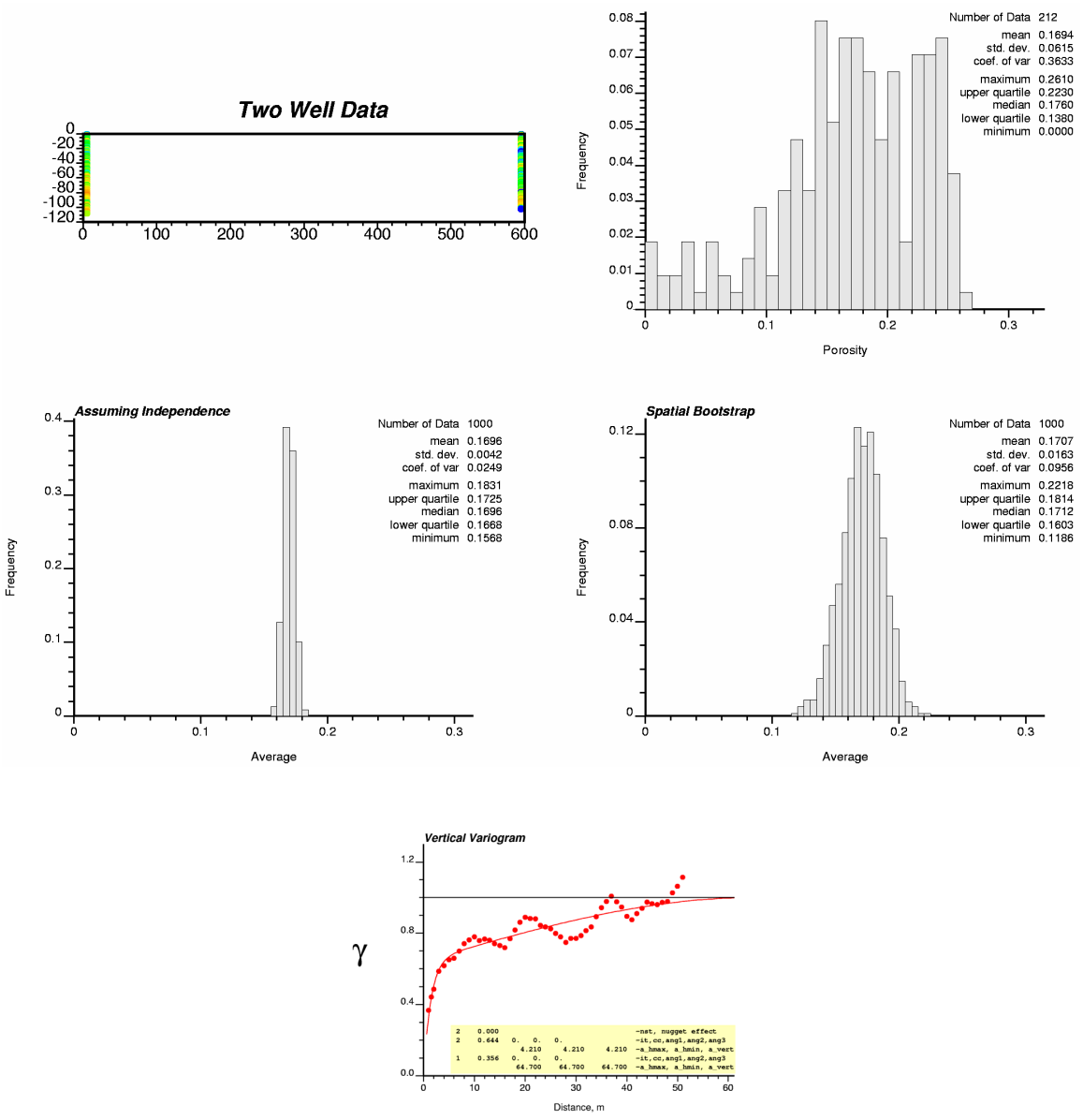

Figure 2: example with twowell.dat. The profile of porosity in the two wells and a histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right. The vertical variogram is fitted and is shown at the bottom.

