2D Markov-Chain Simulation and Prediction of Shallow Subsurface Alluvial Soil Textural Layers

Weidong Li[* a] , Chuanrong Zhang [b], et al.

Department of Geography, University of Wisconsin, Madison, WI 53706
Department of Geography, University of Wisconsin, Milwaukee, WI 53201

Abstract. Characterization of shallow subsurface sedimentary architecture, such as alluvial soil layers, is essential in hydrogeology. Elfeki and Dekking (2001) presented a highly efficient coupled Markov chain model for subsurface characterization with conditioning on well logs. The model is capable of reproducing some major subsurface features at their approximate locations. However, the model shows some obvious limitations, such as layer inclination and under-estimation of minor states, when applied to simulating the structures of alluvial soil textural layers with sparse borehole data. This study presents an idea to mitigate these deficiencies by conditioning two-dimensional simulations of vertical transects on simulated lines generated by one-dimensional Markov chain methods. This method is proven to be effective with some tradeoffs by rigorous testing using different borehole schemes and comparison with simulated realizations from the coupled Markov chain model. Since datasets of densely distributed boreholes are often not available in most real-world applications, the method may enlarge the application scope of the coupled Markov chain methodology in shallow subsurface characterization with sparse borehole data.

Keywords: Markov chain; Alluvial soil layers; Borehole; Simulated lines; Conditional simulation; Subsurface heterogeneity.

1. Introduction

Characterization of sediment layer heterogeneity in shallow subsurface is essential in hydrogeology, pedology, and sedimentology. The information obtained about shallow subsurface formations such as alluvial soil layers in flooding plains, often comes from borehole records (Weissmann et al., 1999; Li et al., 1997). Boreholes provide sufficient knowledge about the vertical variation of the depositional sequences (e.g., sediment textural layers). Borehole data are available in many soil surveys (usually within a 2m depth). But little information is available about the lateral extension of shallow subsurface features due to the difficulty of obtaining this information. Soil scientists mainly rely on the use of soft data (expert knowledge, common sedimentary principles, field profile observation, etc.) as a source of information about the lateral variability of alluvial soil layering. However, the density of boreholes (because of its high cost) is often too low for scientists or surveyors to correctly interpolate the lateral extensions of soil layers. The lateral extension of soil textural layers has strong influence on water and solute transport in the vadose zone (Koltermann and Gorelick, 1996; Feyen et al., 1998; Li et al., 2001). It is necessary to develop effective methods capable of characterizing the spatial heterogeneity of shallow subsurface formations in two- to three-dimensions with limited borehole data.

1.1 A Simple Review of Related Methods

Imitation of spatial heterogeneity of shallow subsurface categorical variables such as sediment textural layers in multi-dimensions is difficult using conventional methods when available survey data is very limited (Koltermann and Gorelick, 1996). For example, indicator-based geostatistical methods (Deutsch and Journel, 1997), which are well-developed and popularly used, do not consider sedimentary probability rules and have difficulty to deal with the abrupt boundaries (McBratney et al., 2000) between layers, therefore, it is difficult for them to honor the transitions between different facies (Murray, 1994). Parameter estimation (e.g., variograms in the lateral direction) is also difficult when borehole data are few (Bierkens and Weerts, 1994; Weissmann and Fogg, 1999). Computation time and parameter preparation workloads are high when many nominal classes are involved and cross-variograms and anisotropies are considered (Ehlschlaeger, 2000; Zhang and Goodchild, 2002). Some recently developed methods can be seen in Carle and Fogg (1996), Weissmann and Fogg (1999), Ehlschlaeger (2000), Bogaert (2002), and Chen and Rubin (2003), which all attempt to avoid or deal with the constraints of indicator geostatistics with different extents of success.

The Markov chain theory represents another approach for heterogeneity characterization. The unique merits of Markov chains are that (1) each single Markov chain can describe a spatially interdependent sequence rule of different states (i.e., classes) in one direction, (2) both auto-correlation and cross-correlation of different states are included in a Markov transition probability matrix, (3) many classes can be dealt with simultaneously, and (4) Markov transition probabilities are relatively more intuitive than indicator variograms; therefore soft information, even expert knowledge, such as facies length, proportions, and juxtaposition relationships, is relatively easier to be incorporated into estimation of Markov transition probability matrices (Rosen and Gustafson, 1996; Weissmann and Fogg, 1999). Although one-dimensional Markov-chain methods have been used widely in different fields such as geology, soil science and ecology (Harbaugh and Bonham-Carter, 1980; Burgess and Webster, 1984a, b; Li et al., 1999; Balzter, 2000), using Markov chains for multi-dimensional simulation, especially for conditional simulation and prediction (i.e., interpolation), is relatively difficult and usually with obvious constraints. A typical earlier work, for example, can be seen in Lin and Harbaugh (1884), who first realized the unconditional multi-dimensional Markov chain simulation of geological formations based on the Swizter's (1965) theorem. In recent years, the Markov random field (MRF) approach (Besag, 1974, 1986; Descombes et al., 1999; Wu et al., 2004) and the hierarchical Markov chain transition probability matrix approach (Johnson et al., 1999; Patil and Taillie, 2001) have found their uses in image processing and landscape ecology. However, these two approaches normally need exhaustive data (i.e., the original image) for simulation; thus, they actually don't provide a tool for spatial prediction from sparse measured data. Multi-dimensional Markov chain methods for conditional simulation and prediction from measured data emerged only recently in geosciences as a tool for characterizing lithofacies (Elfeki and Dekking, 2001; Norberg et al., 2002). Conditional simulation is obviously more interesting and useful because spatial patterns can be imitated to approximate locations with the conditioning of measured data and their spatial uncertainty may be analyzed using occurrence probabilities calculated from multiple realizations (Zhang and Goodchild, 2002).

The main difficulties for using explicit multi-dimensional Markov chains for conditional simulations include: (1) conditioning on multiple boundaries (or measured data) and (2) choosing a suitable simulation ordering in a past-present-future sequence (Koltermann and Gorelick, 1996). Recently, Norberg et al. (2002) further extended the computationally simple Bayesian Markov

random field methodology of Rosen and Gustafson (1996) to simulate lithofacies from sparse data. Although Markov random field methods avoid the aforementioned difficulties, their implementation employs a heavily iterative scheme to obtain the final spatial configuration. The iterative algorithm is slow to converge or even may fail and thus results in extremely high demand in computation. For example, for an area of about 100×100 pixels (i.e., grid cells), the method of Norberg et al. (2002) needs 1.5 to 2.5 days of run time on a SUN workstation for producing a realization. Other deficiencies include the severe under-estimation of minor states and that the model cannot produce realistic maps by unconditional simulation, as mentioned by the authors (Norberg et al., 2002). Additionally, the standard implementation of Markov random fields also cannot reproduce anisotropic structures (Tjelmeland and Besag, 1998). These constraints make it currently not well-suited for simulating categorical variables from sparse data over large areas and for uncertainty analyses, which normally needs many realizations to be generated.

1.2 Review of the Coupled Markov Chain Model

The coupled Markov chain (CMC) model of Elfeki and Dekking (2001) for subsurface lithofacies characterization uses an explicit non-iterative algorithm (i.e., one pass one realization), which makes it highly efficient. Theoretically, the model also represents another way to realize multi-dimensional simulation using Markov chains. More significantly, it presents a solution for approximately conditioning on borehole data. The study cases of Elfeki and Dekking (2001) demonstrated that the CMC model could capture the major features (with long extensions) of subsurface geological formations at their approximate locations when a number of well data were conditioned. This may be an obvious advantage of the model for subsurface imitating (Elfeki and Dekking, 2001, p.586). While the model has some significant merits, obvious deficiencies can be found from simulations using the model, which include (a) layer inclination tendency along the simulation direction, (b) layer discontinuity along boreholes, and (c) underestimation of minor states. The model can generate very imitative realizations when boreholes are densely distributed (otherwise realizations are quite unrealistic) or layers are naturally tilted along the simulation direction and different states account for a similar proportion (i.e., no minor or major states).

