Application of the coupled Markov chain model to categorical soil data: Feasibility and constraints

Weidong Li [1], Chuanrong Zhang [2]
[1] *Department of Mathematics, Statistics and Computer Science, Marquette University*
[2] *Department of Geography, University of Wisconsin- Milwaukee*

Summary
Spatial pattern characterization of categorical soil variables is crucial for predictive soil mapping and environmental management at watershed scales. This study attempts to evaluate the coupled Markov chain (CMC) model for characterizing the complex spatial patterns of categorical soil variables. Two simulation cases about soil types and alluvial soil textural layers are conducted for evaluating the feasibility and constraints of the CMC model for mapping categorical soil variables from survey line samples. Results indicate that the CMC model can capture the spatial patterns of categorical soil variables under a high density of transect line survey data or borehole data. However, when the density of survey data is relatively low, some obvious prediction artifacts such as parcel (or layer) inclination along the simulation direction show, and minor or small states (soil classes) are obviously underestimated. The possible reasons for these constraints are discussed. Such an efficient model is desirable for modeling categorical variables over moderate to large areas, but further efforts are needed to improve it in dealing with the data sparseness and eliminating the prediction artifacts.

Introduction
Categorical soil data, such as subsurface soil layers and surface soil classes, are crucial in land management, precision farming, and environmental analysis and modeling over watershed scales (Kite & Kauwen, 1992; Feyen *et al*., 1998; Li *et al*., 2001; Zhu & Mackay, 2001; Bouma *et al*., 2002). Categorical spatial variables are usually composed of multiple nominal classes with complex spatial patterns. The special features of categorical spatial variables in earth sciences, such as soil types, are revealed in their complex spatial dependence, site-specific sequences, anisotropies, and abrupt boundaries between multinomial classes. This complexity poses challenging problems for many spatial simulation methods, such as indicator geostatistics (Gomez-Hernandez & Srivastava, 1990; Deutsch & Journel, 1998), to effectively mimic them over watershed scales from limited samples (Wingle & Poeter, 1993; Bierkens & Weerts, 1994; Weissmann & Fogg, 1999; Zhang & Goodchild, 2002). Typical problems, for examples, include (i) the high demand in computation (Ehlschlaeger, 1998, 2000), which prevents these methods from applying over watersheds or larger areas for modeling multiple classes with required resolutions, (ii) the difficulty to deal simultaneously with sharp boundaries and auto-correlation (Mowrer & Congalton, 2000; McBratney *et al.*, 2000), and (iii) the difficulty to cope with cross-correlations between multinomial classes within simulation algorithms (Goovaerts, 1997). It is desirable to find a highly efficient method capable of mimicking spatial patterns of categorical soil variables from sampling data so that categorical soil information can be effectively assessed and integrated into environmental modeling. Heuvelink & Webster (2001) presented a thorough review on methods for modeling soil variations. Scull *et al*. (2003) specifically reviewed the methods for predictive soil mapping. Some significant methods proposed for dealing with various categorical soil data and their application cases can be seen in Bierkens & Burrough (1993a, b), Bierkens & Weerts (1994), Carle & Fogg (1996), Lark & Webster (2001), and Zhu et al. (2001).

Markov chain methods represent an important approach for heterogeneity characterization of categorical variables. It was thought that compared with indicator variograms (Goovaerts, 1997; Deutsch

& Journel, 1998), Markov transition probabilities are more intuitive, easier to interpret, and therefore, relatively easier to incorporate soft information, particularly expert knowledge such as facies length, proportions, and juxtaposition relationships, into parameter estimation (Weissmann & Fogg, 1999; Weissmann *et al.*, 1999). However, Markov chains are conventionally used for one dimension. Typical application examples of one-dimensional Markov chains in some fields of geosciences can be seen in Harbaugh & Bonham-Carter (1980) for synthesizing stratigraphic sequences, in Todorovic & Woolhiser (1975) for simulating year precipitation processes, and in Balzter (2000) for species change. Burgess & Webster (1984a, b) first applied Markov chain transition probabilities in soil science to describe the spatial order of parcels of different soil classes in one-dimension. They estimated transition probabilities from linear transects across some soil maps and applied them to optimal sampling strategies for mapping soil types. Li *et al*. (1997, 1999) used one-dimensional Markov chain methods to describe the vertical changes of alluvial soil textural layers and simulate alluvial soil textural profiles. They estimated transition probabilities from a number of soil boreholes, which are thought to be individual realizations of a short Markov chain. Although multi-dimensional Markov chain methods for unconditional simulation have a longer history (Lin & Harbaugh, 1984; Johnson *et al*., 1999), conditional multi-dimensional Markov models emerged only recently in geosciences as a tool for characterizing categorical variables (e.g., lithofacies) (Elfeki & Dekking, 2001; Norberg *et al*., 2002). By now we have not seen any attempts to apply multi-dimensional Markov chain methods in modeling categorical soil variables.

The Markov random field (MRF) model of Norberg *et al*. (2002) is relatively complex and high demanding in computation because of its heavily iterative simulation algorithm. For example, for an area of about 100×100 pixels (i.e., grid cells), 1.5 to 2.5 days of run time on a SUN workstation are needed for producing a realization, as mentioned by the authors. Other deficiencies include (a) the under-estimation of minor states, and (b) the inability to reproduce the transition probabilities of the original map, which makes the method ill-suited for uncertainty analysis by means of repeated simulations, as also mentioned by the authors. These constraints make it currently not suitable for modeling categorical soil variables over large areas with required resolutions from limited samples. On the contrary, the coupled Markov chain (CMC) model of Elfeki & Dekking (2001) uses an explicit non-iterative algorithm (i.e., one pass one realization), which makes it highly efficient. The study cases of Elfeki & Dekking (2001) only demonstrated that the CMC model could capture the major features of subsurface geological formations when a number of well data were conditioned but did not talk about its deficiencies. Some shortcomings, such as layer disconnectedness along borehole lines and under-prediction of minor/small states can be directly seen from the simulated results given by the authors. It is not clear whether or not and if applicable under what conditions such an efficient model can be used to characterize the spatial heterogeneity of categorical soil variables.

