A Middle-Insertion Algorithm for Markov Chain Simulation of Soil Layering

Weidong Li
Department of Geography, Kent State University
Kent 44242, OH, USA
Tel: 1-330-998-0958
weidong6616@yahoo.com

Chuanrong Zhang
Department of Geography, Kent State University
Kent 44242, OH, USA
Tel: 1-330-998-0957
zhangchuanrong@gmail.com

ABSTRACT
Layering is a typical feature of alluvial soils in flooding plains. Simulation of soil layering is crucial to accurately quantifying soil subsurface physical and chemical processes. This paper introduces a middle-insertion (MI) Markov chain algorithm for vertical two-dimensional simulation of soil layering from borehole data based on the Markov chain random field theory. The MI algorithm requires the current pixel being estimated always be located at the middle of two nearest known pixels in a lattice row and thus performs a two-dimensional simulation row by row from top to bottom. The algorithm achieves effective symmetry in the lateral direction through this way, compared with the quasi-symmetry of the recognized alternate advancing (AA) algorithm, which was suggested to avoid pattern inclination in two-dimensional Markov chain simulation. A vertical soil transect is used as a simulation case study to demonstrate the feasibility of the new algorithm. Simulated results show that this fixed-path algorithm can work effectively and when boreholes are relatively sparse it generates relatively better layer continuity in realizations than the AA algorithm does. Therefore, the MI algorithm is suitable for simulation of soil layering.

Categories and Subject Descriptors
G.3 [Probability and statistics]: Markov processes, probabilistic algorithms (including Monte Carlo), stochastic processes, nonparametric statistics.

General Terms
Algorithms, Measurement, Verification.

Keywords
Geostatistics, Markov chain random field, transiogram, simulation algorithm, soil, natural resource evaluation.

1. INTRODUCTION
Characterization of soil subsurface structure heterogeneity (e.g., soil layering) is essential in soil science and hydrogeology because soil subsurface formations determine the preferential paths of water and solute transport and influence their movement in soils [1][2]. The information obtained about soil subsurface formations, such as alluvial soil textural layer structures in flooding plains, often comes from borehole records (usually within a 2-m depth). While borehole records provide sufficient knowledge about the vertical one-dimensional (1-D) variation of soil layers, limited information is available about their lateral extensions in vertical two and three dimensions due to the scarcity of existing borehole data or the high cost of obtaining densely distributed borehole records.

Similar to lithofacies, different alluvial soil layer (textural) types usually exhibit special features such as sharp boundaries, long-range connectivity, asymmetric vertical sequences, juxtaposition tendency and interclass dependences. These features pose a challenge to conventional geostatistical methods for categorical variables. For example, the indicator kriging simulation algorithm - sequential indicator simulation (SIS) is a widely used geostatistical technique for categorical (or indicator) variable modeling. While pointing out many good reasons for applying SIS to simulation of categorical variables, Deutsch [3] admitted that SIS often leads to uncontrolled and geologically unrealistic transitions between the simulated categories (i.e., classes) and the cross correlations between multiple categories is not explicitly controlled. In soil subsurface characterization, our major concerns are that (i) simulated realizations using SIS tend to have dispersed short-scale patterns and (ii) simulated layer class sequences may violate the actual neighboring relationships. Lin [4] pointed out that quantification of soil structure and its impacts on flow and transport in field soils remained unresolved. Bierkens and Weerts [5] ever used SIS to simulate soil subsurface textures in three dimensions and found that simulated results were not satisfactory in representing the connectivity of soil textural layers. To remove the geologically unrealistic short-scale effect generated in SIS-simulated realizations, Deutsch [3] suggested a post-cleaning method. He found that the facies (or layers) in simulated realizations became smoother and less erratic as the cleaning level increased. However, for dealing with complex categorical variables (i.e., multinomial classes), the cleaning level won’t be easy to control, and it is still a reasonable concern that conventional geostatistical methods may have the difficulty in keeping the correct layer sequence both in the vertical and horizontal directions because of the difficulty in incorporating interclass relationships in simulations.
Recently, Li [5] suggested a Markov chain random field (MCRF) theory for building sound Markov chain models for conditional simulation of categorical variables. Both the alternate advancing (AA) path algorithm and the middle insertion (MI) path algorithm were suggested as suitable fixed-paths for Markov chain simulation. An appropriate random-path algorithm was also proposed recently by Li and Zhang [6]. These algorithms are suitable because they are symmetric or quasi-symmetric in the lateral direction (i.e., the simulation proceeding direction for fixed paths) or in all directions (i.e., the simulation proceeding direction for the random path) and thus induce omni-directional correlation of simulated values. Otherwise if a simulation path is not symmetric or quasi-symmetric in the simulation proceeding direction, correlation of simulated values won’t be induced omni-directionally and thus pattern abnormality (e.g., diagonal inclination) will arise [7][8]. Note that here different Markov chain algorithms may mean different specific models because a specific algorithm may need a specific model to implement. Designing suitable algorithms is crucial in generating appropriate spatial patterns in 2-D simulation of unilateral processes (including Markov chains) models, and improper algorithms may generate serious artifacts and even disqualify a model for practical use [8].

