Conditional Simulation

Conditional simulation is the practice of using a fitted geostatistical model to repeatedly simulate whole potential surfaces. Why would we want to do this? It is much simpler to just predict, with standard errors, at each point. Instead, conditional simulation requires that we make full use of the joint distribution of all prediction locations so that we simulate the entire surface that accounts for the correlation among the prediction locations.

As an example, let us suppose that the data in the toy example in Fig. 9.3 represents water depths, on the log scale, across a body of water. Our job is to lay a cable across that body of water, and to do so, we must order the cable to be manufactured. We do not want to have a cable that is too short, as that would require a whole new cable to be built. On the other hand, we do not want the cable to be too long (we are assuming it can be easily shortened after laying it), as that would cost more money. We might consider the length of the kriging line in Figure 9.3 to be our best estimate. However, this would be a mistake. The kriging line is "smoother" than the actual data. That is, you can imagine that any actual value, if we were able to observe it, along the kriging prediction line would be above or below that line, and taken all together, the cable would need to be longer than the kriging line. Moreover, we would rather have the cable be a little longer than needed rather than a little shorter. A conditional simulation attempts to create a surface that could have been observed, often called a "realization" of the surface. In our example, this surface is a line, and the conditional simulation of a line is much rougher than the kriging prediction line, and if we measure the length of the conditional simulation line, it would be longer than the kriging line. Now, if we create 1000 such lines, each a different and equally probable realization, then we can compute the length of cable needed for each realization. From these 1000 lengths, we can compute an empirical distribution of lengths, and from these, choose a value that matches our relative risk for ordering a cable that is too short versus too long. For example, if we wanted to be 95% certain that our cable was long enough, and were able to tolerate that extra expense, we would choose the 950th ordered value from the 1000 lengths computed from the realizations.

The example above is one where a nonlinear mathematical operation is computed on a whole surface. Another example often used in environmental applications is computing the area above a threshold. For example, there might be some contaminant in the environment, and after sampling, we want to estimate the total area impacted above a regulatory threshold. Again, the kriging surface will be too smooth, and because the prediction standard errors are point-wise, it is impossible to obtain correct confidence intervals. Rather, we would create multiple realizations of surfaces, compute the area above the threshold for each surface, and then obtain the distribution of those areas to make our inference.

Conditional simulation was recognized from the outset of geostatistics as an important companion to kriging, and one of the earliest methods for conditional simulation was the turning-bands method (Matheron 1973). An introduction to turning-bands is provided by Mantoglou and Wilson (1982). Here, we will present a hierarchical formulation that makes conditional simulation more transparent.

In Bayesian terminology and methods, conditional simulation is the use of the posterior predictive distribution when using Markov chain Monte Carlo (MCMC) methods. As part of the MCMC chain, hierarchically the algorithm consists of 1) drawing covariance parameters from their posterior distributions, 2) conditional on the covariance parameters, drawing fixed effects from their posterior distributions, and 3) conditional on covariance parameters and fixed effects, drawing predictions from their conditional distributions. We describe a similar algorithm that, rather than being part of an MCMC chain, starts with independent samples of marginal estimates of covariance parameters, and proceeds conditionally from there.

We consider conditional simulation starting with a REML estimate, and it should be obvious how to do it analagously when starting with MLE.

- Preliminary item: If $\tilde{\boldsymbol{\theta}}$ is the REMLE, then compute $\mathcal{I}_R(\tilde{\boldsymbol{\theta}})$ as described in Section 8.7.2.

- Preliminary item: Set $k = 0$.

- Step 1: Set $k = k + 1$.

- Step 2: Sample $\boldsymbol{\theta}^*_k$ randomly from MVN($\tilde{\boldsymbol{\theta}}$,$\mathcal{I}_R(\tilde{\boldsymbol{\theta}})$). If any values of $\boldsymbol{\theta}^*_k$ are outside of their parameter space (e.g., a negative variance value), resample.

- Step 3: Sample $\boldsymbol{\beta}^*_k$ from MVN($\tilde{\boldsymbol{\beta}}_k, [\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}^*_k)^{-1}\mathbf{X}]^{-1}$), where $\tilde{\boldsymbol{\beta}}_k = [\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}^*_k)^{-1}\mathbf{X}]^{-1}\mathbf{X}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}^*_k)^{-1}\mathbf{y}$

- Step 4: Create $\mathbf{R}^*_k$, $\mathbf{R}^*_{\mathbf{uu},k}$, and $\mathbf{R}^*_{\mathbf{yu},k}$ by using $\boldsymbol{\theta}^*_k$ for the joint spatial autocorrelation matrix of $\mathbf{y}$ and $\mathbf{u}$. Let $\tilde{\mathbf{u}}^*_k = \mathbf{X}^T_{\mathbf{u}}\boldsymbol{\beta}^*_k + \mathbf{R}^{*T}_{\mathbf{uu},k}\mathbf{R}^{*-1}_k(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*_k)$ and $\tilde{\mathbf{C}}^*_k$ be the matrix Var($\tilde{\mathbf{u}} - \mathbf{u}$) in Section 9.1 where, in its construction, all covariance parameters are replaced with $\boldsymbol{\theta}^*_k$. Then draw the kth conditional simulation $\mathbf{u}^*_k$ from MVN($\tilde{\mathbf{u}}^*_k, \tilde{\mathbf{C}}^*_k$).

- Go to Step 1.

Formally, suppose that we have some non-linear function f computed on a conditional simulation $c_k = f(\mathbf{u}^*_k)$. Then, for K conditional simulations, we obtain the set $\{c_1, c_2, \ldots, c_K\}$ which can be used for inferences; i.e., mean, median, or mode values and valid confidence intervals can be computed on this set. This is the primary attraction of conditional simulation. If f is the identity function, then the mean of the set $\{c_1, c_2, \ldots, c_K\}$ will converge towards the kriging predictions, and the point-wise standard errors will converge towards the kriging standard errors.

References

Mantoglou, A. and Wilson, J. L. (1982), "The Turning Bands Method for simulation of random fields using line generation by a spectral method," *Water Resources Research*, 18, 1379–1394.

Matheron, G. (1973), "The intrinsic random functions and their applications," *Advances in Applied Probability*, 5, 439–468.