In this paper, we will attempt to extend and directly use the CMC model for modeling categorical soil variables over watershed scales. Study cases, respectively about soil types and alluvial soil textural layers, will be given. Different conditioning schemes (i.e., different densities of sampled line-data) will be used to study the applicable conditions and the constraints for applying the model in producing categorical soil data.

Methods

The CMC model

The detailed explanation of the CMC model can be seen in Elfeki & Dekking (2001). Here we only give a simple introduction and the final equations. Because the CMC model is developed for characterizing subsurface geological formations in vertical sections, simulation is only conditioned on well data and the upper boundary.

The CMC model is composed of a horizontal one-dimensional Markov chain and a vertical one-dimensional Markov chain, which are perpendicular to each other in a two-dimensional domain (Figure 1). The two one-dimensional Markov chains are assumed to be stationary and fully independent of each other. Assuming that X_i and Y_j represent the horizontal one-dimensional Markov chain and the vertical one-dimensional Markov chain, respectively, both defined on the state space $[S_1, S_2, ..., S_n]$. These two Markov

chains form a CMC $Z_{i,j}$ with the joint transition probability 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 $Z_{i,j}$ being expressed as

$$p_{lm,k} = p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m) = \frac{p_{lk}^x \cdot p_{mk}^y}{\sum_f p_{lf}^x \cdot p_{mf}^y}, \quad k = 1, \ldots, n \qquad (1)$$

where p_{lk}^x is the one-step transition probability of the chain X_i in the x-direction (i.e, the horizontal direction for a vertical section), p_{lk}^y is that of the chain Y_i in the y-direction (i.e, the vertical direction for a vertical section), and superscripts x and y stand for directions. Equation (1) is the initial version of the CMC model developed by Elfeki (1996), which was used in Li (1999) with unsatisfied results; here we denote it as CMC_0. This model only conditions to the upper boundary (i.e., surface) and one side boundary (i.e., one well). These two boundaries are also necessary for conducting a two-dimensional simulation using the CMC approach. Simulation is performed from one upper corner (e.g., top-left) to the diagonal corner (e.g., bottom-right).

If a "future" state q at location $Z_{Nx,j}$ is known in the x-direction, to consider its influence to the current unknown state k, the one-dimensional Markov chain in the x-direction, X_i, can be approximately but cheaply conditioned on this future state. The expression of the conditional joint transition probability in such a conditional CMC $Z_{i,j}$ (Figure 1) is given by

$$p_{lm,k|q} = p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m, Z_{N_x,j} = S_q)$$
$$= \frac{p_{lk}^x \cdot p_{kq}^{x(N_x - i)} \cdot p_{mk}^y}{\sum_f p_{lf}^x \cdot p_{fq}^{x(N_x - i)} \cdot p_{mf}^y} \qquad (2)$$

where $p_{kq}^{x(N_x - i)}$ is the (N_x-i) step transition probability of the one-dimensional Markov chain X_i in the x-direction, which can be calculated by imposing a (N_x-i) power to the one-step transition probability matrix (TPM). Other symbols are similar to those in Equation (1). When cell N_x is far from cell i the terms $p_{lq}^{x(N-i)}$ and $p_{fq}^{x(N-i)}$ will have no influence because they are almost equal. However, when simulation gets closer to cell N_x, its state will start to play a role and the simulation result will be affected by the state at that cell (Elfeki & Dekking, 2001).

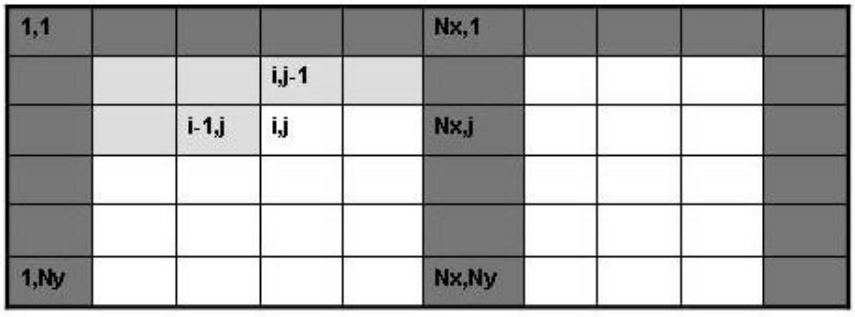

Figure 1 A coupled Markov chain conditioned on borehole lines. Dark gray means known data. Shallow gray means simulated data.

Equation (2) represents the CMC model in Elfeki & Dekking (2001) for subsurface characterization, here denoted as CMC_x. Conditioning on the other boundary and internal well data as future states is obviously a major improvement in this model, which, to a large extent, promotes this model to a two-dimensional stochastic method for conditional simulation. Such a model is directly suitable for simulating

alluvial soil textural layers in vertical transects by conditioning a simulation on soil borehole data, which are equivalent to lithological well data.

For characterizing the spatial distribution of soil types on the ground surface, we use survey lines to replace boreholes. Because survey lines can be obtained in both the *x*-direction and the *y*-direction, the CMC model will be extended to condition on survey line data in these two directions (Figure 2). Such an extension is straightforward. Assuming that S_o is the state at cell $Z_{i,Ny}$ on the opposite boundary in the *y*-direction, the conditional joint transition probability of the CMC with conditioning on future states in the *x*- and *y*-directions can be given as

$$p_{lm,k|qo} = p(Z_{i,j} = S_k \mid Z_{i-1,j} = S_l, Z_{i,j-1} = S_m, Z_{N_x,j} = S_q, Z_{i,Ny} = S_o)$$
$$= \frac{p_{lk}^{x} \cdot p_{kq}^{x(N_x-i)} \cdot p_{mk}^{y} \cdot p_{ko}^{x(N_y-j)}}{\sum_f (p_{lf}^{x} \cdot p_{fq}^{x(N_x-i)} \cdot p_{mf}^{y} \cdot p_{fo}^{y(N_y-j)})} \quad (3)$$

where $p_{ko}^{x(N_y-j)}$ is the (N_y-j) step transition probability of the one-dimensional Markov chain Y_j in the *y*-direction. Other symbols are similar to those in Equation (3). Note that the full independency assumption of the two one-dimensional Markov chains is used in derivation of all of the above three equations.