In this paper, we introduce the MI algorithm for Markov chain simulation of soil layering. We hypothesize that the MI path has some advantages over the AA path in capturing the complex spatial patterns of soil layers in some situations because it contains symmetry and shorter interaction distances with borehole data compared with the AA path. It is expected that such a Markov chain algorithm will be useful in subsurface characterization.

2. METHODOLOGY AND DATA

2.1. The Markov Chain Model

A Markov chain model based on the MCRF theory contains only a single Markov chain [5]. Therefore, MCRF-based models differ from conventional multiple-chain-based Markov chain models. For shallow subsurface characterization, the observed data available for estimating an unknown location come from three directions – left borehole log, right borehole log and the upper boundary. With a fixed-path algorithm, the general Markov chain model for subsurface characterization needs to consider only three nearest known neighbors in three cardinal directions – left, up and right (Figure 1). Assuming the Markov chain moves from location \( u_1 \) to the current unknown location \( u \), and its three nearest known neighbors are \( u_1, u_2 \), and \( u_3 \) in the left, up and right directions, respectively, the conditional probability distribution of the Markov chain (i.e., the random variable \( Z \)) at the unknown location \( u \) can be generally given as

\[
Pr(z(u) = k | z(u_1) = l, z(u_2) = m, z(u_3) = q) = \frac{p_{l,m}^h(h_1) \cdot p_{m,q}^h(h_2) \cdot p_{q,h}^h(h_3)}{\sum_{j=1}^n [p_{l,j}^h(h_1) \cdot p_{m,j}^h(h_2) \cdot p_{q,j}^h(h_3)]}
\]

(1)

where, \( n \) is the number of soil layer types (i.e., classes); \( h_1, h_2 \), and \( h_3 \) represent the distances from the unknown location \( u \) to its nearest known neighbors \( u_1, u_2 \), and \( u_3 \); \( l, m, \) and \( q \) represent the states of the Markov chain at the four locations \( u, u_1, u_2 \), and \( u_3 \); and any one \( p_{l,m}^h(h) \) represents a transition probability from class \( k \) to class \( l \) in the \( i \)th direction over the distance lag \( h \). With increasing lag \( h \), \( p_{l,m}^h(h) \) forms a transiogram – a transition probability diagram. In the Markov chain model, the Markov chain decides its state by interacting with its nearest known neighbors in the three cardinal directions through transition probabilities with different lags, which are provided by transiogram models.

Figure 1. Illustration of the general Markov chain random field model for subsurface characterization. Its specific forms can work with the middle insertion and alternate advancing algorithms. Gray cells represent the three known boundaries (left and right borehole logs and the upper boundary).

Consider a soil vertical section bounded by two boreholes and the ground surface, along which soil type information are all known. The simulation of internal unknown locations in such a soil vertical section must be conditioned on the three boundaries. To use the MI path and conduct the simulation row by row from the top to the bottom, we have \( h_1 = h_3 \) and \( h_2 = 1 \). Thus for such an algorithm Equation (1) specifically changes to

\[
Pr(z(u) = k | z(u_1) = l, z(u_2) = m, z(u_3) = q) = \frac{p_{l,m}^h(h_1) \cdot p_{m,q}^h(h_2)}{\sum_{j=1}^n [p_{l,j}^h(h_1) \cdot p_{m,j}^h(h_2) \cdot p_{q,j}^h(h_3)]}
\]

(2)

where \( h \) is the distance from the current pixel to either of its two nearest known neighbors in a lattice row. Note that in case the number of pixels between two nearest known pixels in a row is even, we make \( h_1 = h_2 = 1 \) (i.e., \( h \) is not absolutely equal for the left and right nearest known locations). Such a small lag difference does not influence the lateral symmetry in a high-resolution simulation that involves a large number of pixels in each row between two neighboring boreholes.