To explain the deficiencies (a) and (b), please note the model employs a fixed asymmetric path (i.e., simulation ordering). Thus, it has to first decide a simulation sequence from the top-left corner to the bottom-right corner or from the top-right corner to the bottom-left corner row by row for conducting a simulation. This asymmetry is further emphasized by the asymmetric conditioning neighborhood (i.e., the immediate top cell and one immediate side cell). A random path that is used in some other random field models to avoid simulation artifacts (Kyriakidis and Dungan, 2001) is not feasible for this model because the generation of the current cell depends on its immediate preceding neighboring cell and its upper cell. Our simulation of alluvial soil transects using this model shows that this ordering problem results in some simulation artifacts – layer inclination along the simulation ordering and layer discontinuity (will be shown in our simulation cases in this paper, also can be seen in Li (1999). Layer discontinuity also can be seen in the simulation cases of Elfeki and Dekking (2001)). Conditioning on future states (i.e, internal boreholes) will mitigate these artifacts. However, the influence of future states on the coupled Markov chain is usually short-distanced, particularly if layers of the states are not very thin and long (i.e., strongly auto-correlated) along the lateral direction; therefore, unless the boreholes is densely distributed or layers are naturally tilted along the simulation direction, these artifacts

cannot be effectively eliminated. Very thin layers (by carefully choosing cell sizes) may not show clear inclination tendency, but layer discontinuity along borehole lines will be stronger. These artifacts mitigate with the increase of the density of boreholes - the conditioning data.

Underestimation of minor states (consequently overestimation of major states) occurs when different states account for different proportions. Some minor states may be missing when conditioning data are too sparse. This point is also mentioned by the authors of the CMC model in another way as "the geological features with short extensions are not very well reproduced" (Elfeki and Dekking, 2001, p.586) and "it is also important to point out that the lithology coded 5 (black) does not appear in any of the wells and so it is reproduced neither in the single realization or in the ensemble average" (Elfeki and Dekking, 2001, p.588) although it is represented in parameters. Underestimation of minor states also occurs in the Markov random field method of Norberg et al. (2002). Norberg et al. (2002) thought that their method might have the tendency of over-estimating spatial dependencies of classes because of the possible existence of phase transitions (Guyon, 1995). This explanation may also apply to the CMC model. The second reason is related with the independency assumption of the two one-dimensional Markov chains, because this assumption for the CMC model to be implemented is also an assumption that cannot be justified. Our observation from our simulations shows that this problem also gradually mitigates with the increase of the density of boreholes - the conditioning data in the CMC modeling. This observation is similar as those mentioned by the CMC model developers, which said that "it is clear that by increasing the number of wells, the simulation results improve and become closer to the original image ('real' formation)" (Elfeki and Dekking, 2001, p.579). Therefore, when boreholes are sparse, the CMC model is not sufficient to generate realistic realizations. Please note, the sparsity is relative to the lengths of layers in the lateral direction. The longer are the lengths of layers, the fewer boreholes are needed to generate realistic realizations. Please also note, whether or not the aforementioned problems are model deficiencies also depend on whether or not these problems are users' concerns.

In general, the CMC model has its obvious advantages in both theory and applications. However, to make such an efficient approach to be applicable for other simulation purposes, such as alluvial soil layers for hydrological modeling, with sparse boreholes, it is necessary to further mitigate the aforementioned deficiencies.

1.3 Objectives

Data sparsity is the normal case in real world applications. It is also desirable to have the subsurface features more reasonably represented with approximate areal proportions and shapes in conditional multidimensional Markov chain simulation if measured data are not sufficient to reproduce them at their accurate locations. In the study we develop an idea - "*conditioning on simulated lines*" generated by one-dimensional Markov chains to simulate alluvial soil textural layers. This idea is based on two points of observations: (1) increasing conditioning data will obviously mitigate the deficiencies of the CMC model and (2) one-dimensional Markov chains do not underestimate minor states. The purpose of using simulated lines is to increase the density of conditioning data for two-dimensional simulation, thus to mitigate the deficiencies of the CMC model. Such an idea provides a simple and intuitive solution for mitigating the deficiencies of the CMC model with tradeoffs. It may enlarge the application scope of the CMC theory for different study purposes with sparse borehole data. The theoretical foundation of this method is still the CMC theory of Elfeki and Dekking (2001). This paper will concentrate on graphically

testing the application of this idea and related conditions and constraints by simulating an alluvial soil transect with different borehole schemes.

2. Methods

To realize this idea, we need one-dimensional Markov chain methods to first generate a number of simulated lines and then use the CMC model to fill the left unknown cells. Therefore, such an idea is actually a mixture of one-dimensional Markov chains and the CMC that needs to be further extended.

2.1 One-dimensional Markov Chains for Generating Simulated Lines

Assuming $(X_i)_{0\le i\le N}$ is a discrete, stationary, and first-order one-dimensional Markov chain defined on the state space [S_1, S_2, ..., S_n], we can express this one-dimensional Markov chain as

$$p(X_i = S_k \mid X_{i-1} = S_l, ..., X_0 = S_r) = p(X_i = S_k \mid X_{i-1} = S_l) = p_{lk} \tag{1}$$

where p_{lk} is the one-step transition probability from state S_l to state S_k, which means that the occurrence of state S_k in cell i only depends on the state S_l in the previous cell i-1 but does not depend on states in cells further removed. The one-step transition probability matrix (TPM) contains all one-step transition probabilities between different states, which together describe a Markov chain. A one-step TPM is expressed by

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \dots & \dots & \dots & \dots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix} \tag{2}$$

where n is the number of states. Equation (1) will be used to simulate vertical lines, i.e., false boreholes.

But if a future state at X_N is known as S_q, its influence can be accounted for by "*conditioning*" the Markov chain on that state through multi-step transition probabilities. A one-dimensional Markov chain conditioned on a future state (Elfeki and Dekking, 2001) can be given by

$$p(X_i = S_k \mid X_{i-1} = S_l, X_N = S_q) = p_{lk|q} = \frac{p_{lk} \cdot p_{kq}^{(N-i)}}{p_{lq}^{(N-i+1)}} \tag{3}$$

where $p_{kq}^{(N-i)}$ is a (N-i)-step transition probability, $p_{lq}^{(N-i+1)}$ is a (N-i+1)-step transition probability, and $p_{lk|q}$ is the probability of cell i to be in state S_k, given that the previous cell i-1 is in state S_l and the future cell N is in state S_q. Please note that an m-step TPM can be calculated by imposing a power of m to the one-step TPM, i.e.,

$$P^{(m)} = P^m = \underbrace{P \cdot P \cdot \dots \cdot P}_{m} \tag{4}$$

The influence of the future state on the Markov chain is normally short-distanced because of the stationary property of a Markov chain. That is, a known cell only influences the determination of a cell that is close to it by transition probabilities. When the m in Equation (4) is large enough, the m-

step TPM will reach a status that it has equal transition probability values in each column. These transition probabilities are called stationary probabilities, which can be expressed by

$$W = (w_1, w_2, \ldots, w_n) = \lim_{m \to \infty} P^{(m)} \tag{5}$$

Thus, in Equation (3), when cell N is far from cell i the terms $p_{lq}^{(N-i+1)}$ and $p_{kq}^{(N-i)}$ have little influence on $p_{lk|q}$ because they both will be almost equal to the same stationary probability w_q. However, when the simulation gets closer to cell N, its state will start to play a role and the simulation result will be affected by the state at that cell.