In such a CMC model, survey lines in the *x* and *y* directions will divide a two-dimensional domain into sub-domains, which we will call "windows" in this paper. Here we denoted this CMC model as CMC_{xy}. Two-dimensional simulation can be performed in each window using Equation (3) with conditioning on the four window boundaries (Figure 2).

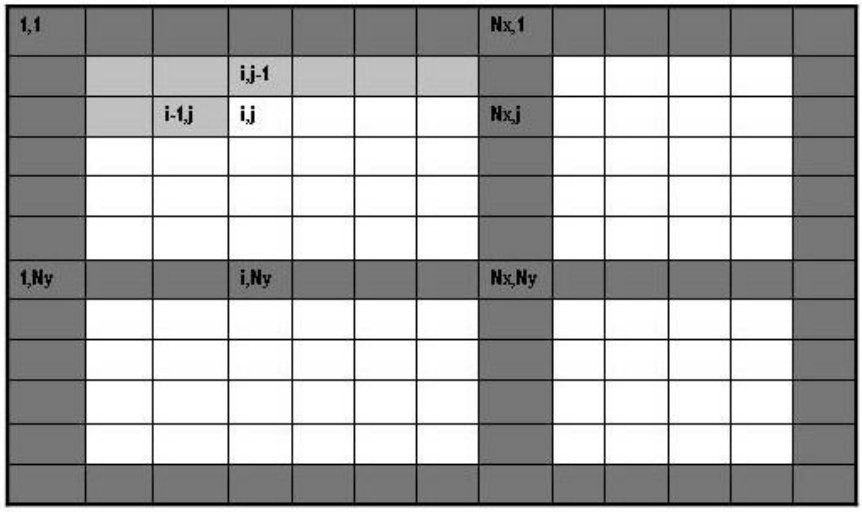

Figure 2 A coupled Markov chain conditioned on survey lines in two directions. Dark gray means known cells. Shallow gray means simulated cells.

Parameter inference

The main input parameters are the one-step TPMs of the two one-dimensional Markov chains. Other input parameters include the number of states, the number of the domain grid columns and the number of the domain grid rows. The cell size and simulation domain extents are needed only for discretizing the simulation domain and displaying realization images with the true length and width. In the following simulated images we will only show the grid column and row numbers.

For a soil system represented by a Markov chain, one has to first define the set of possible states of the system, [S_1, S_2,…, S_n], and the one step transition probabilities (i.e., the two TPMs) for the two one-dimensional Markov chains. The state space can be determined according to actual needs, rather than the soil taxonomy. For example, soils can be according to the research purpose (e.g., hydrologic modeling)

classified into several types or classes. The transition probabilities can be determined by superimposing a lattice on soil maps that are representative for the simulation and counting the state changes in different directions. The cell size (square or rectangle) should not be larger than the smallest parcel size to be shown in simulated realizations and should be the same for both parameter estimation and simulation.

The transition frequencies between the states in the *x*- and *y*-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 can be obtained by dividing the transition frequencies with the total number of transitions as below:

$$p_{ij} = T_{ij} \Big/ \sum_{j=1}^{n} T_{ij} \tag{4}$$

where, T_{ij} is the transition frequency from state i to state j in the *x*-, and *y*-direction on the lattice. Joint transition probabilities for a conditional CMC can be further calculated based on the above equations.

Borehole data recording soil layers normally has no gap. For modeling soil types, data for conditioning should also be continuous (i.e., no gap) lines. The requirement of conditioning on survey lines is special comparing with other methods, but might be advantageous given the prominence of transect sampling. Please note that TPMs may be estimated from various soft data, as mentioned by many researchers (Rosen & Gustafson, 1996; Weissmann & Fogg 1999; Weissmann *et al*., 1999). Here we will not explore this subjective problem but rather only test the feasibility of applying this method on categorical soil data modeling.

Simulation procedure

Monte-Carlo sampling method is used to conduct stochastic simulations using the above models. A simulation procedure involves the following set of instructions, which is similar to those introduced by Elfeki & Dekking (2001):

Step 1: The two-dimensional domain to be simulated is discretized using a predefined grid cell size.

Step 2: Survey line data are inserted into the simulation area (including outer boundaries). Note, internal survey lines (or boreholes) will divide a simulation area into windows, and the simulation will proceed in each window one by one in the same way.

Step 3: Generate the internal unknown cells with numbers (i, j), i=2, ..., $N_{x\text{-}1}$ and j=2, ..., $N_{y\text{-}1}$ in each window row by row using the conditional joint transition probability distribution.

Step 4: The procedure continues until all unknown cells in the two-dimensional domain are visited.

Step 5: Repeat step 3 and 4 to produce the next realization.

Probability map

We will use probability maps to show the spatial uncertainty of every sate (soil or layer type), namely how likely a component occurs at every location in a simulated area. A probability map is calculated like this: When a state occurs at a location (i.e., a cell) in a realization, it is counted. By dividing the total number of occurrence of a state at a location counted from many realizations with the realization number, we can get an occurrence probability of the state at the location. Thus, a probability map can be obtained for each state through visualizing all its probability values on every location. The provided probability maps in the study are calculated from 100 realizations.