2.2. Simulation Algorithm

The simulation process of filling a lattice row using the MI path algorithm is demonstrated in Figure 2. In a row, we fill unknown pixels recursively by choosing the middle pixel between two nearest known pixels from left to right. Thus, for a high-resolution simulation that involves many pixels in each row, the symmetry in the lateral direction can be well achieved. Such symmetry achieved by the MI path represents an advantage over the AA path, which achieved only quasi-symmetry in the lateral direction. In addition, because the largest lag \( h \) used in the MI path algorithm is equal to a half of the longest borehole interval, in simulation this path uses transition probability values with shorter lags compared with that used in the AA path algorithm. For transiogram models, the larger lag means the less reliability.
Therefore, this is the second advantage of the MI path.

To conduct simulation, we need to first estimate transiogram models from experimental transiograms, which can be directly estimated from sample borehole data. For vertical 2-D simulation, experimental transiograms in the vertical (from bottom to top) direction and in the lateral (from left to right) direction (see the directions in Figure 1) are needed. After transiogram models are obtained, conditional simulation can be performed on a rectangular lattice, where borehole logs have been inserted into some columns at the corresponding locations. The MI simulation algorithm contains the following steps:

- **Step 1**: Discretize a soil vertical section into a rectangular lattice with the required pixel size.
- **Step 2**: Insert borehole data and the ground surface data into the corresponding cells in the lattice.
- **Step 3**: Search the second row (the first row is known as the upper boundary) from left to right for an unknown pixel that is located at the middle between two known nearest pixels. Note that if the number of pixels between two nearest known pixels in a row is even, set \( h = h + 1 \).
- **Step 4**: Estimate the value of the unknown pixel using Equation (2) and Monte Carlo sampling by conditioning on its left, right and upper known neighbors. Transition probabilities with required lags are fetched from corresponding transiogram models.
- **Step 5**: Continue Steps 3 and 4 until all unknown pixels are visited in the row.
- **Step 6**: Move to the next row to fill all unknown pixels in the same way, until unknown pixels in all rows of pixels are visited. Thus a realization is generated.
- **Step 7**: Repeat Steps 3 to 6 to generate the next realization.

**Figure 2. Illustration of the middle insertion path algorithm:** Unknown pixels in a lattice row are filled recursively in the simulation process. Gray cells represent known pixels (including simulated pixels); white cells represent unknown pixels; the arrow points to the current location being estimated and the current location is always located at the middle of two nearest known pixels.

**2.3. Study Site and Borehole Data**

A long alluvial soil transect in the North China Plain was surveyed. The soil transect has a length of 5250-m and a depth of 2-m. Soil layers were classified into four classes – sand, sandy loam, loam, and clay (represented as 1, 2, 3, and 4, respectively). Because the soils in the plain are alluvial soils, generated by a series of floods in the geologic history, the soil layer boundaries between different textures are normally clear. We chose 22 borehole records as a dense dataset. The averaged borehole interval is about 254-m. From the dense dataset, we further chose 11 borehole records as a sparse dataset, with an averaged interval of about 525-m. The soil layer classes at the ground surface were assumed known (they can be easily obtained in field survey). For the simulation purpose, the soil transect was simply discretized into a lattice of 310 columns and 43 rows, with unequal pixel sizes in the vertical and lateral directions (i.e., about 17-m × 4.6-cm) so that the long transect could be conveniently displayed. Note that the vertical simulation resolution (i.e., pixel size) is not necessary to be the same as the vertical observation interval because transiogram models are continuous curves and have no the limitation to pixel sizes. This is also one of the advantages of transiograms over conventionally used transition probability matrices.

**2.4. Estimation of Transiogram Models**

Experimental transiograms were directly estimated from borehole logs. After experimental transiograms are estimated, the next step is to perform their joint modeling subset by subset using suitable mathematical models [9]. Here a subset refers to all the experimental transiograms headed by the same class. Joint modeling is necessary to meet the constraint conditions of transiogram models for Markov chain simulation, such as non-negative and summing-to-one conditions. Because there are totally four layer classes in this case study, 16 experimental transiograms in the lateral direction and 16 in the vertical direction need to be estimated from borehole data. Note that MCRF models for subsurface characterization do not need transiogram models in opposite lateral directions because they are well generalized.

**Figure 3. Four experimental auto/cross-transiograms in the west-to-east direction and fitted models.** The scale along h axis is represented by the numbers of pixel lengths. Label “p(2,3)” refers to the cross-transiogram from class 2 (sandy loam) to class 3 (loam).

