# Two-dimensional Markov Chain Simulation of Soil Type Spatial Distribution

Weidong Li,* Chuanrong Zhang, James E. Burt, A.-Xing Zhu, and Jan Feyen

## ABSTRACT

**Soils typically exhibit complex spatial variation of multi-categorical variables such as soil types and soil textural classes. Quantifying and assessing soil spatial variation is necessary for land management and environmental research, especially for accurately assessing the water and solute transport processes in watershed scales. This study describes an efficient Markov chain model for two-dimensional modeling and simulation of spatial distribution of soil types (or classes). The model is tested through simulations of a simplified soil map. The application of the model for predictive soil mapping with parameters estimated from survey lines is explored. Analyses of both simulated maps and associated semi-variograms show that the model can effectively reproduce observed spatial patterns of soil types and their spatial autocorrelation given an adequate number of survey lines. This indicates that the model is a feasible method for modeling spatial distributions of soil types (or classes) and the transition probability matrices of soil types in different directions can adequately capture the spatial interdependency relationship of soil types. The model is highly efficient in terms of computer time and storage. The model also provides an approach for assessing the uncertainty of soil type spatial distribution in areas where detailed survey data are lacking. The major constraint on applications of the model at this stage is that the minor soil types are relatively underestimated when survey lines are too sparse.**

Complex spatial variation of multi-categorical soil variables, such as soil types and soil textural classes, is a typical feature of soils in the real world. On the one hand, traditionally the information on the spatial distribution of soil types can only be obtained by detailed field surveys, and soil maps are drawn according to experts' empirical judgment based on visual field observations and visual interpretation of air photos and topographic maps. For some regions with limited physical access, or without enough survey data, the soil distribution is difficult to assess. On the other hand, we are still short of suitable mathematical methods to quantitatively characterize the spatial distribution of categorical soil variables such as soil types and various soil classes. Because an understanding of the spatial distribution of categorical soil variables is crucial to soil management and environmental research (Kite and Kauwen, 1992; Zhu and MacKay, 2001; Bouma et al., 2002), it is essential to develop suitable mathematical models for characterization of the spatial distribution of such variables. At present, such research is still very rare in soil science literature.

For characterizing the spatial correlation of categorical variables in geosciences, the main descriptive tools currently used are *indicator variograms* provided by indicator geostatistics and *transition probability matrices* (TPMs) provided by Markov chains. Currently, indicator geostatistics (Journel, 1983), especially the sequential indicator simulation (Deutsch and Journel, 1997), are more widely used. Indicator methods usually deal with multiple classes by considering each class binomially and using indicator variogams class-by-class to represent spatial correlation. This approach has been proven suitable for modeling cutoffs (i.e., thresholds) of continuous variables (Goovaerts, 1997, 1999; Brus et al., 2002). But for categorical variables that are normally classified into multinomial classes with complex spatial dependences, indicator geostatistics seem insufficient to capture the complex spatial patterns of multinomial classes with limited measured data (Bierkens and Weerts, 1994; Ehlschlaeger, 2000; Weissmann and Fogg, 1999; McGwire and Fisher, 2001). For example, indicator geostatistics have difficulties in dealing with sharp boundaries and autocorrelation of nominal classes simultaneously (Mowrer and Congalton, 2000; McBratney et al., 2000), coping with anisotropies in multinomial classes (Wingle and Poeter, 1993; Ehlschlaeger, 1998), respecting the juxtaposition relationships between classes (Weissmann and Fogg, 1999), and integrating of expert knowledge (Carle and Fogg, 1996; Scull et al., 2003; Weissmann and Fogg, 1999). They are also highly demanding in computation when the number of classes is large (Zhang and Goodchild, 2002), which hinders application over large areas and in high-resolution simulation.

The Markov chain theory is a stochastic process theory, which describes how likely one state is to change to another state through one or more time or space steps. The one-dimensional Markov-chain method has been widely used in geology to simulate stratigraphic sequences since 1960s (Harbaugh and Bonham-Carter, 1980; Krumbein, 1968). It also has been used in soil science to describe the spatial order of parcels of different soil classes (Burgess and Webster, 1984a, 1984b) and the vertical spatial change of textural layers in alluvial soils (Li et al., 1997, 1999) in one-dimension. Although one-dimensional Markov chain is simple and easy to use, extending it into multidimensions for conditional simulation is difficult because of the difficulties of conditioning on measured data and choosing a suitable simulation ordering.

"Unlike one-dimensional application of Markov chains, two- and three-dimensional applications are difficult because there is not

W. Li and J.E. Burt, Dep. of Geography, Univ. of Wisconsin, Madison, WI 53706; C. Zhang, Dep. of Geography and Geology, Univ. of Wisconsin, Whitewater, WI 53190; A.-Xing Zhu, State Key Lab. of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science, Beijing 100101, China; J. Feyen, Institute for Land and Water Management, Catholic Univ. of Leuven, B-3000 Leuven, Belgium. Received 7 Apr. 2003. *Corresponding author (li2@wisc.edu).

Published in Soil Sci. Soc. Am. J. 68:1479–1490 (2004).
© Soil Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

**Abbreviations:** CMC, coupled Markov chain; TMC, triplex Markov chain; TPM, transition probability matrix.

> an easily identifiable ordering of values in a past-present-future sequence (A.G. Journel, personal communication 1995)."
>
> —Koltermann and Gorelick (1996, p. 2631)

But recently, there have been attempts to formulate new models for the use of Markov chains in two- to three-dimensional lithologic characterization in geology. One approach is to integrate transition probabilities of Markov chains into the framework of indicator geostatistics for lithofacies simulation in three-dimensions (Carle and Fogg, 1996, 1997). Thus, compared with conventional indicator geoststistics, this approach can better represent the spatial cross-correlations between different states, such as the state sequence asymmetry, which are especially prominent in the vertical direction of lithofacies. The second approach implements the Markov random field theory of Besag (1974). Applications of this approach have mainly appeared in other fields such as image restoration. Its application in geosciences is limited by deficiencies such as extremely high demand in computation and underestimation of infrequent states when simulation is conditioned on sparse data (Norberg et al., 2002). A third approach directly couples two one-dimensional Markov chains and uses the joint probability distribution to perform two-dimensional simulation (Elfeki, 1996; Elfeki and Dekking, 2001). This method also cannot meet the requirements for predictive mapping of categorical soil variables for reasons discussed below. In general, currently it seems the methods suitable for two-dimensional modeling of soil type (or class) spatial distribution from sparse survey data are rare because of the complex patterns of multinomial soil classes and the difficulties in dealing with anisotropies, connectedness and boundaries of soil class parcels.