Simulation cases

Case one – a soil map with seven classes

First we simulate a soil map with a length of 8km and a width of 1.7km. This map is cut from a watershed soil map in a river basin of Belgium. The soil is classified into seven types. The soil map is shown in Figure 3, top. We can see that the soil types have very complex spatial patterns and that different types account for different areal proportions, namely, some types occur frequently but some other types occur infrequently. The map is discretized into a 160×34 grid with a cell size of 50m. Input parameters, i.e.,

TPMs directly estimated from the map are given in Table 1. We use the map as a reference to test whether the CMC_{xy} model can work with soil class map simulation.

Table 1 Input parameters estimated from the original soil type map

	States: 7							Grid column number: 160						Grid row number: 34
	TPM[a] in the *x*-direction							TPM in the *y*-direction						
Soil type	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	.838	.030	.055	.010	.046	.012	.009	.838	.023	.068	.003	.048	.012	.008
2	.202	.614	.019	.013	.133	.006	.013	.173	.669	.000	.000	.135	.015	.008
3	.098	.004	.804	.009	.023	.051	.011	.103	.012	.801	.012	.013	.050	.008
4	.033	.000	.041	.633	.131	.077	.086	.041	.012	.016	.592	.114	.102	.122
5	.050	.010	.002	.014	.849	.049	.025	.067	.013	.003	.014	.835	.042	.027
6	.029	.000	.051	.025	.083	.732	.080	.028	.003	.050	.048	.107	.708	.056
7	.040	.000	.004	.055	.139	.090	.672	.031	.007	.007	.051	.113	.109	.682

[a] Transition probability matrix.

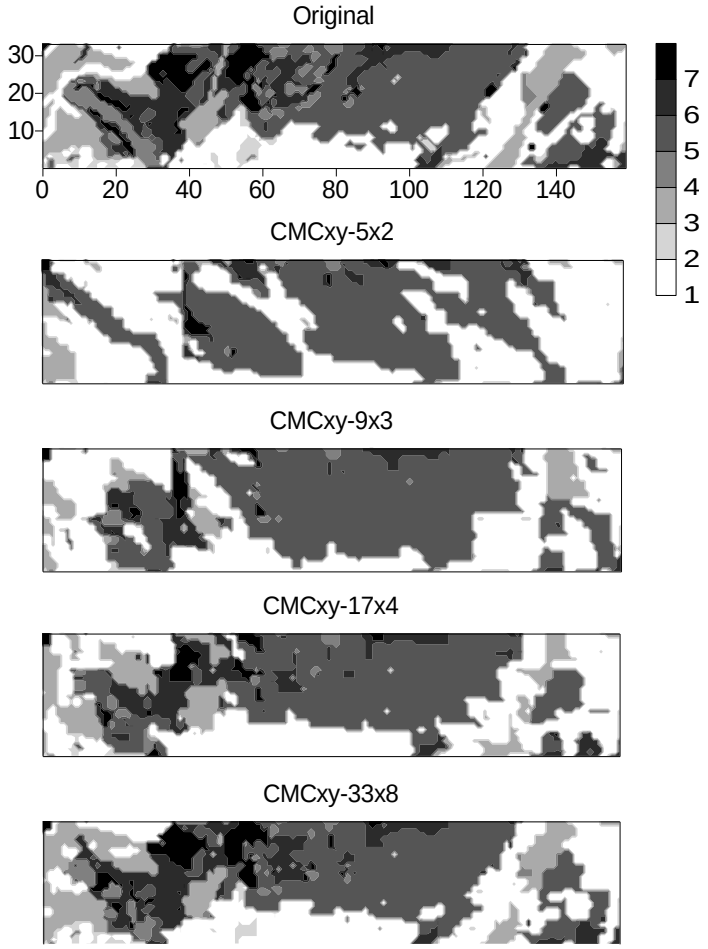

Figure 3 Simulated realizations of a soil map with 7 classes using the CMC_{xy} model under different conditioning schemes. Symbols of realizations represent the model used and numbers of survey lines in the *y*-direction and *x*-direction. Corresponding survey line intervals for the four realizations from the second to the last one are approximate 2000m, 1000m, 500m and 250m. Scales on the original map represent grid numbers.

Simulated realizations under different conditioning schemes (they are also the sampling schemes) are given in Figure 3. The survey line intervals used for this conditional simulation are approximately 2000m, 1000m, 500m, and 250m, respectively for the four conditioning schemes, i.e., 5×2, 9×3, 17×5, and 33×8, which represent the survey line number in the *y* direction and the line number in the *x* direction (see the

realization symbols in Figure 3). We can see that when the survey line interval is long the patterns in simulated realizations (see Figure 3, CMC_{xy}-5×2 and CMC_{xy}-9×3) are not reasonable: parcel inclination is clear, at some places the steep parcel disconnectedness appears, and major soil types (e.g., type 1 and 5) are over-predicted. But when the density of survey lines increases, i.e., decreasing the survey line interval, simulated soil patterns gradually become close to the reference map. The realization with a 250m survey line interval (see Figure 3, CMC_{xy}-33×8) resembles the reference map well, except for some very fine features. The parcel inclination artifact is not completely eliminated even with a very high density of survey line data (see the saw-teeth along the simulation direction in the left parts of the bottom two realizations in Figure 3).

The occurrence probability maps of specific soil types (Figure 4), which are calculated from 100 realizations, indicate that how possible one soil type predicted from survey line data occurs at each location. The parcel inclination problem also can be seen in some probability maps when survey lines are sparse. From the probability maps it can be seen that the minor soil type 3 (Figure 4, left column) is under-predicted and that the major soil type 1 is over-estimated when survey lines are sparse. But this problem gradually lessens with increasing the density of survey lines. Figure 5 gives the areal proportions of different soil types under different conditioning schemes. It can be seen that the predicted areal proportions of different soil types at the 250m survey line interval are close to the original data.