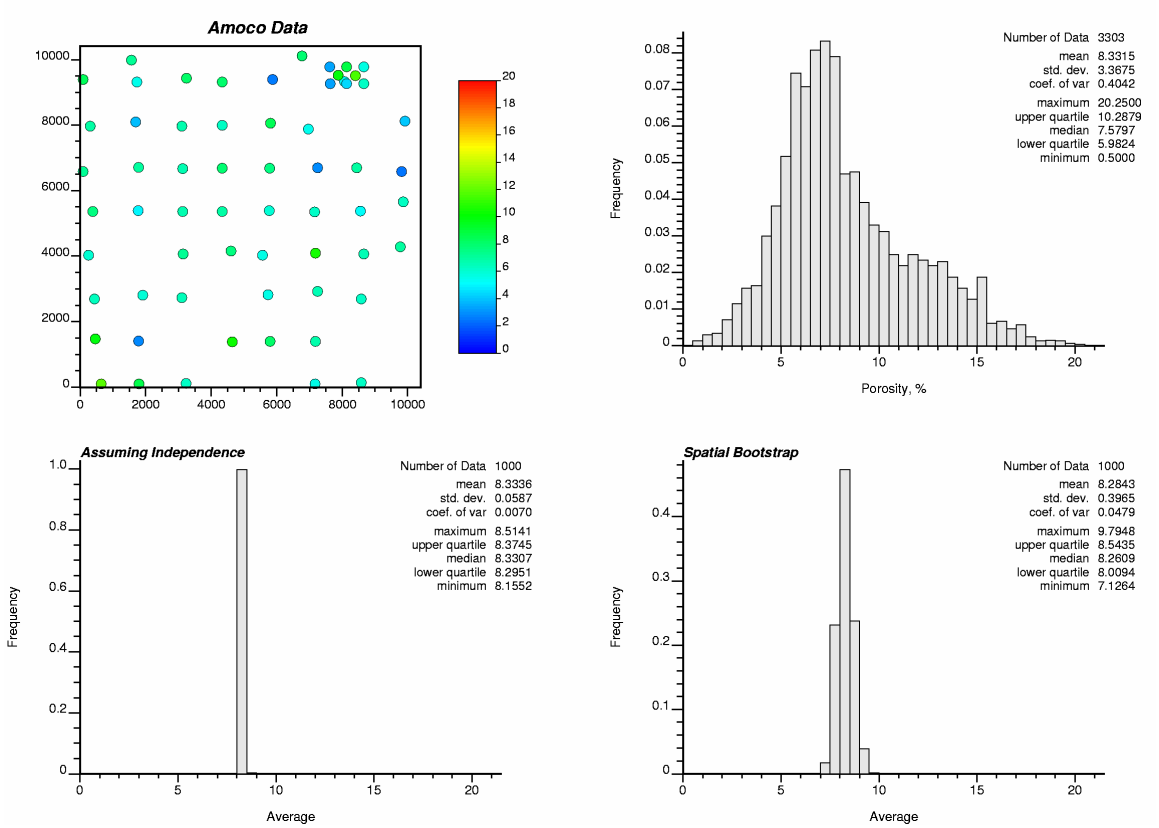

Figure 3: example with Amoco data. A slice through the data and a histogram of the data are shown on the top. The uncertainty in the mean assuming independence is shown in the lower left. The spatial bootstrap results are shown in the lower right.

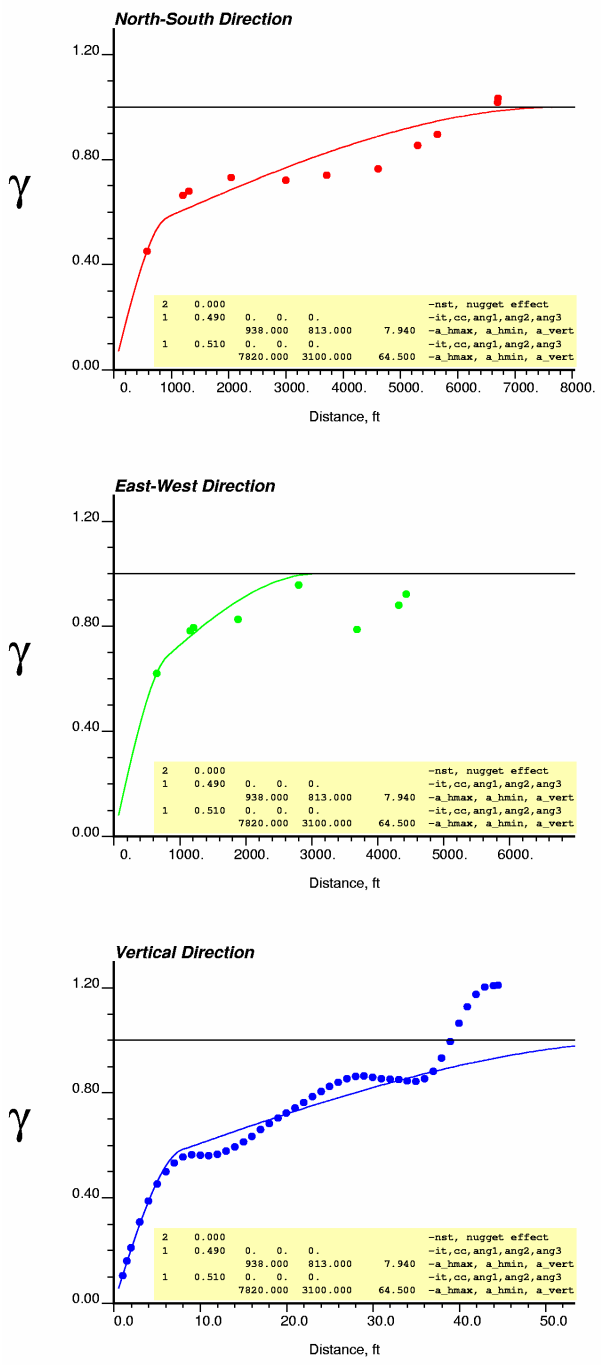

Figure 4: 3-D variograms for the Amoco data. The horizontal anisotropy N-S versus E-W was found by visual inspection. The variograms were fit with `varfit`.