Equation (3) will be used to generate lateral lines by conditioning on boreholes and simulated vertical lines if there are. In addition to one one-step TPM, it needs to know the initial point and the end point (as a future state) for conducting such a one-dimensional simulation.

2.2 Coupled Markov chain with Conditioning on Borehole Data

Here we give a simple introduction of the CMC model. For details see Elfeki and Dekking (2001). Considering a vertical one-dimensional Markov chain $(X_i)_{0 \le i \le N_x}$ and a horizontal one-dimensional Markov chain $(Y_j)_{0 \le j \le N_y}$ are coupled to form a two-dimensional Markov chain ($Z_{i,j}$) on a lattice (Fig. 1), and only considering state transition from state S_l at $Z_{i-1,j}$ and state S_m at $Z_{i,j-1}$ to the same state S_k at $Z_{i,j}$, if the two one-dimensional chains are assumed to be independent of each other, the joint transition probability can be given as

$$\begin{aligned} p_{lm,k} &= p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m) \\ &= C \cdot p_{lk}^h \cdot p_{mk}^v = \frac{p_{lk}^h \cdot p_{mk}^v}{\sum_f p_{lf}^h \cdot p_{mf}^v} \qquad k = 1, \ldots, n \end{aligned} \tag{6}$$

where C is a normalizing constant, which arises because we only consider the transitions from S_l and S_m to the same state S_k. The superscripts h and v in the above equation represents the directions of Markov chains, i.e., horizontal and vertical, respectively. The subscripts l, k, m and f all represent states in the state space. The above Equation (6) actually represents the unconditional (on future states) coupled Markov chain model developed by Elfeki (1996). Because no influence of future states is considered in this model, unless all states have similar areal proportions minor states may be strongly underestimated or even missing in simulated results and layers are also strongly inclined if they are not very thin (Li, 1999) or strongly discontinuous if they are thin (Elfeki and Dekking, 2001, p. 586) when boreholes are sparse.

Suppose the horizontal chain is conditioned to the known future state S_q at $Z_{N_x,j}$ on the right boundary of a window (see Fig. 1), by applying Equation (3) to Equation (6), the expression of the conditional joint transition probability in the coupled chain can be given by

$$\begin{aligned} p_{lm,k|q} &= p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m, Z_{N_x,j} = S_q) = C' \cdot p_{lk|q}^h \cdot p_{mk}^v \\ &= \frac{p_{lk|q}^h \cdot p_{mk}^v}{\sum_f p_{lf|q}^h \cdot p_{mf}^v} = \frac{p_{lk}^h \cdot p_{kq}^{h(N_x-i)} \cdot p_{mk}^v}{\sum_f p_{lf}^h \cdot p_{fq}^{h(N_x-i)} \cdot p_{mf}^v} \end{aligned} \tag{7}$$

where the symbols have same meanings as they have in Equation (3) and (6). The above equation (7) represents the CMC model developed by Elfeki and Dekking (2001) for simulations of geological cross-sections. The added feature of "conditioning on borehole data" is the core function for

conducting conditional simulations and mitigating the simulation artefacts. This conditioning was done in an approximate but computationally cheap way as the one in one-dimension shown in Equation (3). Elfeki and Dekking (2001) also mentioned another possible conditioning way, i.e., conditioning a current state to a future (known) state and the past states in the previous row together. However, this more exact and also more complex conditioning way was not integrated into the CMC model. The reason may be that it is difficult to implement and also too time-consuming in computation.

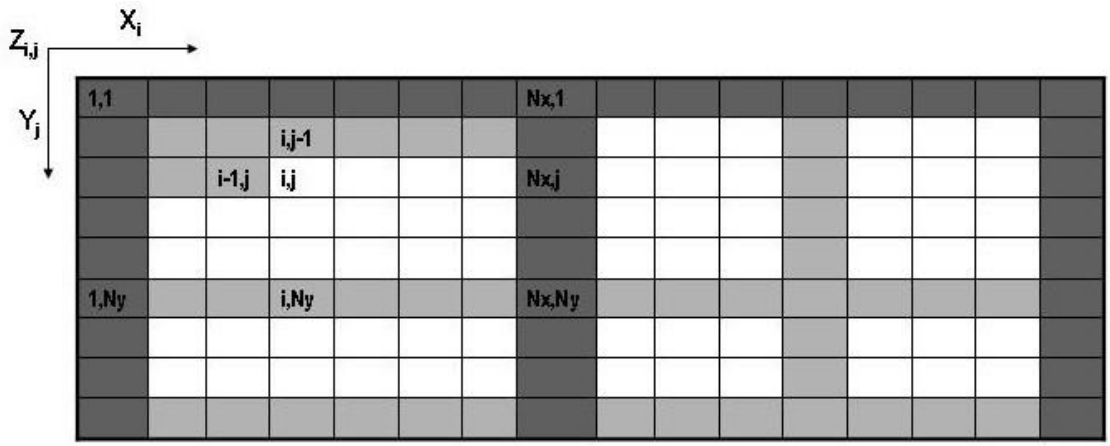

Figure 1. A coupled Markov chain with conditioning to boreholes and simulated lines. Dark gray color represents known data (top boundary and boreholes). Light gray represents simulated lines using the one-dimensional Markov chain methods and already filled cells by the coupled Markov chain.

2.3 Conditioning the Coupled Markov Chain to Lateral Line Data

Normally, the lateral information for subsurface is difficult to acquire; therefore assuming measured data in the lateral direction for conditioning is not suitable. With the idea of 'conditioning on simulated lines', it becomes realistic to extend the above CMC model to condition on future states on lateral lines. Assuming that S_o is the known state at cell Z_{i,N_y} on the bottom boundary of a simulation window and applying Equation (3) to Equation (7) to replace the transition probability in the vertical direction, we have the conditional joint transition probability of the coupled chain as

$$p_{lm,k|qo} = p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m, Z_{N_x,j} = S_q, Z_{i,N_y} = S_o)$$

$$= C'' \cdot p^h_{lk|q} \cdot p^v_{mk|o} = \frac{p^h_{lk|q} \cdot p^v_{mk|o}}{\sum_f p^h_{lf|q} \cdot p^v_{mf|o}} = \frac{p^h_{lk} \cdot p^{h(N_x-i)}_{kq} \cdot p^v_{mk} \cdot p^{v(N_y-j)}_{ko}}{\sum_f (p^h_{lf} \cdot p^{h(N_x-i)}_{fq} \cdot p^v_{mf} \cdot p^{v(N_y-j)}_{fo})} \quad (8)$$

This extension is very straightforward and is still based on the CMC theory, but it will make the CMC model not an exact example of unilateral Markov field any more (Galbraith and Walley, 1976; Pickard, 1980).

The above equation (8) will be used in our proposed method. The condition for the model is that we must have lateral lines whose states are already known before performing two-dimensional simulation using the above equation to fill the windows.

2.4 The Proposed Method

The proposed method is composed of the one-dimensional Markov chains in Equation (1) and (3) and the extended CMC in Equation (8). For the convenience to present it, we denote this method as mCMC (modified coupled Markov chain model). Inserting simulated lines, particularly simulated vertical lines, is useful only when the measured boreholes are not sufficient for the CMC model to generate satisfied realizations, which may depend on different application cases and purposes. Therefore, the proposed method is only used for dealing with insufficient borehole data. To produce a realization, we will first consider inserting simulated vertical lines between boreholes if it is necessary because of the excessive sparseness of measured boreholes, and then insert simulated lateral lines by conditioning them on the boreholes and simulated vertical lines. These simulated lines together with the hard boreholes partition the simulation domain into small "*windows*". Finally the extended CMC model in Equation (8) performs to fill in windows with conditioning on its known boundaries. If boreholes are not excessively sparse (e.g., the borehole interval is not obviously less than the mean-length of layers), don't insert simulated vertical lines. The basic points for inserting simulated lines will be further discussed later.