Although the CMC_{xy} model has difficulty to provide correct areal proportions of different soil types and the predicted soil type patterns have some simulation artifacts, the simulated results given in this study at survey line intervals of about or less than 500m should be useful for predictive soil mapping. They may not be suitable to be directly used as soil maps, however, with local soil experts' help, high quality soil type maps may be generated based on the simulated results. Problem is that the required density of survey lines may be too high for real world soil surveys. But when the survey line interval is too big, e.g., 1000m or larger for this study case, the simulated results will be useless or even misleading.

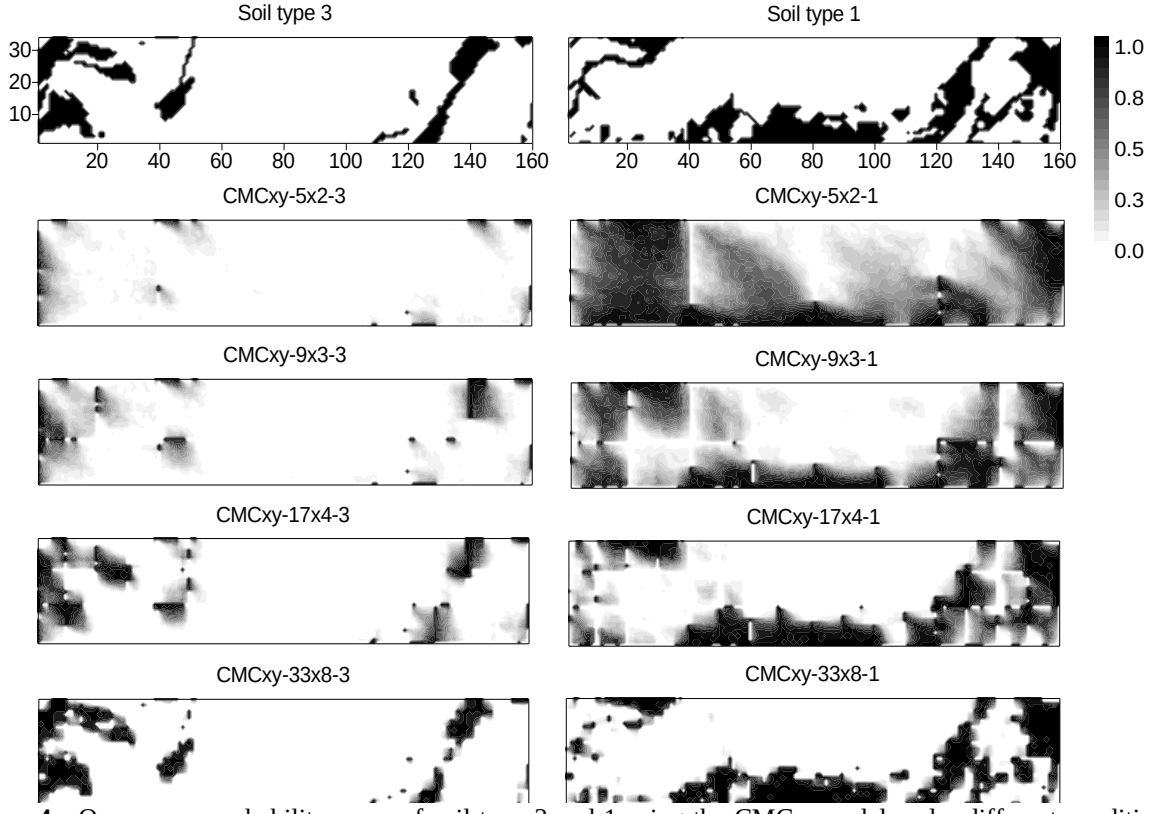

Figure 4 Occurrence probability maps of soil type 3 and 1 using the CMC_{xy} model under different conditioning schemes. Symbols represent the model used, survey line numbers in the *y*-direction and *x*-direction, and soil type.

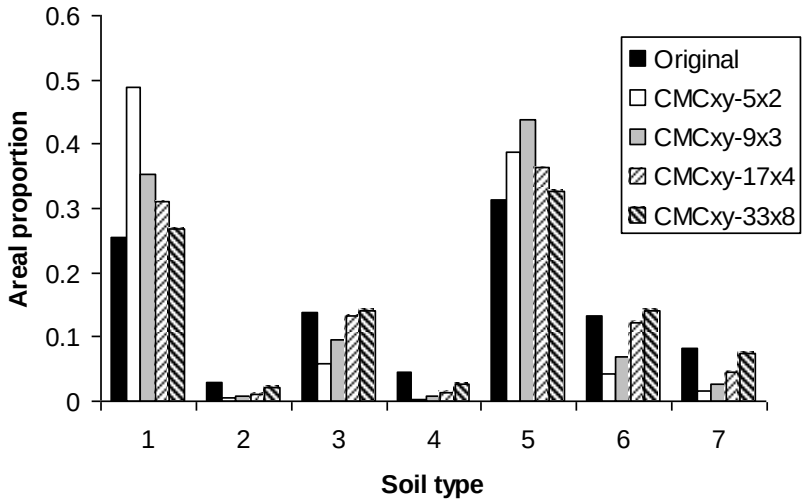

Figure 5 Areal proportions of different soil types in simulated results using the CMC_{xy} model under different conditioning schemes. Symbols represent the model used and survey line numbers in the *y*-direction and *x*-direction.

Case two – an alluvial soil transect with 4 textural layer types

The selected alluvial soil transect comes from the North China plain. It has a length of about 5250m and a depth of 2m, with four types of alluvial textural layers. The four types of layers, denoted as 1, 2, 3, and 4, account for 30.8%, 28.0%, 17.2%, and 24.0% of the whole transect area, respectively. So the type 3 occurs infrequently relative to others, and the type 1 is contrary. These soil textural layers have strong horizontal extensions. Input parameters estimated from the soil transect are given in Table 2.

Table 2 Input parameters estimated from the original alluvial soil transect

	States: 4				Grid column number: 310			Grid row number: 43
	TPM [a] in the horizontal direction				TPM in the vertical direction			
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.