The coupled Markov chain (CMC) model developed by Elfeki (1996) to characterize the heterogeneity of geological formations was quite simple. Two one-dimensional Markov chains in the *x*-direction and the *y*-direction, respectively, which were assumed independent of each other, were coupled together. Simulation was done from one corner to the other diagonal corner. But Li's (1999) application of the method to simulation of soil type distribution and alluvial soil textural profiles showed that the method was not practical for categorical soil variables. Three problems were found in his simulation: (i) the simulated soil parcels or layers are always inclined in the direction of simulation; (ii) if some component accounts for a small proportion of the area (occurs sparsely relative to others) it will be seriously underrepresented or even disappear in simulated realizations; (iii) when a new layer (or parcel) appears in a realization it occurs abruptly along survey lines and shows disconnectedness of component layer or parcels. Elfeki and Dekking (2001) recently extended the method to simulate geologic sections by conditioning on future states (well data), which largely improved the practicality of the method for characterization of subsurface lithofacies. They showed that the major layers of lithofacies could be captured in simulated realizations when a number of wells were conditioned. But the aforementioned problems still exist to some extent when the density of boreholes is not high enough. One of the main reasons for these constraints is the asymmetric property of the model. However, the CMC does have its outstanding merits such as high efficiency and explicitness.

Recognizing the merits and remaining problems of the CMC model, we propose a triplex Markov chain (TMC) approach to mitigate the shortcomings of the CMC model. The TMC approach employs two related CMCs, which proceed alternately in a simulation domain. It extends the CMC model's ability to condition on future states to permit conditioning on neighboring survey line data in four directions. Thus, the aforementioned Problems (i) and (iii) existing in the CMC model are overcome and the Problem (ii) is mitigated. (As with the CMC model, the Problem (ii) disappears if a suitable amount of measured conditioning data is used.) The TMC retains the advantages of the CMC model, such as explicitness and high efficiency. While other spatial stochastic models mainly condition simulations to data at dispersed single sampling points, the TMC model conditions simulations to survey lines (outer boundaries and internal lines). Given the prominence of transect surveys, this may be more realistic for real-world applications of the model to categorical soil variables.

This study applies the TMC approach to characterizing spatial variability in soil types. The purpose is to introduce a fast simple and effective method for predictive mapping of categorical soil variables, and for obtaining uncertainty estimates of predicted values. Different simulation schemes are used in our simulations to test the feasibility of the TMC approach in modeling soil type spatial heterogeneity and to explore its deficiencies.

## MATERIALS AND METHODS

### The Triplex Markov Chain Model

For the details of the CMC model, see Elfeki and Dekking (2001). The following section introduces the TMC model. The TMC methodology uses two CMCs constructed from three independent one-dimensional first-order Markov chains. Consider three one-dimensional stationary Markov chains $(X_i^{\mathrm{L}})$, $(X_i^{\mathrm{R}})$, and $(Y_i)$ all defined on the state space [1, 2, ..., *n*]. The symbol $(Y_i)$ represents a *y*-direction chain. The $(X_i^{\mathrm{L}})$ and $(X_i^{\mathrm{R}})$ represent an *x*-direction chain (i.e., from left to right) and an $x'$-direction (i.e., anti-*x*-direction, from right to left) chain, respectively (Fig. 1). Then the $(Y_i)$ and $(X_i^{\mathrm{L}})$ are coupled together to form one CMC $(Z_{ij}^{\mathrm{L}})$ from left to right, and at the same time $(Y_i)$ and $(X_i^{\mathrm{R}})$ are coupled together for the other CMC $(Z_{ij}^{\mathrm{R}})$ from right to left. The two CMCs proceed on a two-dimensional domain at opposite directions alternately. This defines the TMC used in this study.

Consider the two-dimensional domain of cells as shown in Fig. 1. This domain is partitioned by survey lines (including outer boundaries) into many small "windows." The four boundaries of each window are known as parts of survey lines. In each window, each cell has a column number *i* and a row number *j*. Also consider a given number of soil types or classes (1, 2, …, *n*) occurring in the domain (These soil types are coded as, for example, 1 for Type 1, 2 for Type 2, and 3 for Type 3, etc., in data). The first-order Markov chain $(Y_i)$ describes the spatial change process of soil types in this two-

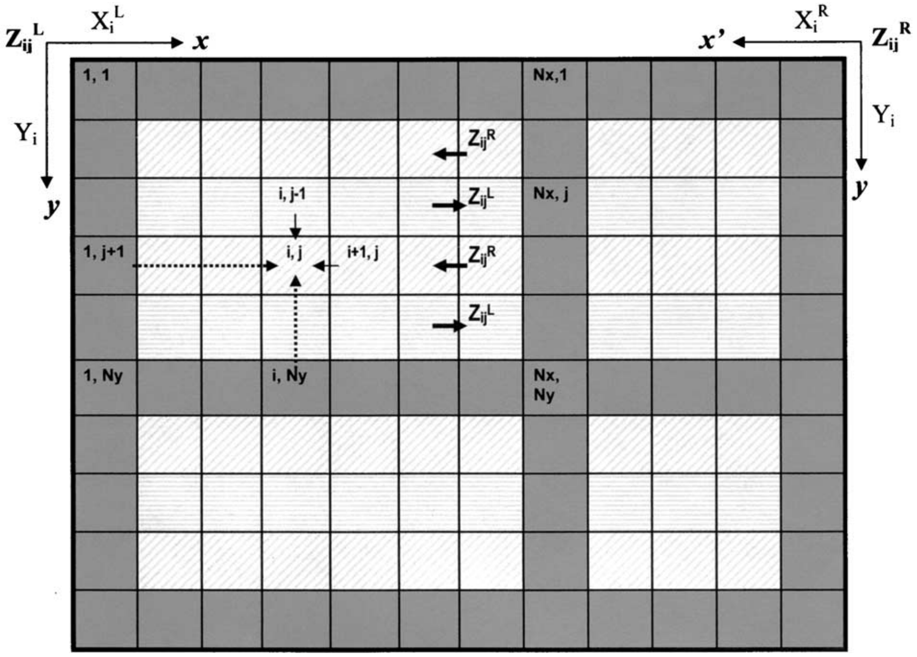

**Fig. 1. A triplex Markov chain is applied to each window (light gray cells) of a two-dimensional domain. Simulation is conditioned on window boundaries, that is, survey lines (dark gray cells).**

dimensional domain in the $y$-direction, and the $(X_i^{\mathrm{L}})$ and $(X_i^{\mathrm{R}})$ describe the spatial change processes of soil types in the $x$-direction and the $x'$-direction.