Figure 3 shows four experimental auto/cross-transiograms in the west-to-east direction estimated from borehole records. Note that simulations use only a very short section of vertical transiogram models close to their origins [i.e., (0, 0) for cross-transiograms and (0, 1) for auto-transiograms], which is a less than 4.6-cm lag here. In the lateral direction, the same 22 borehole records are used to estimate transiogram models. Because the number of data-pairs decreases and the estimated transition probabilities deviate more from the truth with increasing lag, estimated transition probability values at high lags were ignored and only the first 10

\[
p(1,1) = \begin{bmatrix}
0.3 & 0.3 & 0.4 & 0.6 & 0.8 \\
0.3 & 0.4 & 0.6 & 0.8 & 1.0 \\
0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
\end{bmatrix}
\]

\[
p(2,3) = \begin{bmatrix}
0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\
0.2 & 0.3 & 0.4 & 0.5 & 0.6 \\
0.3 & 0.4 & 0.5 & 0.6 & 0.7 \\
0.4 & 0.5 & 0.6 & 0.7 & 0.8 \\
0.5 & 0.6 & 0.7 & 0.8 & 0.9 \\
\end{bmatrix}
\]

\[
p(3,2) = \begin{bmatrix}
0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
0.4 & 0.6 & 0.8 & 1.0 & 0.0 \\
0.6 & 0.8 & 1.0 & 0.0 & 0.0 \\
0.8 & 1.0 & 0.0 & 0.0 & 0.0 \\
1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\
\end{bmatrix}
\]

\[
p(4,1) = \begin{bmatrix}
0.1 & 0.2 & 0.3 & 0.4 \\
0.2 & 0.3 & 0.4 & 0.5 \\
0.3 & 0.4 & 0.5 & 0.6 \\
0.4 & 0.5 & 0.6 & 0.7 \\
0.5 & 0.6 & 0.7 & 0.8 \\
\end{bmatrix}
\]
transition probability values in each experimental transiogram are used for model fitting. In fact, simulations used only the low-lag section (i.e., the section with lags shorter than a half of the largest borehole interval) of lateral transiogram models, which is 800-m in the dataset. The exponential model and the spherical model [9] were used to fit experimental transiograms. Although these two simple models have difficulty in capturing the complex shapes of experimental transiograms at high lags, they are usually acceptable for fitting the low-lag section.

3. RESULTS AND ANALYSIS
We simulated soil layers using both the AA path and the MI path algorithms. Simulations were conditioned on the two borehole datasets and the top boundary of the vertical transect. Figure 4 shows simulated results conditioned on 11 boreholes. The prediction maps were based on the maximum occurrence probabilities estimated from 100 realizations using the MI and AA algorithms, respectively. The borehole interval is 525-m (or 31 data columns), which is a normal interval used in soil survey. Both simulated realizations and the prediction map based on maximum probabilities displayed the long-range layer features with clear boundaries. The two different algorithms both captured similar patterns but the MI path algorithm generated relatively better layer connectivity in the lateral direction (see the thin layer in the low-right place in realizations in Figure 4). This should be related to the better lateral symmetry achieved by the MI algorithm and the shorter transiogram lags used by the algorithm.

Figure 4. Simulated results conditioned on 11 boreholes (about 525-m interval) and the top boundary. Label “MI R1” refers to the first realization using the MI algorithm.

But when we increased the number of boreholes to 22, that is, decreased the borehole interval to about 254-m (or 15 data column), simulated realizations by the two algorithms showed no difference. This means that when conditioning data are dense, data play the major role in determining simulated results and the specific algorithms used will not have much influence as long as the general model is the same. It should be noted that the directional asymmetry of layer occurrence sequence in the vertical direction was well kept and the layer boundaries were clear in all simulated realizations no matter how many boreholes were conditioned. These features should be attributed to the incorporation of interclass relationships described by cross-transiogram models and the non-linear characteristic of the MCRF-based Markov chain models.

4. CONCLUSION
The MCRF-based MI algorithm for Markov chain simulation of subsurface heterogeneity from borehole data was introduced. Simulations of a long soil vertical transect conditioned on two borehole datasets of different densities show that the algorithm can effectively capture subsurface layer structures of soils such as long-range features, asymmetric layer sequence and clear layer boundaries. Compared to the AA algorithm, the MI algorithm can generate realizations with relatively better layer connectivity when boreholes are relatively sparse. But when boreholes are dense, they generate similar results. This means the MI algorithm provides a new method for Markov chain simulation of subsurface heterogeneity and it may represent a better solution than the AA algorithm.

Figure 5. Simulated results conditioned on 22 boreholes (about 254-m interval) and the top boundary. Label “AA R2” refers to the second realization using the AA algorithm.

7. REFERENCES