Because it is not suitable to assume that subsurface lateral information can be available, we only use the CMC_x model to do this simulation. Figure 6 shows simulated realizations of the transect under different conditioning schemes, i.e., 4, 7, 17 and 32 boreholes, which correspond with borehole internals of about 1750m, 875m, 330m and 170m, respectively. It can be seen that when boreholes are sparse the simulated realizations clearly have the layer inclination tendency, and the minor (i.e., infrequent) layer type 3 is under-predicted. But when the density of boreholes is dense (see the bottom two realizations in Figure 6), the layer inclination problem is not clear. The complex spatial patterns of soil textural layers are effectively mimicked in the last realization with very dense borehole data. This situation also can be seen in the occurrence probability maps of layer type 2 and 3 in Figure 7.

The occurrence probability maps of individual layer types show that different realizations under the same conditioning scheme are quite similar. The uncertainty of occurrence of each type of soil layers decreases with the increase of density of boreholes. The under-prediction of the minor layer type 3 under low densities of boreholes can be seen in the left column of Figure 7. Figure 8 displays the histogram of areal proportions of different layer types in simulated results by different densities of boreholes.

Obviously, it can be seen that layer type 1 and 3 are abnormally predicted when boreholes are not enough. With the predicted artifacts and the abnormal estimation of major and minor types, the simulated results with inadequate borehole data will be useless or misleading. Therefore, to produce useful or valuable subsurface soil layer data, borehole data must be adequate, i.e., the required density of boreholes should be guaranteed. In this simulation case, the borehole interval should be about 400m or less. This density may be a quite high requirement for field soil surveys over large areas.

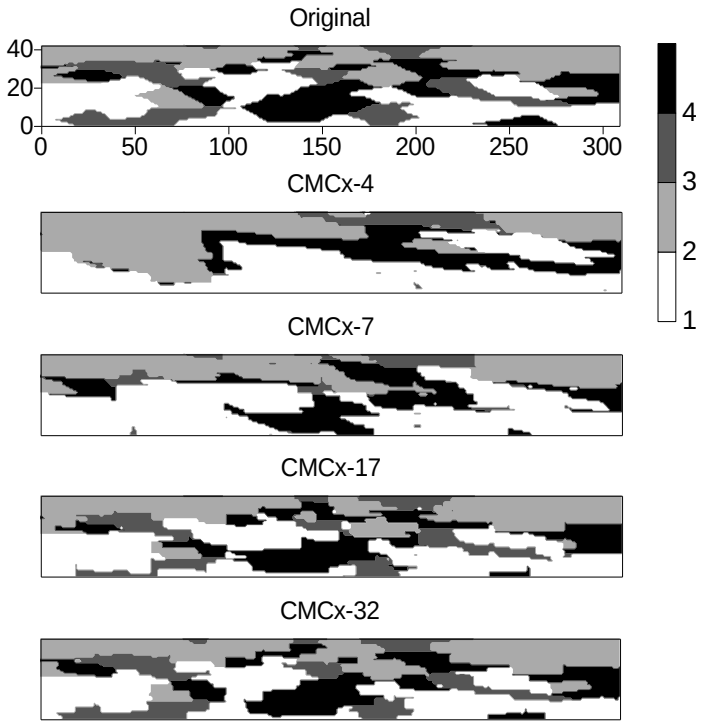

Figure 6 Simulated realizations of an alluvial soil transect with 4 types of textural layers using the CMC_x model under conditioning schemes. Symbols of realizations represent the model used and borehole numbers. Corresponding borehole intervals for the four realizations from the second to the last one are approximate 1750m, 875m, 330m and 170m. Scales on the original map represent grid numbers.

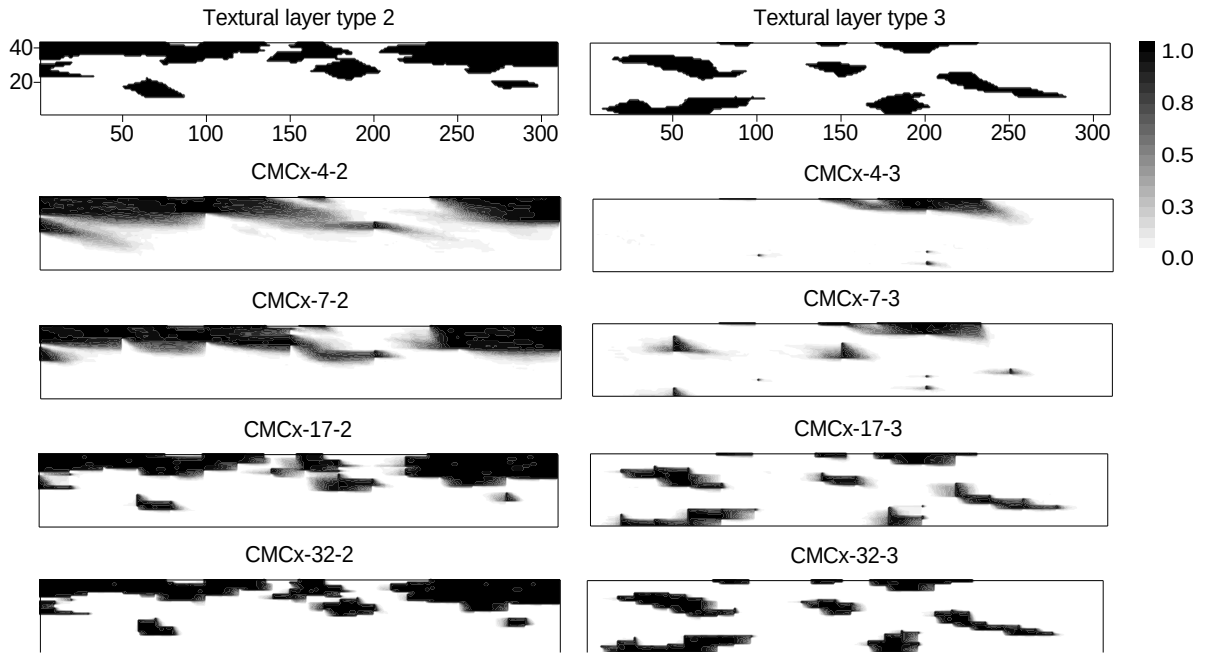

Figure 7 Occurrence probability maps of soil textural layer type 2 and 3 using the CMC_x model under different conditioning schemes. Symbols of realizations represent the model used, borehole numbers, and layer type.