The transition probability of the $(X_i^{\mathrm{L}})$ chain with conditioning to future states can be expressed as

$$p^L_{lk|q} = p(X_i = k|X_{i-1} = l, X_{N_x} = q) = \frac{p^L_{lk}\, p^{L(N_x-i)}_{kq}}{p^{L(N_x-i+1)}_{lq}} \quad [1]$$

where $p^L_{lk}$ is a one-step transition probability from state $l$ to state $k$ in the $x$-direction, $p^{L(N_x-i)}_{kq}$ is a $(N_x - i)$-step transition probability, $p^{L(N_x-i+1)}_{lq}$ is a $(N_x - i + 1)$-step transition probability, and $p_{lk|q}$ is our target, the probability of Cell $i$ to be in State $k$, given that the previous cell $i - 1$ is in State $l$ and the future cell $N_x$ is in State $q$. When Cell $N_x$ is far from Cell $i$ the terms $p^{L(N_x-i)}_{kq}$ and $p^{L(N_x-i+1)}_{lq}$ have little influence because they both will be almost equal to the same stationary probability. 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. Similarly the transition probability of the $(Y_i)$ chain with conditioning to future states is expressed as

$$p^Y_{mk|o} = p(Y_j = k|Y_{j-1} = m, Y_{N_y} = o) = \frac{p^Y_{mk}\, p^{Y(N_y-j)}_{ko}}{p^{Y(N_y-j+1)}_{mo}} \quad [2]$$

By forcing the two one-dimensional chains $(Y_i)$ and $(X_i^L)$ to move to the same states, say $k$, in Cell $(i, j)$ (Fig. 1), for the CMC $(Z^L_{ij})$, we have its conditional joint transition probability

$$P^L_{lm,k|qo} = Cp^Y_{mk|o}p^L_{lk|q} = \frac{p^L_{lk}\, p^{L(N_x-i)}_{kq}\, p^Y_{mk}\, p^{Y(N_y-j)}_{ko}}{\sum_{f=1}^{n} (p^L_{lf}\, p^{L(N_x-i)}_{fq}\, p^Y_{mf}\, p^{Y(N_y-j)}_{fo})} \quad [3]$$

where $k = 1, \ldots, n$. $C$ is a normalizing constant, which arises because we only consider the transition from State $l$ at $\mathrm{Z}_{i-1,\ j}$ and State $m$ at $\mathrm{Z}_{i,\ j-1}$ to the same State $k$ at $\mathrm{Z}_{i,j}$. Here, the $C$ is expressed as

$$C = \left(\sum_{f=1}^{n} p^L_{lf|q}\, p^Y_{mf|o}\right)^{-1} \quad [4]$$

The process $(Z^R_{ij})$ obeys a similar rule except for an opposite proceeding direction. Similarly, we can get $p^R_{lm,k|qo}$ for $(Z^R_{ij})$. Thus, the conditional joint probability pair $(p^L_{lm,k|qo}, p^R_{lm,k|qo})$ represents the TMC model.

In such a two-dimensional domain, the two CMC processes cannot proceed in the same row. Rather, they proceed alternately in different rows of each window (Fig. 1). Each process will condition on the states in the upper row produced by the other process (except for the top boundary), the preceding state produced by the same process, and known future states (i.e., boundaries of each window) in the $x$- or $x'$-direction and the $y$-direction. In the second line of each window the process will condition on the upper boundary. Thus, a simulation can be done window by window and in each window row by row, and all the survey line data in the domain are used for conditioning.

## Inference of Statistical Parameters

A Markov chain is completely described when the state space, TPM and initial probabilities are given. For a soil system represented by a Markov chain, one has to first define the set of possible sates (i.e., types) of the system, [1, 2,…, $n$], and the transition probability, $p_{lk}$, for transitions from State $l$ to State $k$ in one step. The state space can be determined according to the actual need. For example, for hydrologic modeling soils might be grouped into just a few classes, such as sand, loam, and clay, etc. The transition probabilities can be determined by superimposing a lattice on the soil map and counting the state changes in different directions. The cell size (square or rectangle in this study for simplicity) should not be larger than the smallest parcel size to be shown in simulated realizations and must be the same for both parameter estimation and simulation. In practical applications, where exhaustive data about a study area are not available, transition probabilities can be directly estimated from survey lines. When survey lines are very sparse, insufficient for parameter estimation, soft information such as existing paper or digital maps (maybe hand delineated and low quality), information derived from analogous areas, and expert knowledge may be incorporated

(Rosen and Gustafson, 1996; Weissmann and Fogg, 1999; Elfeki and Dekking, 2001).

The transition frequency matrices between the soil types in the $x$-, $x'$-, or $y$-direction can be calculated by counting the times of a given type (e.g., $l$) followed by itself or the other type (e.g., $k$) in the direction on the lattice, and then the one-step TPMs (for one-dimensional first-order Markov chain) can be obtained by dividing the transition frequencies with the row total number of the transition frequency matrices as below:

$$p_{lk} = T_{lk} / \sum_{k=1}^{n} T_{lk} \qquad [5]$$

where, $T_{lk}$ is the transition frequency from State $l$ to State $k$ in the $x$-, $x'$- or $y$-direction on the lattice. When we obtain the TPM from the transition frequency matrix in the $x$-direction, the TPM in the $x'$-direction also can be obtained because the transition frequency matrix in the $x$-direction is transpositive of the transition frequency matrix in the $x'$-direction, that is,

$$T^{x}_{lk} = T^{x'}_{kl} \qquad [6]$$

Multistep transition probabilities for a one-dimensional chain can be calculated by self-multiplication of the one-step transition probability matrix. Conditional joint transition probabilities for a conditioned CMC can be further calculated based on Eq. [3].

## Simulation Procedure

Monte-Carlo simulations were used to generate multiple realizations using the above TMC model. The complete procedure consists of the following steps:

Step 1: Discretize the area to be simulated using a grid (Fig. 1).
Step 2: Insert survey line data in boundary cells and internal cells for conditioning the simulation.
Step 3: Generate the unknown cells in each window bounded by survey lines with numbers $(i, j)$, $i = 2,..., N_x - 1$ and $j = 2,..., N_y - 1$ row by row using the conditional joint probability distribution $p^{L}_{lm,k|qo}$ from left to right and $p^{R}_{lm,k|qo}$ from right to left, alternately for different rows.

Step 4: Repeat the procedure until all windows are fully filled.

Step 5: Generate another realization using Steps 3 and 4.

## Simulation Examples

Our simulations were based on a section of a digital soil map of a river basin in Belgium, with a length of 8 km and a width of 1.7 km (Fig. 2). More than 40 soil types were shown on the original map. For simplicity's sake we merged similar soil types to give a total soil of seven types. This is warranted whenever two or more soil classes have nearly identical values of whatever property is of interest (e.g., hydraulic conductivity). On a more practical level, merging will sometimes be necessary to ensure a manageable number of categories in the transition probability matrices. Here the small number of classes was chosen mainly for clarity of presentation—numerical constraints would admit a much larger number of classes. The specific soil types are themselves of no particular interest in this study. They are just used here to show that spatial heterogeneity of soil types or classes can be characterized using the TMC model. The soil map is discretized into a 160 × 34 grid with a cell size of 50 m (Fig. 2). We will simulate the whole map—a larger map and its left half—a smaller map.

