SEQUENTIAL SPATIAL SIMULATION USING LATIN HYPERCUBE SAMPLING

PHAEDON C. KYRIAKIDIS
Department of Geography, University of California Santa Barbara, Ellison Hall 5710, Santa Barbara, CA, 93106-4060, U.S.A.

Abstract. An efficient method is proposed for generating realizations from an arbitrary multivariate distribution using sequential simulation and Latin hypercube sampling. In a spatial context, this efficiency entails a reduction of sampling variability in statistics of spatially distributed model outputs when the inputs are realizations of random field models. The proposed method yields an unbiased reproduction of a target semivariogram, even for a small number of realizations, and consequently can be used for enhanced uncertainty and sensitivity analysis in complex spatially distributed models. In addition, the method is simple enough to be incorporated in virtually any geostatistical software for sequential simulation.

1 Introduction

Monte Carlo simulation is routinely used for uncertainty and sensitivity analysis of model outputs in a wide spectrum of scientific disciplines (Morgan and Henrion, 1990). Any realistic uncertainty analysis, however, calls for the availability of a representative distribution of such outputs, and can become extremely expensive in terms of both time and computer resources in the case of complex models and simple random (SR) sampling. This problem is far more pronounced for spatially distributed models, due to the large number of correlated (regionalized) variables comprising each input parameter map to such models, e.g., 3D rasters of hydraulic conductivity used for simulation of flow and transport in porous media.

An intelligent alternative to SR sampling is Latin hypercube (LH) sampling, a special case of stratified random sampling, which yields a more representative distribution of model outputs (in terms of smaller sampling variability of their statistics) for the same number of input simulated realizations. Analytical results demonstrating the efficiency of LH over SR sampling from univariate distributions are given in the (now classic) paper of McKay et al. (1979). A more recent comprehensive review of LH sampling for uncertainty and sensitivity analysis in complex systems can be found in Helton and Davis (2003).

LH sampling from a multivariate distribution, i.e., the task of inducing correlation in LH samples, is an important research theme in risk analysis and reliability

engineering (Haas, 1999), which becomes critical in a spatial context for ensuring unbiased outputs of complex spatially distributed models. This paper makes a novel contribution to the literature of spatial uncertainty analysis, by proposing a simple and efficient method for sequential LH sampling from random field models.

2 Latin hypercube sampling

Consider a set of \( K \) independent continuous RVs \( \{ Y_k, k = 1, \ldots, K \} \), with \( F_{Y_k}(y_k) = \text{Prob}\{Y_k \leq y_k\} \) denoting the cumulative distribution function (CDF) of the \( k \)-th RV \( Y_k \). Simple random (SR) sampling of \( N \) realizations from RV \( Y_k \) proceeds by first generating a \( (N \times 1) \) vector \( u_k = [u_k(n), n = 1, \ldots, N]' \) of uniform random numbers in \([0,1]\), which are treated as simulated probability values, and then transforming \( u_k \) into a \( (N \times 1) \) vector \( y_k = [y_k(n), n = 1, \ldots, N] \) of simulated realizations as: \( y_k = F^{-1}_{Y_k}(u_k) \), using the inverse CDF \( F^{-1}_{Y_k} \) of RV \( Y_k \).

Latin hypercube (LH) sampling of \( N \) realizations from the \( k \)-th RV \( Y_k \) calls for generating, independent of vector \( u_k \), a \( (N \times 1) \) vector \( p_k = [p_k(n), n = 1, \ldots, N]' \) of random permutations of \( N \) integers \( \{1, 2, \ldots, N\} \). A \( (N \times 1) \) vector \( z_k = [z_k(n), n = 1, \ldots, N]' \) of stratified realizations is then obtained as (McKay et al., 1979):

\[
    z_k = F^{-1}_{Y_k}\left(\frac{p_k - u_k}{N}\right) \tag{1}
\]

where the argument \((p_k - u_k)/N\) of the inverse CDF \( F^{-1}_{Y_k} \) ensures that the simulated probability values for the \( k \)-th RV \( Y_k \) are stratified, i.e., fall in \( N \) different probability strata. The monotonic transformation of the simulated probabilities incurred by the inverse CDF \( F^{-1}_{Y_k} \) does not ruin stratification, which entails that each entry of vector \( z_k \) (each simulated value) falls within a different stratum in the original variable space, no matter the distributional form of \( F_{Y_k}(y_k) \). The independence of vectors \( p_k \) and \( u_k \) ensures that there is a uniform probability \( 1/N \) for a simulated value within a particular stratum, i.e., there is no systematic placement of simulated values at the edges of strata. Variations of the above basic LH sampling procedure to further control sampling variability include variance reduction techniques, such as antithetic and control variates, as well as correlated sampling (Ang and Tang, 1984; Switzer, 2000).

A naive application of the above LH sampling procedure to correlated RVs fails to induce any correlation in the simulated values, simply because vectors \( p_k \) and \( u_k \) for the \( k \)-th RV \( Y_k \) are generated independent of other such vectors for other RVs. From these two sources that contribute to lack of correlation, the most important one is the vector \( p_k \) of random permutations because it dictates the strata within which the entries of \( u_k \) are distributed. To date, the most widely used method for generating LH samples from correlated RVs with a given rank correlation coefficient is the distribution-free method of Iman and Conover (1982). This method, however, is prohibitive for a large number \( K > 10,000 \) of RVs (typically the case in a spatial setting) because it calls for the Cholesky decomposition of an extremely large \((K \times K)\) variance-covariance matrix.