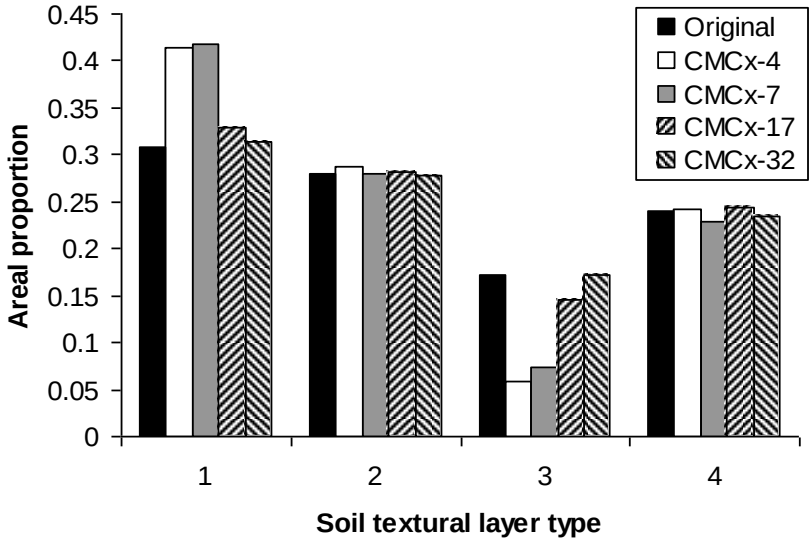

Figure 8 Areal proportions of different soil textural layer types in simulated results using the CMC_x model under conditioning schemes. Symbols represent the model used and borehole numbers.

Discussions

From the above simulated results, we can see that the CMC method with suitable extensions (i.e., CMC_x for vertical transects and CMC_{xy} for surface soil class maps) indeed can capture the spatial patterns of categorical soil variables if there is a set of densely measured line-data available. But if the density of line-data is low, three constraints appear, which include (i) under-prediction of minor/small states, (i) parcel (or layer) inclination along the simulation direction, and (iii) parcel (or layer) disconnectedness along survey lines (or boreholes). There is no doubt that the CMC method has important values due to its ability to capture major subsurface lithological layers, and it is also true that the so-called constraints of a

method are actually relative to different application purposes. But it would be very desirable to have such an efficient model also being able to reproduce spatial patterns of categorical (not just soil) variables with correct proportions of all states and without the prediction artifacts under low densities of line-data; after all, sparse data are the usual cases in the real world applications.

The parcel/layer inclination is clearly related with the simulation ordering, i.e., from one corner to another diagonal corner row by row. Koltermann & Gorelick (1996) ever mentioned the two difficulties in applying Markov chains to multi-dimensional simulation: one is the difficulty of conditioning on multi-boundaries, and the other is the ordering problem, as stated: "*Unlike one-dimensional application of Markov chains, two- and three-dimensional applications are difficult because there is not an easily identifiable ordering of values in a past-present-future sequence* (A. G. Journel, personal communication 1995)" (Koltermann & Gorelick, 1996, pp.2631). With the extensions of conditioning on well/borehole data or survey-line data, it can be seen that the CMC method has preliminarily resolved the problem of conditioning on measured data. But the ordering problem is not addressed, which results in an asymmetric simulation algorithm in the CMC model and the difficulty of retaining class parcel anisotropies in realizations. The parcel inclination is always along the simulation direction, i.e., parcels are inclined from top-left to bottom-right if the simulation is along this direction, or inclined from top-right to bottom-left if the simulation ordering is changed to this direction (Li, 1999). If layers (or parcels) are naturally inclined along the simulation direction (i.e., the so-called preferred inclination), their anisotropy might be well maintained (See the hypothetical example in Elfeki & Dekking (2001)). The MRF method (Norberg *et al*., 2002) avoids this ordering problem by using a very time-consuming iterative algorithm.

The parcel/layer disconnectedness along survey lines (or boreholes) similarly results from the simulation ordering, although using line-data for conditioning may be also part of the reason. The steep side of a parcel is always the left side of the parcel if the simulation ordering is from top-left to bottom-right, and vice versa. This problem is strong in the CMC_0 (Li, 1999), but mitigated by conditioning on line-data as "future" states in the CMC_x and CMC_{xy}. However, when survey lines (or boreholes) are sparse, this artifact still occurs.

As to the under-prediction of minor states (correspondingly, the over-prediction of major states), we think the main reason may be the full independency assumption of the two one-dimensional Markov chains in the *x* and *y* directions that was used to derive the model. Similar problem also occurs in the MRF model of Norberg *et al*. (2002) but the reason is not sure. They thought that multi-dimensional Markov chains themselves may have the tendency of over-estimating spatial dependencies of classes. If this is the case, the over-estimated spatial dependencies should be rather the auto-correlations than the cross-correlations of classes.

Although all these three constraints can lessen with increasing the conditioning data, the required data density for a satisfied simulation is usually not realistic for many soil survey datasets. The under-estimation of minor/small states is a big problem unless users' attention is the major states. The prediction artifacts, particularly the parcel/layer inclination, will also disqualify the simulated data when surveyed line-data for conditioning a simulation is not adequate. Therefore, resolving or further lessening these constraints in the CMC method will be the key tasks to raise the model to a widely applicable level for characterizing categorical soil variables.