2.5 Parameter Inference

A Markov chain is described by its state space, one-step TPM and initial state. The state space can be determined according to the actual need. For example, shallow subsurface alluvial soil textural layers in a soil transect can be classified into several classes (i.e., types) based on their hydrologic properties for the purpose of hydrologic modeling. The cell size should be the same for both parameter estimation and simulation. The initial state becomes known if one boundary point is known.

To estimate parameters from existing maps or known cross-sections, we need to superimpose a lattice on them. The cell size should not be larger than the smallest parcels we want to show in simulated realizations. The transition frequencies between the states in the horizontal or vertical direction can be calculated by counting the times of a given state (e.g., S_i) followed by itself or the other states (e.g., S_j) in the direction on the lattice, and then the one-step transition probabilities (for the one-dimensional Markov chain in that direction) can be obtained by dividing the transition frequencies with the total number of transitions as below:

$$p_{ij} = T_{ij} \Big/ \sum_{j=1}^{n} T_{ij} \tag{9}$$

where, T_{ij} is the one-step transition frequency from state i to state j in horizontal or vertical direction on the lattice. Conditional joint transition probabilities for the CMC can be further calculated based on Eq. (8).

In practical applications, an original map is normally unavailable for parameter estimation; thus, TPMs have to be estimated from soft information and hard data, such as expert knowledge, borehole data, existing maps delineated by other methods, or even information derived from similar areas. The surface boundary, if needed in a simulation, normally can be obtained from field transect survey or existing maps. When the number of boreholes is not sufficient, the vertical TPM estimated from boreholes may not be reliable and expert knowledge has to be used for adjustment. Although Markov chain TPMs are obviously more intuitive than indicator variograms, how to estimate reliable TPMs from soft information still depends on users' knowledge acquirement about their study area, and a trial-and-adjustment process may be needed. Rosen and Gustafson (1996) and Weissmann et al. (1999) all presented some suggestions for using soft information for parameter estimation in Markov chain modeling. The critical knowledge for estimating a TPM

includes: (1) proportions of categories, (2) mean lengths of certain categories in the direction, and (3) facies juxtapositional tendencies (Weissmann et al., 1999). Further research is needed to build a complete procedure easy to follow. In this study, to test the proposed method, we use one-step TPMs that are directly estimated from the reference map by counting state changes on the lattice, as normally done in model testing (Lin and Harbaugh, 1984; Elfeki and Dekking, 2001) so that the TPMs are representative for the simulations.

2.6 Implementation Procedures

Monte-Carlo sampling is used to conduct stochastic simulations. A simulation procedure for the proposed model is comprised of the following steps:

Step 1: The two-dimensional domain to be simulated is discretized as a grid.

Step 2: Known conditioning data (e.g., top boundary and borehole data) are inserted into their locations in the domain.

Step 3: If the upper boundary is not known, simulate it: if no point data available, use Equation (1) to extrapolate and the initial point can be chosen randomly or from the estimated proportions of different states; but if there are some point data along the upper boundary of the transect, use Equation (3) to interpolate between two points. If the boreholes are too sparse, consider inserting vertical lines between boreholes using the one-step Markov transition probabilities in Equation (1). Then use the one-dimensional Markov chain method in Equation (3) to produce bottom boundary and some interior lateral lines according to the predefined positions and number. Inserting how many simulated lines and where inserting them are decided by users. Note, interior boreholes and simulated data lines will split a domain into small windows (see Fig. 1).

Step 4: Generate the left unknown cells within each window row by row from the top-left corner to the bottom-right corner using the conditional joint transition probabilities (see Eq. (8))

Step 5: The procedure continues until all unknown cells in the two-dimensional domain are visited.

Step 6: Repeat step 3 to step 5 to produce the next realization.

Step 7: Calculate probability maps of each states and output results.

2.7 Study Example

A 5,250m long and 2m deep alluvial soil transect from the North China Plain is used as a study example to test the model (Fig. 2). This transect includes four states (i.e., four types of sediment textural layers), namely, sand, sandy loam, loam, and clay, denoted as 1, 2, 3 and 4, respectively, with strong layer extensions in the lateral direction. State 1, 2, 3, and 4, respectively, account for 30.84%, 28.02%, 17.18% and 23.96% of the transect area. So state 3 is relatively a minor one (i.e., infrequently occurring). State 2 dominates the upper part and state 1 the bottom part. The vertical transect is discretized into a grid of 310 columns and 43 rows. TPMs are directly estimated from this transect by counting transitions between different states on the grid and using Equation (9). The estimated TPMs in the horizontal (from left to right) direction and the vertical (from top to bottom) direction are shown in Table 1. Proportions of different states are given in Table 2. The TPMs directly estimated from the sample transect will be used as input parameters for testing the model.

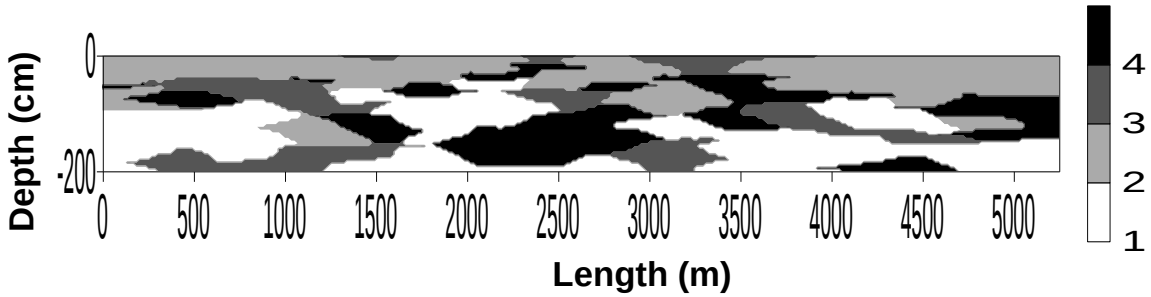

Figure 2. An alluvial soil transect with four types of sediment textural layers from the North China Plain.

Table 1. Input parameters estimated from the alluvial soil transect in Figure 2

	States: 4				Number of columns: 310			Number of rows: 43
	TPM[a] in the horizontal direction				TPM in the vertical direction			
Soil textural layer type	1	2	3	4	1	2	3	4
1	.9731	.0057	.0106	.0106	.9196	.0041	.0411	.0352
2	.0013	.9781	.0053	.0153	.0211	.8741	.0240	.0808
3	.0216	.0100	.9632	.0052	.0790	.0384	.8358	.0468
4	.0140	.0105	.0070	.9685	.0822	.0542	.0273	.8363

[a] Transition probability matrix

2.8. Simulation Schemes

We will simulate the soil transect using the proposed method and compare simulated results with those simulated using the previous CMC model to demonstrate the effectiveness of the proposed method for sparse borehole data. Simulations will be done under different conditioning schemes with respect to different numbers of boreholes and simulated lines. Vertical lines will be inserted only when the boreholes are very sparse or no boreholes are available. The data lines (including boreholes and simulated lines) will be distributed uniformly within the transect (i.e., with approximately equal intervals). Specific simulated schemes for each simulation will be shown in the labels of realizations and probability maps so that they are clearer to be seen. We will simulate 100 realizations for each simulation, but only give the first realization and the corresponding probability map of the most infrequently occurred state – the state 3 for most simulations. For all conditioning schemes, state proportions are averaged from the first 100 realizations in each simulation and corresponding computer run times are recorded.

2.9. Probability Maps

Probability maps will be used to show how likely a state occurs at every location in the simulated realizations or whether multiple realizations display similar patterns. A probability map is calculated as follows: When a state occurs at a location in a realization, its indicator value is denoted 1, otherwise 0. By dividing the sums of indicator values from multiple realizations with the number of realizations, we can get a probability map of a state with probability values from 0.0 to 1.0, which shows the probabilities of a state occurring at every location. The probability maps provided here are calculated from the first 100 realizations of each simulation.