## Simulation Schemes

Different schemes will be used in our simulations to display the feasibility and the practicality of the TMC model. To test the model, we will use the TPMs directly estimated from exhaustive data—the original maps to simulate the same areas represented by the maps. To show the practical use of the model, we will use the TPMs estimated from only the survey lines. Different densities of survey lines (i.e., different survey line intervals) will be used. Survey lines are distributed with approximately equal intervals (about 1000, 500, or 250 m) in our simulations.

## Input Parameters and Output Results

Input parameters for a simulation include TPMs in the $x$, $x'$ and $y$ directions, the numbers of grid columns and grid rows (or cell size, length, and width) of the discretized simulation area, and the number of soil types, plus transects used for conditioning. Output results include realizations, occurrence probability maps of each soil types, the soil map estimated from maximum occurrence probabilities (i.e., the so-called optimal interpolation map), and related statistics. For each simulation under a conditioning scheme, we will generate 100 realizations but only display the first realization. To show the simulated results in multiple realizations and how likely a soil type occurs on each location (i.e., grid cell), we use occurrence probability maps. Occurrence probability maps are calculated as follows: When a soil type occurs at a location in one realization, it is counted, otherwise not. By dividing the occurrence number of a soil type at a location by the number of realizations, we can get the occurrence probability of the soil type at the location. Thus, we can get an occurrence probability map for each soil type. The provided probability maps are calculated from 100 realizations.

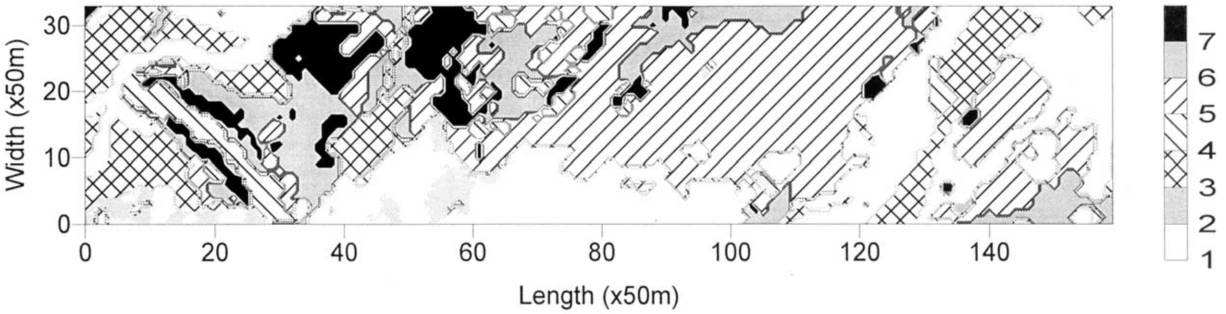

**Fig. 2. A simplified soil map with seven soil types. This map is discretized into a 160 × 34 grid with a cell size of 50 m. Note: The length should be multiplied by 50 to obtain the correct length values.**

# RESULTS AND DISCUSSION

## Model Testing

We first simulate the larger soil map using TPMs directly estimated from the map. This guarantees that the TPMs are correct, which means that deviations between observed and simulated patterns can be ascribed to failings of the model and/or inadequate conditioning data. Figure 3 shows simulated results using three different survey line intervals. Input parameters estimated from the original map of Fig. 2 are given in Table 1. All three realizations show resemblance to the original map's spatial patterns of soil types, despite the very different patterns in the left and right parts. But it can be seen that the major (frequently occurring) Soil Type 5 is overestimated in the realization with sparser survey lines (i.e., a survey line interval of 1000 m); consequently, some infrequently occurring soil types, such as Soil Type 6, are underrepresented. The occurrence probability maps of Soil Type 6 under different conditioning schemes show that it is underestimated when survey lines are too sparse. This can be seen more clearly from the statistical proportions of different soil types in Table 2. The representation problem is quickly mitigated when the survey line interval decreases to about 500 and 250 m. The simulated results are very similar to the original map under the survey line interval of 250 m.

**Table 1. Input parameters (one-step transition probability matrices [TPMs] and grid information) estimated from the original soil map in Fig. 2.†**

| Soil type | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| | | | **TPM in the $x$-direction** | | | | |
| 1 | .838 | .030 | .055 | .010 | .046 | .012 | .009 |
| 2 | .202 | .614 | .019 | .013 | .133 | .006 | .013 |
| 3 | .098 | .004 | .804 | .009 | .022 | .052 | .011 |
| 4 | .033 | .000 | .041 | .633 | .131 | .077 | .086 |
| 5 | .050 | .010 | .002 | .014 | .849 | .049 | .026 |
| 6 | .029 | .000 | .051 | .025 | .083 | .732 | .080 |
| 7 | .040 | .000 | .004 | .055 | .139 | .090 | .672 |
| | | | **TPM in the $x'$-direction** | | | | |
| 1 | .828 | .023 | .053 | .006 | .061 | .015 | .013 |
| 2 | .259 | .614 | .019 | .000 | .108 | .000 | .000 |
| 3 | .101 | .004 | .823 | .014 | .005 | .050 | .003 |
| 4 | .057 | .008 | .029 | .633 | .098 | .073 | .102 |
| 5 | .037 | .012 | .010 | .019 | .849 | .035 | .037 |
| 6 | .022 | .001 | .053 | .026 | .114 | .727 | .056 |
| 7 | .029 | .004 | .018 | .047 | .096 | .129 | .678 |
| | | | **TPM in the $y$-direction** | | | | |
| 1 | .838 | .023 | .068 | .003 | .048 | .012 | .008 |
| 2 | .173 | .669 | .000 | .000 | .135 | .015 | .008 |
| 3 | .103 | .012 | .801 | .012 | .013 | .050 | .008 |
| 4 | .041 | .012 | .016 | .591 | .114 | .102 | .122 |
| 5 | .067 | .013 | .003 | .014 | .835 | .042 | .027 |
| 6 | .028 | .003 | .050 | .048 | .107 | .708 | .056 |
| 7 | .031 | .007 | .007 | .051 | .113 | .109 | .682 |

† States: 7; grid columns: 160; grid rows: 34.