Additionally, from a large number of simulations we did, we found that the CMC_x produces worse results than the CMC_{xy} for simulating soil classes even with dense survey lines (Of course, it is also not necessary since we can use the CMC_{xy}); this is easy to understand because there is no other data in the *y*-direction being conditioned except for one necessary boundary. But the CMC_x model aims to work for simulating soil transects with a number of densely distributed boreholes. The obvious fact is that soil layers have strong extension in the lateral direction (i.e., a large ratio of layer length to width) but soil class parcels on ground surface usually have no such a characteristic. Elfeki & Dekking (2001) also demonstrated that major lithofacies layers can be captured when the layers are very thin and long (almost across the whole transect) with a number of well data. The reason may be that the strong extension of subsurface layers in the lateral direction results in a strong influence of borehole data (i.e., the strong

auto-correlation in the lateral direction, which reflects on the TPM in the lateral direction with very large diagonal probability values p_{ii}).

Conclusions

The CMC method is evaluated for their possible application to characterize spatial heterogeneity of soil classes and layers. Results show that with a set of densely measured (or surveyed) line-data for conditioning simulations the complex spatial patterns of categorical soil variables – surface soil types (using CMC_{xy}) and subsurface soil textural layers (using CMC_x) can be mimicked with abrupt boundaries and approximate anisotropies. However, when the density of line-data becomes relatively low, the simulation effectiveness decreases and realizations become unrealistic. Three obvious constraints can be identified from simulated results when the conditioning line-data are relatively sparse. These constraints include the parcel/layer inclination, the under-prediction/estimation of minor/small states, and the parcel/layer disconnectedness in simulated realizations. The possible reasons for these shortcomings are discussed. The CMC method is efficient because of the non-iterative simulation; in this study, generating 100 realizations of the soil type map needs only about 15 to 20 minutes on a PC, and only several minutes are needed for producing 100 realizations of the alluvial soil transect.

Because the required density of survey line or bore data for producing reasonable realizations, particularly satisfactory spatial patterns, of categorical soil variables is normally quite high, and the datasets in real world applications usually may not reach to such a density, this method at its current stage still has difficulties to meet our needs for predictive soil mapping. If random sample point data were used, the required sample density for producing reasonable realization maps would have been much higher. Therefore, further efforts are needed to solve its drawbacks and eliminate its prediction artifacts.

References

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

Bierkens, M.F.P. & Burrough, P.A. 1993a. The indicator approach to categorical soil data: I. Theory. *Journal of Soil Science*, **44**, 361-368.

Bierkens, M.F.P. & Burrough, P.A. 1993b. The indicator approach to categorical soil data: II. Application to mapping and land use suitability analysis *Journal of Soil Science*, **44**, 369-381.

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

Bouma, J., van Alphen, B.J. & Stoorvogel, J.J. 2002. Fine tuning water quality regulations in agriculture to soil differences. *Environmental Science and Policy*, **5**, 113-120.

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

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

Carle, S.F. & Fogg, G.E. 1996. Transition probability-based indicator geostatistics. *Mathematical Geology*, **28**, 453-477.

Deutsch, C. V. & Journel, A. G. 1998. *GSLIB: Geostatistics Software Library and user's guide*. Oxford University Press, New York.

Ehlschlaeger, C. R. 1998. *The stochastic simulation approach: Tools for representing spatial application uncertainty*. PhD dissertation, University of California, Santa Barbara (http://faculty.wiu.edu/CR-Ehlschlaeger2/older/dissertation/fullDissertation.html).

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

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

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

Feyen, J., Jacques, D., Timmerman, A. & Vanderborght, J. 1998. Modeling water flow and solute transport in heterogeneous soils: A review of recent approaches. *Journal of Agricultural Engineering Research*, **70**, 231-256.

Gomez-Hernandez, J.J. & Srivastava, R.M. 1990. ISIM3D: An ANSI-C three-dimensional multiple indicator conditional simulation program. *Computer & Geosciences*, **16**, 395-440.

Goovaerts, P. 1997. *Geostatistics for natural resources evaluation*. Oxford University Press, New York.

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

Heuvelink G.B.M. & Webster, R. 2001. Modeling soil variation: past, present, and future. *Geoderma*, **100**, 269-301.

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

Kite, G.W. & Kauwen, N. 1992. Watershed modeling using land classification. *Water Resource Research*, **28**, 3193-3200.

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

Lark, R.M. & Webster, R. 1999. Analysis and elucidation of soil variation using wavelets. *European Journal of Soil Science*, **50**, 185-206.

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

Li, W., Li, B. & Shi, Y. 1999. Markov-chain simulation of soil textural profiles. *Geoderma*, **92**, 37-53.

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

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

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

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

Mowrer, H.T. & Congalton, R.G. (Eds). 2000. *Quantifying spatial uncertainty in natural resources: theory and applications for GIS and remote sensing*. Ann Arbor Press, Chelsea, Michigan.

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

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

Scull, P., Franklin, J., Chadwick, O.A. & McArthur, D. 2003. Predictive soil mapping: a review. *Progress in Physical Geography*, **27**, 171-197.

Todorovic, P. & Woolhiser, D.A. 1975. A stochastic model of n-day precipitation. *Journal of Applied Meteorology*, **14**, 17-24.

Weissmann, G. S., Carle, S. F. & Fogg, G. E. 1999. Three-dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics. *Water Resources Research*, **35**, 1761-1770.

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

Wingle, W.L. & Poeter, E.P. 1993. Uncertainty associated with semivarograms used for site simulation. *Ground Water*, **31**, 725-734.

Zhu, A.X. & Mackay, D.S. 2001. Effects of spatial detail of soil information on watershed modeling. *Journal of Hydrology*, **248**, 54-77.

Zhu, A.X, Hudson, B., Burt, J., Lubich, K. & Simonson, D. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, **65**, 1463-1472.

Zhang, J. & Goodchild, M. 2002. *Uncertainty in geographical information*. Taylor & Francis, New York.