3. Results

3.1 Zero Borehole

The CMC model of Elfeki and Dekking (2001) is not suitable for unconditional simulation, not only because it has to be conditioned on three boundaries, but also because of the deficiencies mentioned above. Unconditional simulation may be useful for understanding subsurface formations when there is no measured data available for conditioning but parameters may be estimated in other ways (e.g., from soft information). By inserting simulated lines generated by one-dimensional Markov chain methods, unconditional simulation can be done in the proposed method.

Table 2. Averaged area proportions of different components in simulated realizations under different conditioning schemes.

Model	No. of realizations	Conditioning schemes			Component proportions				Borehole interval (m)	Computer run time (Minute)[d]
		BH[a]	SVL[b]	SLL[c]	1	2	3	4		
Original	-[e]	-		-	.308	.280	.172	.240	-	-
mCMC	100	0	17	10	.338	.260	.173	.229	-	2
CMC	100	2	-	-	.321	.416	.037	.226	5250	14
mCMC	100	2	7	4	.362	.290	.132	.216	5250	2
mCMC	100	2	15	9	.354	.257	.170	.219	5250	2
CMC	100	4	-	-	.412	.290	.058	.241	1750	5
mCMC	100	4	9	4	.388	.251	.131	.230	1750	3
mCMC	100	4	13	9	.351	.254	.155	.241	1750	2
CMC	100	7	-	-	.418	.283	.074	.225	875	3
mCMC	100	7	6	4	.399	.250	.131	.220	875	3
mCMC	100	7	6	9	.360	.249	.155	.236	875	2
CMC	100	17	-	-	.331	.281	.145	.243	385	2
mCMC	100	17	0	4	.336	.271	.156	.237	385	3
mCMC	100	17	0	9	.311	.289	.164	.236	385	3
CMC	100	32	-	-	.314	.279	.173	.235	170	1
mCMC	100	32	0	4	.320	.274	.172	.234	170	2
mCMC	100	32	0	9	.308	.283	.175	.233	170	2

[a] Number of boreholes. [b] Number of simulated vertical lines. [c] Number of simulated lateral lines. [d] Recorded by computer. [e] Not applicable.

The advantage of the proposed mCMC method is the ability of conditioning on measured data, which is derived from the CMC methodology. Therefore we will not do completely unconditional simulation. Figure 3 presented simulation results using the mCMC method, which are only conditioned on 17 point data on ground surface; no borehole is used. All outer boundaries and inner lines are simulated by using one-dimensional Markov chain methods. Labels on realizations and probability maps denote the method used, number of boreholes, simulated vertical lines and lateral lines, and state (e.g., mCMC-0-17×10-3 means using the

proposed method, 0 borehole, 17 simulated vertical lines, 10 simulated lateral lines, and the state 3). This convention will be kept in all the following figures.

From Figure 3 it can be seen that plausible realizations (Two are shown here) can be generated. The probability map of state 3 shows the influence of the point data on the ground surface, where state 3 occurs at several points. State 4 is not occurred on the ground surface, therefore there are no preferred occurring locations in the probability map of state 4. The averaged areal proportions of different states in realizations are very close to the original data (Table 2). Computer run time (for 100 realizations) is just two minutes.

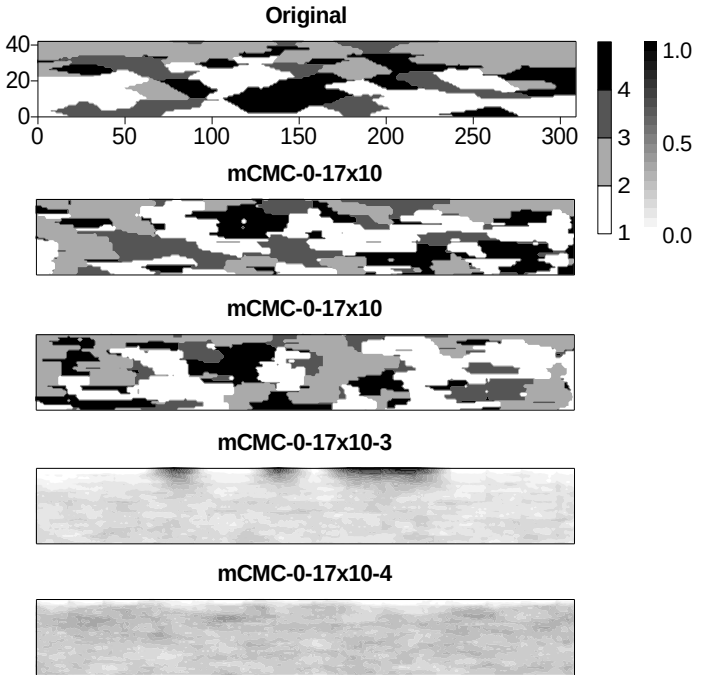

Figure 3. Simulated results conditioned only on surface point data using the mCMC model. All data lines are uniformly distributed (i.e., with equal intervals). Labels on realizations represent method used, number of boreholes, simulated vertical lines and simulated lateral lines. The last number in labels of probability maps represents the state (class) number. Following figures will follow this convention.

3.2 Two Boreholes

Because the upper and two side boundaries are necessary in the CMC model of Elfeki and Dekking, for the convenience of comparing results, all simulations in the following will assume the three boundaries are known.

Figure 4 gives simulated results with two boreholes, i.e., the two side boundaries. This means that the borehole interval is about 5250m, a very spare dataset. It can be seen that the layer inclination and under-estimation of state 3 are clear in the simulated results from the CMC model, but with help of simulated lines, these deficiencies are largely mitigated and the realizations becomes reasonably plausible. It also can be seen that the numbers of simulated vertical lines and lateral lines have influence on simulated realizations. State proportions of simulated results using the mCMC method (Table 2) are more close to the original data.

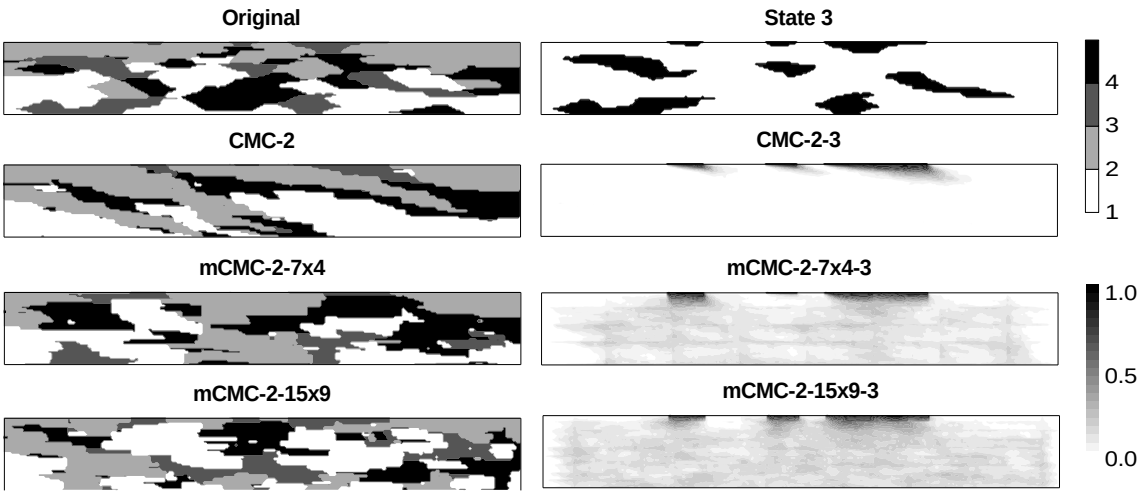

Figure 4. Simulated results using two boreholes (and the upper boundary) and different conditioning schemes. The top-right map is an indicator map of state 3. The upper boundary will always assumed known in the following figures. Every displayed realization is the first one in the simulation, same for the following figures.