The TMC is also quite efficient, as judged by computer execution time. The simulation time used by an ordinary personal computer for generating 100 realiza-

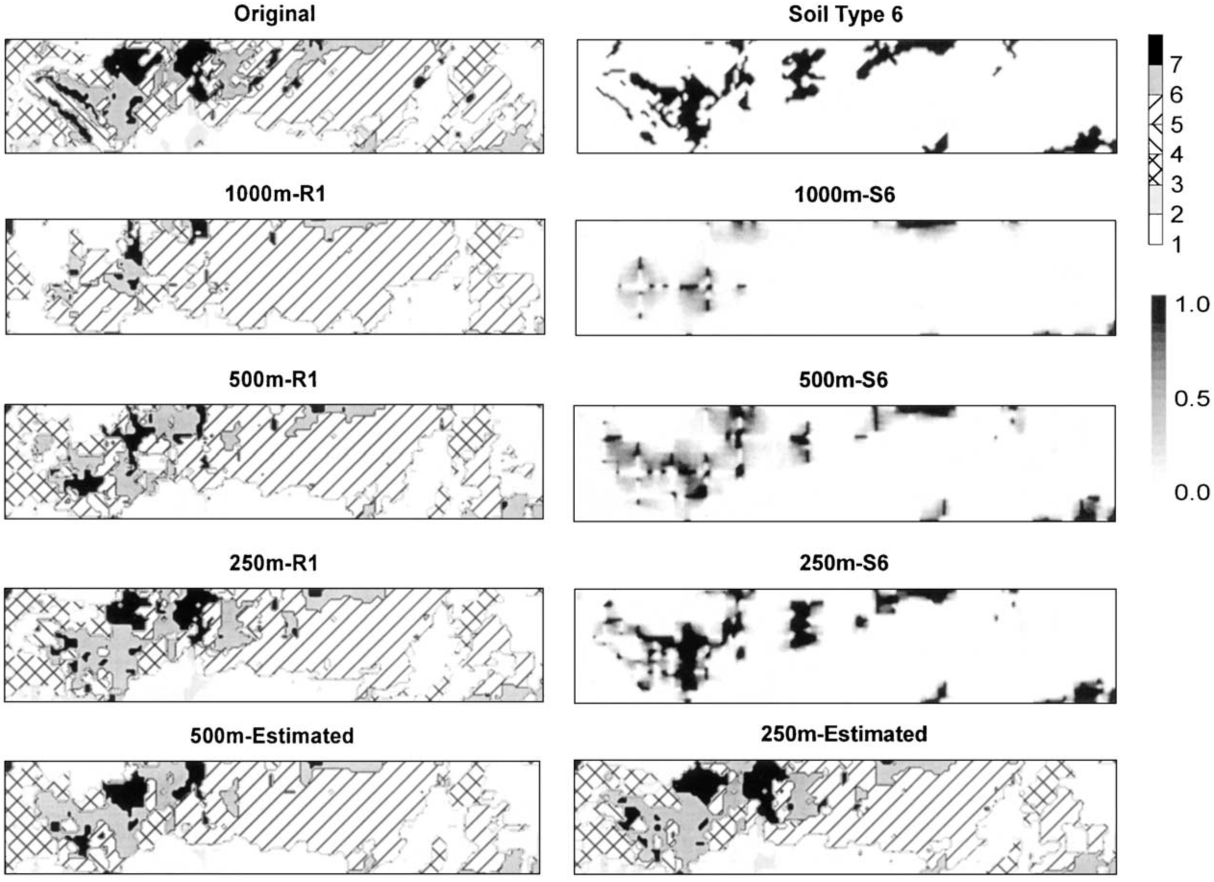

**Fig. 3. Simulated results of the soil type distribution in the study area of Fig. 2 under different conditioning schemes. Labels 1000m, 500m, and 250m represent conditioning schemes used, that is, survey line intervals. R1 means the first simulated realization based on the corresponding survey line interval. S6 means Soil Type 6. The bottom row gives the estimated soil map based on maximum occurrence probabilities.**

**Table 2. Proportions of different soil types in the original soil map (the whole map) and averaged from 100 simulated realizations for each simulation scheme (corresponds to Fig. 3).**

| Survey lines | | Proportions of different soil types | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Columns × rows | Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Run time† |
| | m | | | | | | | | min |
| | | | | Original | | | | | |
| — | — | .2557 | .0296 | .1385 | .0452 | .3143 | .1335 | .0832 | — |
| | | | | Simulated‡ | | | | | |
| 9 × 3 | 1000 | .3272 | .0081 | .0969 | .0084 | .4678 | .0645 | .0272 | 18 |
| 17 × 5 | 500 | .2880 | .0128 | .1341 | .0176 | .3748 | .1223 | .0505 | 14 |
| 33 × 8 | 250 | .2701 | .0207 | .1397 | .0271 | .3278 | .1410 | .0736 | 10 |

† Time required for generating 100 realizations on an EPoX personal computer (Processor: AMD Athlon [tm] XP 2600; CPU: 1.916 GHZ; Memory: 1.048 GB. Memory frequency: 200 MHZ; 128.0 MB RAM). Computer program is written in Fortran 90.

‡ Simulated results are averaged values from 100 realizations, same in the following other tables.

tions of the larger map is given in Table 2. Simulation time never exceeded 20 min, with execution time decreasing as the density of survey lines increases (i.e., simulation windows become smaller).

Clearly, a single realization only gives one of an infinite number of spatial patterns that might occur under a given configuration of survey lines and parameters. The occurrence probability map of a soil type reflects frequency of occurrence over many realizations, and thus provides a measure of uncertainty for any realization. From the occurrence probability maps of Soil Type 6 in Fig. 3, it can be seen that the occurrence of a soil type gradually becomes more certain with the increase of the density of survey lines. The estimated maps based on the maximum occurrence probabilities of soil types at each cell represent the optimal spatial interpolation result of this model (Fig. 3, bottom row), which imitate the original map very well. Since soil types have different spatial distributional patterns and some soil types occur very sparsely in the simulation area, it is difficult to completely represent all soil types and the fine features of soil distributions unless abundant conditioning data are available.

Looking at the full study area, it is seen that the left and right halves exhibit very different patterns, with the left half having a considerably more complex pattern of soils. This suggests that a single set of TPMs might not be appropriate for the entire region. Accordingly, the left half was selected for another simulation, with results as shown in Fig. 4. It can be seen that the underestimation of minor types is clearly not so pronounced under the same density of survey lines. This occurs because the parameters are relatively more representative for the local simulation area (Table 3). Clearly, the occurrence of Soil Type 6 increases in simulated realizations under the survey line interval of 1000 m (see the probability map 1000m-S6 in Fig. 4) compared with the corresponding results in Fig. 2. But when survey lines are relatively dense, simulated results are similar with those in Fig. 3. Therefore, the resemblance between simulated realizations and the original map is not only decided by the parameters but also decided by the abundance of conditioning data (survey lines in this case). The results shown in Fig. 4 also indicate that for a large area with clearly different spatial patterns in different subareas it is better to use different TPMs for the different subareas.

To supplement these largely qualitative results, we have also computed indicator variograms for observed and simulated maps. Indicator variograms are widely used to characterize spatial autocorrelation and cross-autocorrelation in categorical variables (e.g., see Goovaerts, 1997). Here, working with seven soil types, we computed the full complement of (7 × 8)/2 = 28 variograms from the observed map in Fig. 4, which we take as "truth." These 28 variograms can be compared with variograms computed from any realization to assess that realization's agreement with the real-world distribution of variance and cross-variance as a function of spatial scale (varying lag). Figure 5 shows just a subset of 10 comparisons based on the simulated realization 500m-R1 in Fig. 4, where the survey line interval was about 500 m. In all graphs solid points represent the original map and simulation results are drawn as a solid line. In preparing Fig. 5, we selected all seven univariate variograms for the seven soil types, and only three of the 21 cross-variograms between the seven soil types. Looking first at the univariate variograms (Fig. 5a–g), we see that most soils have a distinct pattern of spatial autocorrelation, and in all cases the model does a reasonably good job of capturing the observed pattern. This is, of course, just one realization. Different realizations (not shown) have variograms that differ in their details, but they are similarly consistent with the observed variograms. The cross-variograms (Fig. 5h–j) were subjectively chosen from the 21 available according to various criteria. Figure 5h is a worst-case example; chosen because it reflects the poorest fit between observed and simulated results. Looking at Fig. 5h, we see that although there is good agreement in the sign (mostly negative), the simulated cross-variogram for Soil Types 1 and 6 departs from the observed by a factor of about one-third for lags larger than 20 or so (1000 m and larger). By contrast, Fig. 5i shows an extremely good fit at all lags, although there it has considerably less structure than Fig. 5h. The last graph (Fig. 5j) was selected because it has the most complex observed cross-variogram. We see that despite the highly variable relationship between covariance and lag, there is excellent agreement between observed and simulated curves. These results suggest that the TMC model

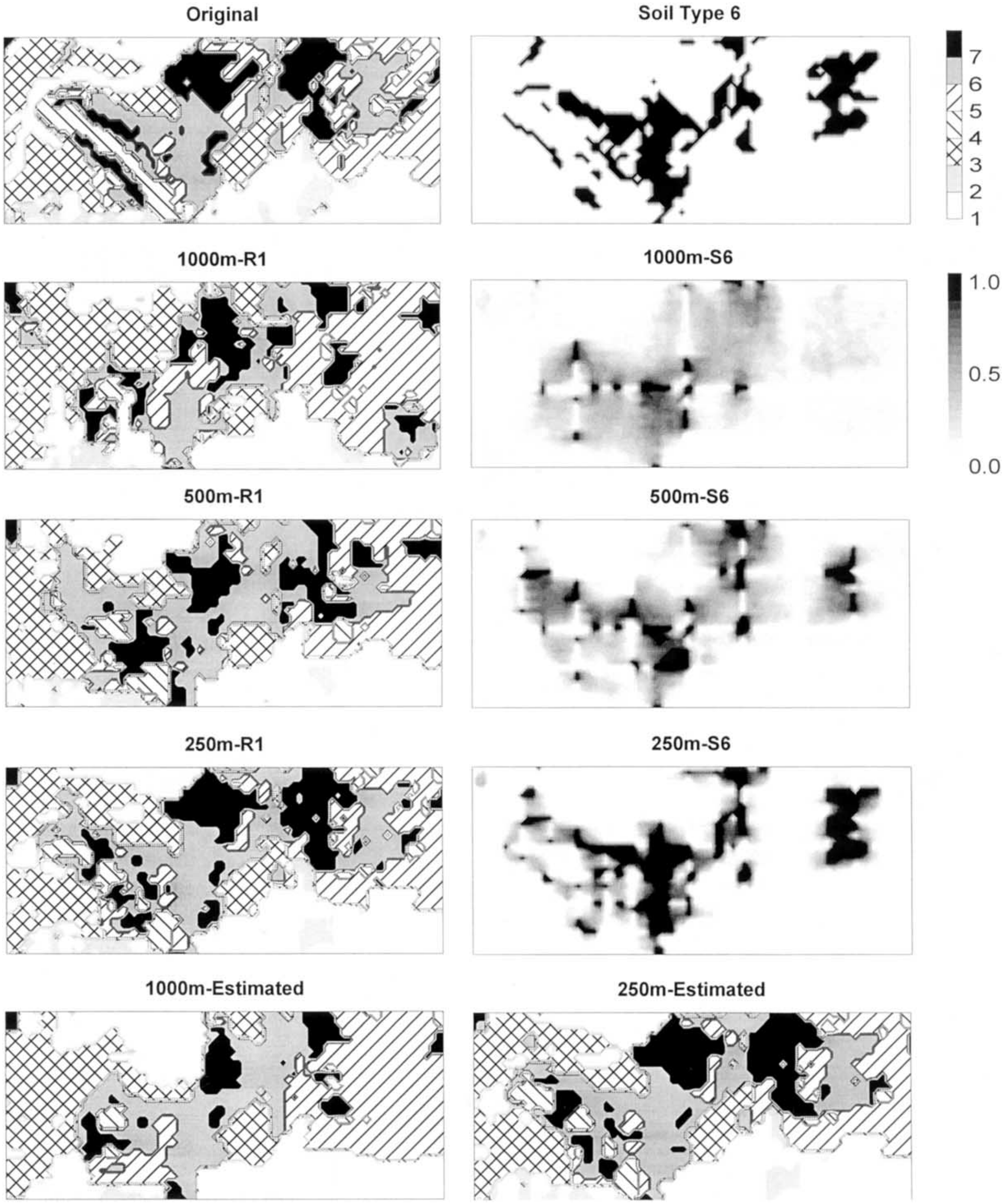

**Fig. 4.** **Simulated results of the soil type distribution in the left half of the study area under different conditioning schemes. Labels 1000m, 500m, and 250m represent conditioning schemes used, that is, survey line intervals. R1 means the first simulated realization based on the corresponding survey line interval. S6 means Soil Type 6. The bottom row gives the estimated soil map based on maximum occurrence probabilities.**

is at least potentially capable of representing complex patterns of autocorrelation and cross-correlation, when supplied adequate conditioning data.

In general, from the testing results in Fig. 3, 4, and 5 we can see that the complex spatial patterns of soil types with abrupt boundaries can be mimicked with a sufficient number of survey lines using the TMC model. The main constraint is the over (or under) representation problem of major (or minor) soil types which occurs obviously when survey lines are too sparse. Of course, if different soil types account for a similar proportion in the simulation area, there will be not such problem.