3.3 Four Boreholes

Simulated results using 4 boreholes are given in Figure 5. This means a borehole interval of about 1750m. It can be seen that under this sparsity of borehole data the simulated results from the CMC model still have clear prediction artifacts, i.e., the layer inclination along the simulation ordering, and the minor state 3 is clearly under-estimated (correspondingly major states such as the state 1 is over-estimated). Simulated lines clearly mitigate these problems. The conditioning scheme with the simulation on the bottom row of Figure 5 gives more reasonably plausible realizations. The influence of the 4 boreholes can be seen in probability maps of the state 3, which show that the occurrence of the state is more certain near boreholes.

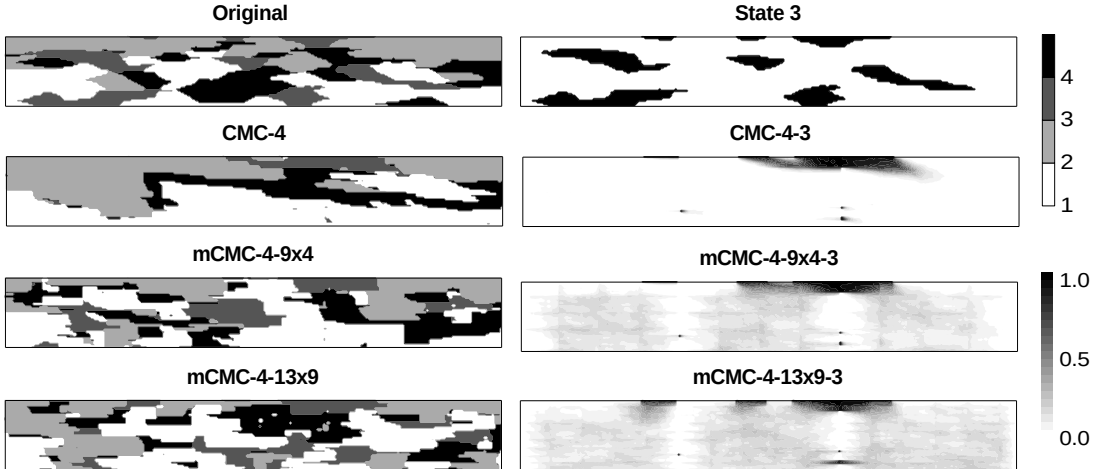

Figure 5. Simulated results using four boreholes but different conditioning schemes.

3.4 Seven Boreholes

In Figure 6 the number of boreholes is increased to 7, which means the borehole interval decreases to about 875m. We still can see the layer inclination problem (see Fig. 6, second row, left) and the under-estimation of the minor state 3 (see Fig. 6, second row, right) in the simulated results from the CMC model. But with inserted lines used in the mCMC method, the simulation effectiveness is improved a lot: it can be seen that with this density of boreholes the state 3 is more fairly represented in realizations and simulated realizations with suitable numbers of simulated lines (see Fig. 6, bottom) are also closer to the original.

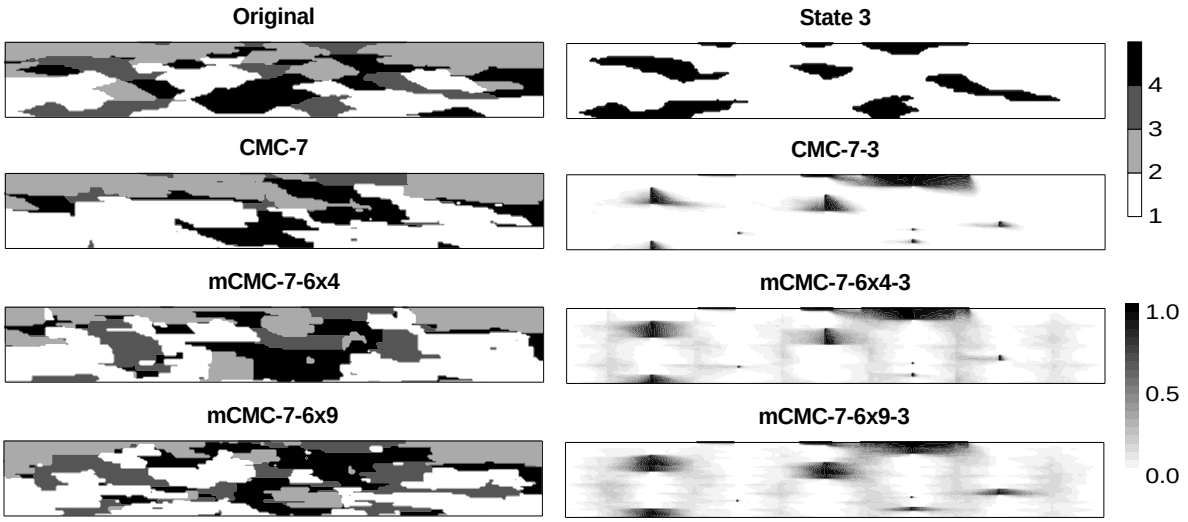

Figure 6. Simulated results using seven boreholes but different conditioning schemes.

3.5 Seventeen Boreholes

We use 17 boreholes in simulations shown in Figure 7. The borehole interval decreases to 385m. It can be seen that the CMC model has overcome the layer inclination problem under this density of boreholes, and the simulated realization is quite similar to the original. The under-estimation of minor states (and the over-estimation of major states) is also not obvious (Table 2). But inserting simulated lateral lines still have positive effect, no matter in mimicking the spatial patterns and representing areal proportions of different states.

But under this density of boreholes, inserting vertical lines is not suitable anymore, because the borehole interval is already obviously less than the layer mean lengths of most states, which means that the influences of neighboring boreholes on the simulations already can reach to each other. Under this situation inserting vertical lines will disturb the influence of borehole data because the inserted vertical line has no spatial correlation with the neighboring boreholes.

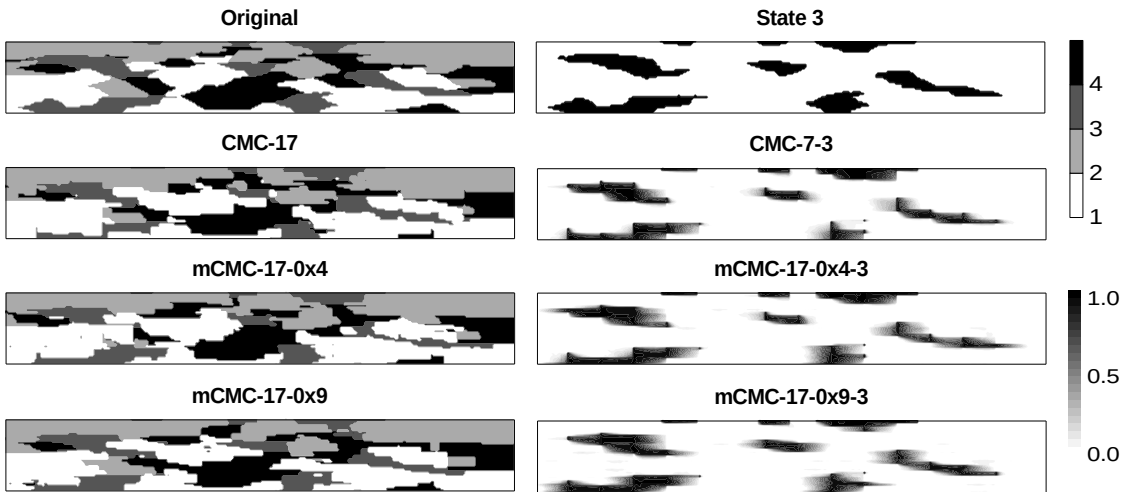

Figure 7. Simulated results using 17 boreholes but different conditioning schemes

3.6 Thirty-two Boreholes

When the density of boreholes further increases to an extremely high density, for example, the borehole interval decreases to 200m or so, the CMC model will be sufficient in mimicking the spatial patterns and representing the proportions of different states in the original soil transect. Inserting simulated lateral lines will be not very necessary. Figure 8 shows simulated results with 32 boreholes, i.e., a borehole interval of about 170m. We can see that the simulated realizations with or without simulated lateral lines are similar and all closely resemble the original (Table 2). But with the simulated lateral lines, it seems that layers in simulated realizations are more similar with their counterparts in the original map in terms of layer shape. In general, when boreholes are sufficient for the CMC model, inserting lateral lines will not help much but also has no bad effect.