## Practical Use

The quality of simulated results not only depends on the sufficiency of survey data but also depend on the

**Table 3. Proportions of different soil types in the original soil map (the left half) and averaged from 100 simulated realizations for each simulation scheme (corresponds to Fig. 4).**

| Survey lines | | Proportions of different soil types | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Columns × rows | Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Run time |
| | m | | | | | | | | min |
| | | | | Original | | | | | |
| – | – | .2112 | .0522 | .1784 | .0764 | .1633 | .1733 | .1451 | – |
| | | | | Simulated | | | | | |
| 5 × 3 | 1000 | .2586 | .0225 | .1986 | .0243 | .2100 | .1670 | .1190 | 9 |
| 9 × 5 | 500 | .2285 | .0328 | .1951 | .0377 | .1815 | .1883 | .1362 | 7 |
| 17 × 8 | 250 | .2147 | .0442 | .1956 | .0484 | .1570 | .1888 | .1513 | 6 |

quality of parameters. With decreasing density of survey lines, the relative importance of parameter estimates grows as the importance of observations shrinks. The simulations above assumed perfect knowledge of parameters acquired from an exhaustive map database. In practical applications one will often not have even a low-quality map as a source of exhaustive data for the study area. In that case parameter estimates are most easily acquired from the survey lines themselves. This is suitable when the density of survey lines is sufficient to cover all soil types and possible transitions. But when survey lines are too sparse parameters obtained from the survey lines may not be reliable. For example, transitions between soils that happen not to appear in the sample will be assigned a transition probability of zero. In addition, a small sample implies few replicates of transitions, which in turn means that the probability estimates will have a large standard error. When dealing with the excessively sparse survey lines, soft information, particularly expert knowledge can in theory be used to adjust the parameters. This is an inherent advantage of Markov chains because a one-step TPM is relatively intuitive (Weissmann and Fogg, 1999). However, incorporating soft information in Markov chain modeling requires expert knowledge about the study area involving typical parcel size, shape, orientation, and juxtaposition (Rosen and Gustafson, 1996). Obviously, such expert knowledge will often be unavailable, but it nevertheless is a potential solution to the problem of inadequate data. Other than to mention it as a possible resource, we will not consider the use of soft information in this paper. Alternatively, parameters can be estimated from survey lines in a similar but larger area. Clearly, survey lines in that larger area will be longer, resulting in more transitions from and to low-frequency types. Thus although the density of lines might remain low, the parameter estimates can be more reliable.

Figure 6 displays simulated results of the larger soil map using the TPMs estimated from survey lines, the interval of which is shown in the labels of realizations. The proportions of soil types estimated from survey lines and those averaged from realizations are given in Table 4. It can be seen that the estimated proportions from survey lines deviate more or less from the actual proportions in the original map. But simulated results (Fig. 6 and Table 4) are still similar with those using the parameters estimated from the exhaustive original map provided in Fig. 3 and Table 2. These simulated results in Fig. 6 are quite satisfactory, except for the same under (or over) estimation problem of minor (or major) soil types, which is obvious when survey lines are relatively sparse.

Similarly as the model testing, we simulate the smaller soil map using parameters estimated from survey lines. The results are presented in Fig. 7 and Table 5. It can be seen that the results has no obvious difference compared with those using the parameters estimated from the exhaustive original map provided in Fig. 4 and Table 3. The soil spatial patterns are well represented in simulated realizations and estimated maps from maximum occurrence probabilities under survey line intervals of about 500 and 250 m.

But in Fig. 7 no simulated results under the sparsest scheme (i.e., the survey line interval of 1000 m) are given. In fact, the simulation under this scheme breaks off because the parameters estimated from survey lines under this scheme are insufficient. The reason is that many transition probabilities related with rarer soil types (e.g., Soil Type 2 and 4) in the TPMs estimated from the survey lines are zero; thus for some cells, no non-zero joint transition probabilities from their conditioning neighbors to themselves can be found to determine their states. Therefore, when a study area is small and survey lines are excessively sparse, directly using parameters estimated from few survey lines may not be feasible.

## CONCLUSIONS

We have introduced a new model—the TMC, which is based on the CMC model, for stochastic simulation of categorical soil variables. The model was illustrated and tested for its ability to simulate the spatial distribution of soil types. The model requires both a set of parameters, in the form of a transition probability matrix, and a set of survey line data used to condition the results. To isolate the role of conditioning data, we first performed simulations using known TPMs acquired from an exhaustive map survey. Testing shows that the TMC model can mimic the observed spatial patterns of soil types very well when supplied an adequate number of survey lines. But when survey lines are too sparse, minor soil types are obviously underestimated in realizations. Not surprisingly, the model performs better when spatial patterns are uniform across the study area, and less well when there are large differences from one subregion to another. For demonstrating the model's practicality in real world applications, where no map survey is available and the TPMs are unknown, the