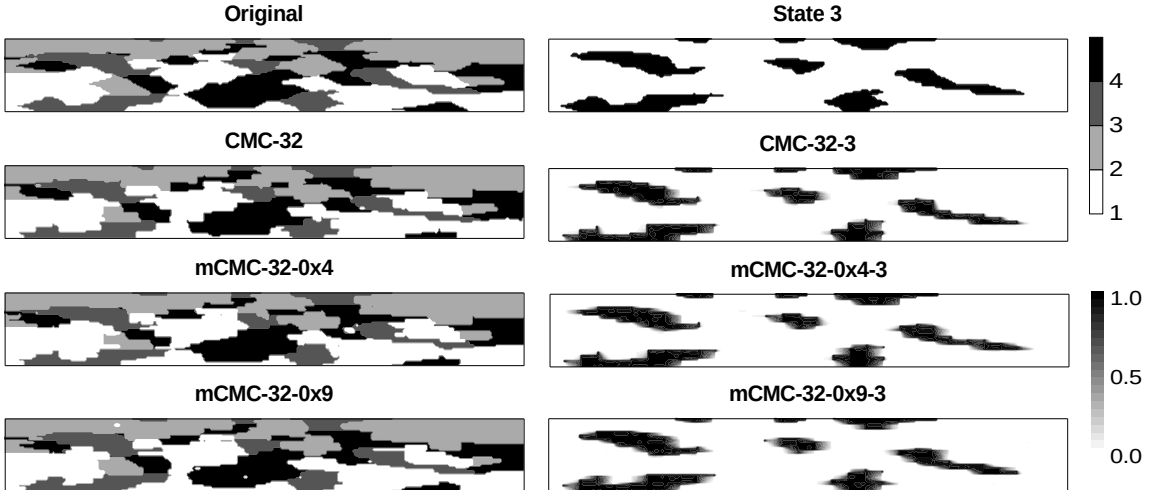

Figure 8. Simulated results using 32 boreholes but different conditioning schemes

4. Discussion

4.1 Tradeoffs

In the above simulations, with the number of boreholes gradually increasing it is demonstrated that simulated lines have obvious effect in mitigating the deficiencies of the CMC model. However, this mitigation comes with tradeoffs in the possible disturbance of spatial relationship of alluvial soil textural layers because the inserted simulated lines are not spatially correlated. The consequences are that long layers may be shortened because of the simulated vertical lines and thick layers may be thinned because of the simulated lateral lines. This tradeoff is clear when boreholes are very sparse and simulated vertical lines are inserted. When boreholes are not very sparse, boreholes may effectively hold the vertical sequence of layers (without fragmentation) because simulated lateral lines are actually conditioned on boreholes. Therefore, cautiously choosing the number of simulated lines according to the borehole density is necessary to avoid serious negative consequence. How to effectively correlate the simulated lines each other and with boreholes is an issue in next step of study.

4.1 Suitable Numbers of Simulated Lines

The suitable numbers of simulated vertical lines and lateral lines for the proposed method depend on specific applications; the same is true with the sufficient number of boreholes for the CMC model. Our simulations indicate that when a straight layer is crossed by more than two boreholes, this layer will usually be captured between the boreholes in the CMC model. This means that the influence of a borehole can be as far as the half length of the layer in the lateral direction. Therefore, if the borehole interval is obviously less than the length of most short layers, the number of boreholes will probably be sufficient for generating realizations without obvious inclination tendency.

This observation provides the evidence for suitable numbers of simulated vertical lines. This means that if simulated vertical lines are used, the suitable interval between simulated vertical lines or between simulated vertical lines and boreholes should be close to the layer mean-length. Particularly the interval should not be obviously shorter than the layer mean-length; otherwise, the simulated layers may be shortened by the excessive simulated vertical lines, because the simulated vertical lines have no spatial correlation with each other and with boreholes. Therefore, once borehole intervals are less than the layer mean-length, simulated vertical lines are not needed any more.

For simulated lateral lines, similar principle should be taken, i.e., the interval of simulated lateral lines should be close to the mean-thickness of layers. The interval should not be obviously thinner than the mean-thickness; otherwise the simulated layers may be fragmentized if boreholes cannot keep the thickness of layers. Inserting simulated lateral lines has no bad impacts when boreholes are sufficient or close to be sufficient. On the contrary, it still helps in shaping the simulated layers. The reason is that the densely distributed boreholes can keep the layer un-fragmentized by their influence.

Due to the mean length and thickness of layers are usually not known, users may have to try and see for a satisfied simulation scheme. Different users may have different requirements in capturing the major subsurface features and how close the areal proportions of different states in realizations should be to their expected values. Therefore, the so-called deficiencies are also relative to different users and application purposes.

4.2 Stationarity and Local Stationarity

Stationarity is a necessary assumption in geostatistics to account for the parameter estimation from spatial data because no repetitive data on the same location can be acquired. Such an assumption is suitable for a relatively small area or when spatial patterns are relatively identical in the study area. For a large area where spatial patterns may be very different in different subareas, to represent the different spatial patterns in different subareas, multiple sets of parameters may be necessary. Thus, when a large area is divided into subareas for simulation, the stationarity assumption only apply in each subarea. The tradeoff is that the workload and information needed for parameter estimation will increase. This, is always a tradeoff to be considered by users.

From the reference soil transect we used in this paper, it can be seen that the spatial patterns are obviously different in the upper half and the lower half. In the upper part, the state 2 dominates and state 1 seldom occurs; in the lower part, the situation is opposite. This means the stationarity assumption does not apply to the whole transect. If two sets of parameters are estimated, one for the upper part and the other for the lower part, there is no doubt that the simulated realizations will reflect this difference (i.e., layer type 1 will seldom occur in the upper part and layer type 2 will seldom occur in the lower part in realizations when borehole are few).

5. Conclusion

An idea – using simulated lines generated by one-dimensional Markov chain methods to increase the conditioning data for two-dimensional simulation using the extended CMC model, is presented with simulations of an alluvial soil transect. The main purpose is to deal with sparse borehole data in characterization of shallow subsurface alluvial soil textural layers. Therefore, the proposed method, which is based on the CMC theory, serves as a simple complement to the CMC model for different application purpose. In the proposed method, the CMC model is further extended to condition on future states in the vertical direction with the support of the proposed idea. The simulated lines, including simulated vertical lines and simulated lateral lines, are first inserted in a two-dimensional domain before two-dimensional simulation is conducted using the extended CMC model. These simulated lines further partition the simulation domain, which is already partitioned by boreholes, into smaller windows, and then two-dimensional simulation can be performed in each window. Through this way, the deficiencies of the CMC model, which are obvious in our simulation case when boreholes are sparse, are largely mitigated without increasing the number of borehole data. Consequently, hard data sufficiency is not a required condition any more in the proposed method for generating plausible realizations without obvious unpreferred artifacts and strong under-estimation of infrequent states.

An alluvial soil transect with four types of sediment textural layers is used as a simulation example for comparing the performance of the proposed method with that of the CMC model. Different conditioning schemes are used for simulations. For most schemes, the first realization and the corresponding probability map of the most infrequent state are displayed. Results demonstrate that the proposed method can produce more plausible realizations than the CMC model with the same number of boreholes when boreholes are sparse. The proposed method also can be used to produce realizations without borehole data.

Cautions must be taken for inserting simulated lines, because the simulated lateral lines have no spatial correlation between them in the vertical direction and the simulated vertical lines have no spatial correlation between them or with boreholes in the lateral direction. Inserting simulated vertical lines should be particularly cautious. The principle for inserting simulated lines should be that the line (including boreholes for vertical lines) interval shouldn’t be obviously shorter (or

thinner) than the mean-length (or mean-thickness) of layers; otherwise the layers may be seriously shortened (or thinned). The tradeoff for this simple and intuitive strategy is discussed.

A typical feature of the CMC model is its high efficiency in performance. The proposed method still keeps this advantage. No extra parameters are needed. The computation time of the proposed method does not increase or even decreases compared with the CMC model (see Table 2). Only minutes are needed for generating 100 realizations in the simulation case of this study.

While the proposed method provides a simple solution to mitigate the deficiencies of the CMC model for alluvial soil textural layer simulation, there are obvious tradeoffs. Another obvious shortcoming for the proposed method is that we have not developed an accurate answer for deciding how many simulated lines to insert, which has to depend on users.

References

Balzter, H., Markov chain models for vegetation dynamics, *Ecological Modelling*, *126*(2-3), 139-154, 2000.

Besag, J., Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Royal Stat. Soc.*, Series *B*, *36*, 192-236, 1974.

Besag, J., On the statistical analysis of dirty pictures (with discussions), *J. Royal Stat. Soc., Series B*, 48(3), 259-302, 1986.

Bierkens, M.F.P., and H.J.T. Weerts, Application of indicator simulation to modelling the lithological properties of a complex confining layer, *Geoderma*, *62*, 265-284, 1994.

Bogaert, P., Spatial prediction of categorical variables: The Bayesian Maximum Entropy approach, *Stoch. Env. Res. Risk A.*, *16*(6), 425-448, 2002.

Burgess, T.M., and R. Webster, Optimal sampling strategies for mapping soil types: I. Distribution of boundary spacings, *Journal of Soil Science*, *35*, 641-654, 1984a.

Burgess, T.M., and R. Webster, Optimal sampling strategies for mapping soil types: II. Risk functions and sampling intervals, *Journal of Soil Science*, *35*, 655-665, 1984b.

Carle, S. F., and G. E. Fogg, Transition probability-based indicator geostatistics, *Math. Geol.*, *28*, 453-477, 1996.

Chen, J., and Y. Rubin, An effective Bayesian model for lithofacies using geophysical data, *Water Resour. Res.*, *39*, 1118, 2003.

Deutsch, C. V. and A. G. Journel, *GSLIB: Geostatistics Software Library and user's guide*, 340 pp., Oxford Univ. Press, New York, 1997.

Descombes, X., R.D. Morris, J. Zerubia, and M. Berthod, Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood, *IEEE transactions on image processing*, *8*(7), 954-963, 1999.

Ehlschlaeger, C. R., Representing uncertainty of area class maps with a correlated inter-map cell swapping heuristic, *Computers, Environmant and Urban Systems*, *24*, 451-69, 2000.

Elfeki, A. M., *Stochastic characterization of geological heterogeneity and its impact on groundwater contaminant transport*, Ph.D. thesis, Delft University of Technology, Balkema publisher, The Netherlands, 1996.

Elfeki, A. M., and F. M. Dekking, A Markov chain model for subsurface characterization: theory and applications, *Math. Geol.*, *33*, 569-589, 2001.

Feyen, J., D. Jacques, A. Timmerman, and J. Vanderborght, Modeling water flow and solute transport in heterogeneous soils: A review of recent approaches, *J. Agric. Eng. Res.*. *70*, 231-256, 1998.

Galbraith, R.F., and Walley, D., On a two-dimensional binary process, *Journal of Applied Probability, 13*, 548-557, 1976.

Guyon, X., *Random fields on a network*, 255 pp., Springer, New York, 1995.

Harbaugh, J. W., and G. F. Bonham-Carter, *Computer simulation in geology*, 321 pp., Wiley-Interscience, New York, 1980.

Johnson, G.D., W.L. Myers, and G.P. Patil, Stochastic generating models for simulating hierarchically structured multi-cover landscapes, *Landscape Ecology*, *14*, 413-421, 1999.

Kyriakidis, P.C., and J.L. Dungan, A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions, *Environmental and Ecological Statistics, 8*, 311-330, 2001

Koltermann, E. C., and S. M. Gorelick, Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resour. Res., 32*, 2617-2658, 1996.

Li, W., 2-D stochastic simulation of spatial distribution of soil layers and types using the coupled Markov-chain method, 28 pp., Postdoctoral research report No.1, Institute for Land and Water Management, K.U. Leuven, 1999.

Li, W., B. Li, Y. Shi, and D. Tang, Application of the Markov-chain theory to describe spatial distribution of textural layers, *Soil Sci., 162*, 672-683, 1997.

Li, W., B. Li, Y. Shi, D. Jacques, and J. Feyen, Effect of spatial variation of textural layers on regional field water balance, *Water Resour. Res., 37*, 1209-1219, 2001.

Lin, C., and J.W. Harbaugh, *Graphic display of two- and three-dimensional Markov computer models in geology*, Van Nostrand Reinhold Company, New York, 1984.

McBratney, A.B., I.O.A. Odeh, T.E.A. Bishop, M.S. Dunbar, and T.M. Shatar, An overview of pedometric techniques for use in soil survey, *Geoderma, 97*, 293-327, 2000.

Murray, C.J., Identification and 3D modeling of petrophysical rock types, In *Geostatistics for Reservoir Geology*, edited by J. Yarus and R. Chambers. AAPG Mem., Am. Assoc. of Pet. Geol., Tulsa, Okla, 1994.

Norberg, T., L. Rosen, A. Baran, and S. Baran, On modeling discrete geological structure as Markov random fields, *Math. Geology, 34*, 63-77, 2002.

Patil, G.P. and C. Taillie, A multiscale hierarchical Markov transition matrix model for generating and analyzing thematic raster maps, *Environmental and Ecological Statistics, 8*, 71-84, 2001.

Pickard, D.K., Unilateral Markov fields, *Advances in Applied Probability, 12*, 655-671, 1980.

Rosen, L., and G. Gustafson, A Bayesian Markov geostatistical model for estimation of hydrogeological properties, *Ground water, 34*, 865-875, 1996.

Switzer, P., A random set process in the plane with a Markovian property (Note), *Ann. Math. Stat., 36*, 1859-1863, 1965.

Tjelmeland, H., and J. Besag, Markov random fields with higher-order interactions, *Scand. J. Stat., 25*, 415-433, 1998.

Weissmann, G. S., S. F. Carle, and G. E. Fogg, Three-dimensional hydofacies modelling based on soil surveys and transition probability geostatistics, *Water Resour. Res., 35*, 1761-1770, 1999.

Weissmann, G. S., and G. E. Fogg, Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework, *Journal of Hydrology, 226*, 48-65, 1999.

Wu, K., N. Nunan, J. W. Crawford, I. M. Young, and K. Ritz, An efficient Markov chain model for the simulation of heterogeneous soil structure, *Soil Sci. Soc. Am. J., 68*, 346-351, 2004.
Zhang, J., and M. Goodchild, *Uncertainty in geographical information*, Taylor & Francis, New York, 2002.