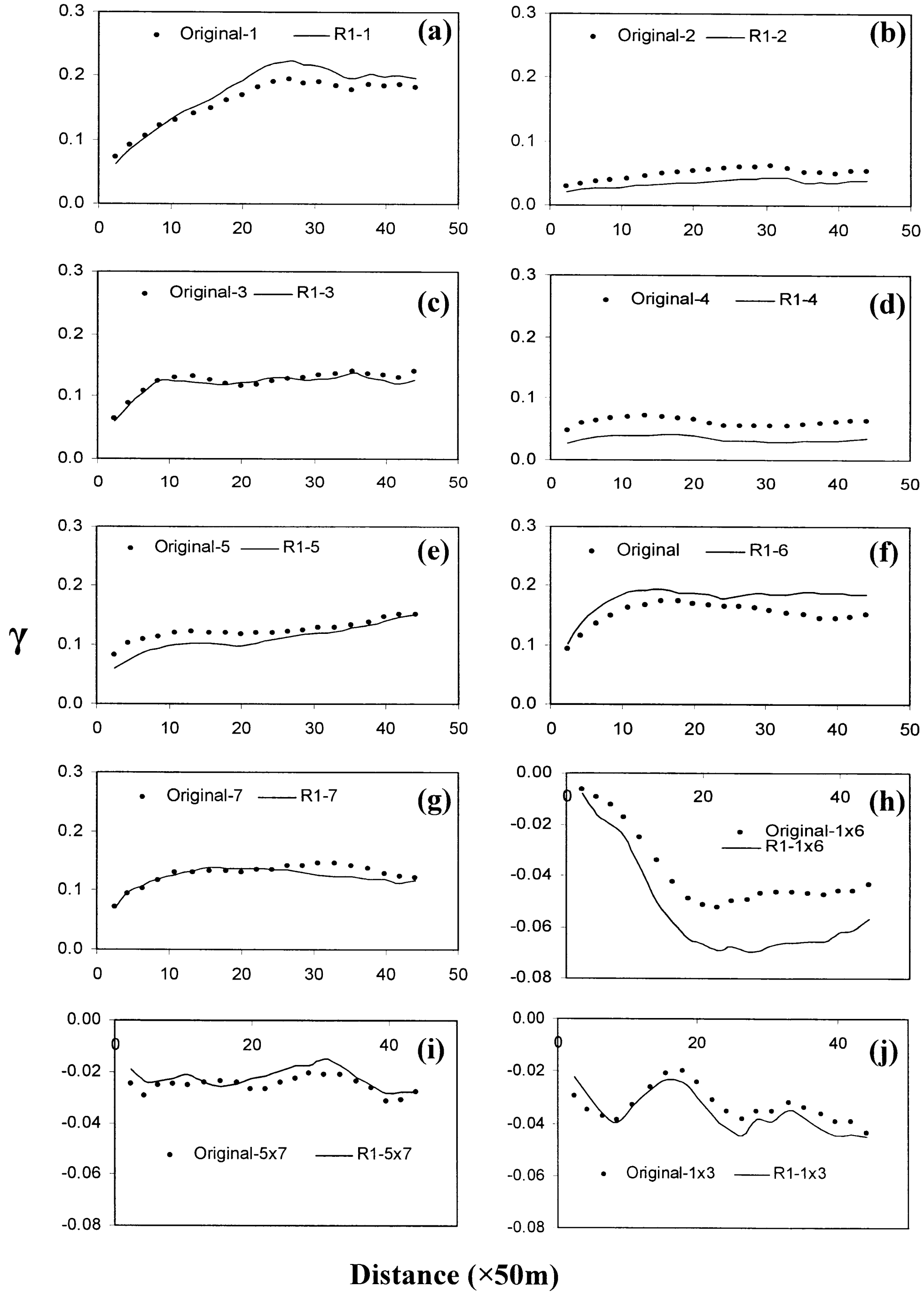

**Fig. 5. Indicator variograms and cross-variograms calculated from the original map and one realization (500m-R1) in Figure 4. Graphs (a) to (g) are indicator variograms of individual soil types and Graphs (h) to (j) are indicator cross-variograms between soil types. Graph legends represent the related maps and soil types; for example, R1-3 means Soil Type 3 in the simulated realization map R1, and Original-1 × 6 means Soil Type 1 vs. Type 6 in the original soil map. Note: The length should be multiplied by 50 to obtain the correct length values.**

same soil map and its subareas were simulated using parameters estimated from survey lines. The simulated results were very similar to those obtained using perfect knowledge of parameters, which shows that satisfactory results can be obtained using only survey lines of moderate density. But when survey lines are very sparse (e.g., 1000-m interval for the smaller soil map), parameters cannot be reliably estimated for all possible transitions in the simulation. As a remedy we proposed (but did not examine) that expert knowledge might be used to adjust preliminary estimates obtained from sparse data, or that parameters values might be borrowed from other areas.

From these simulations, we find the following merits of the TMC model for predictive soil mapping: 1. The model is robust, in that it is not very sensitive to input parameters—the TPMs. As the density of survey lines increases, there is a marked decline in the importance

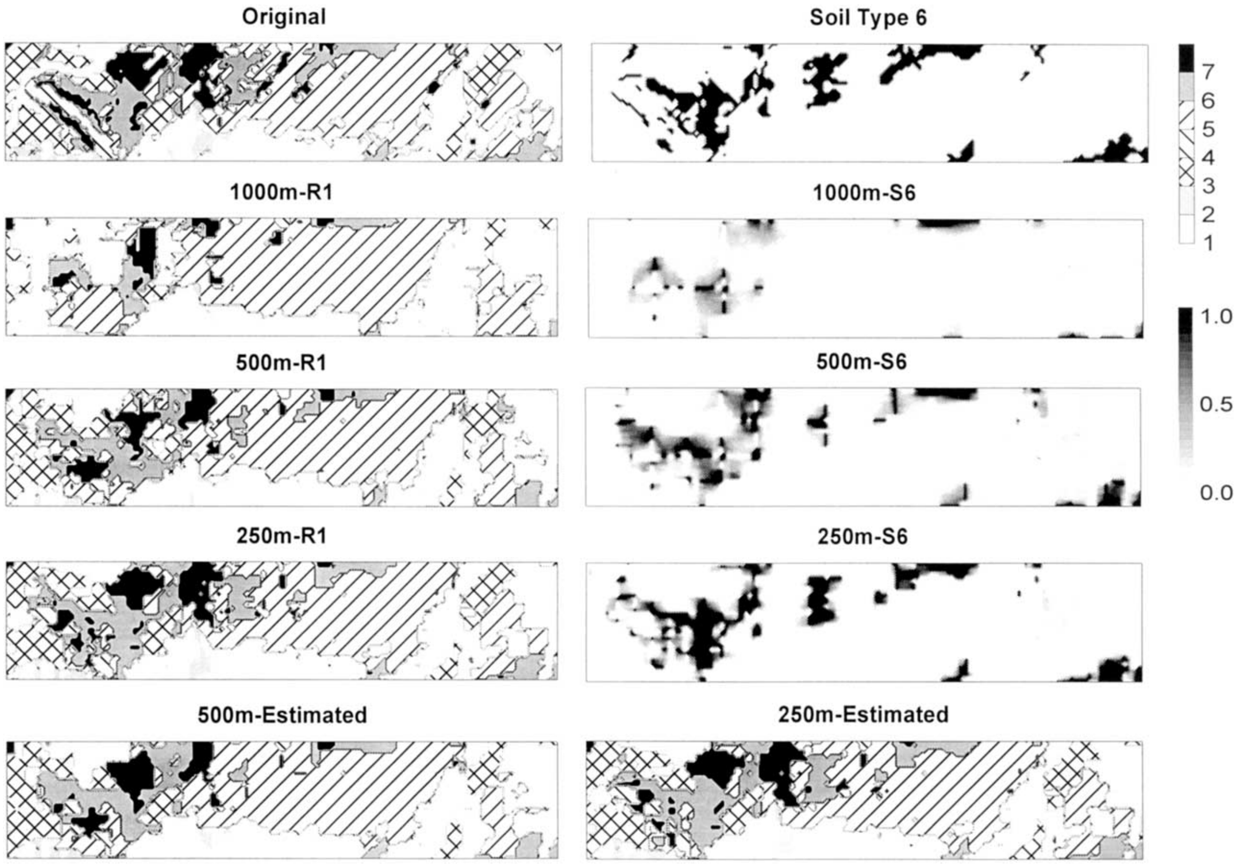

**Fig. 6. Simulated results of the soil type distribution in the study area under different conditioning schemes. Labels 1000m, 500m, and 250m represent conditioning schemes used, that is, survey line intervals. R1 means the first simulated realization based on the corresponding survey line interval. S6 means Soil Type 6. The bottom row gives the estimated soil map based on maximum occurrence probabilities. Parameters (i.e., one-step transition probability matrices) for each simulation are directly estimated from the survey lines used in the simulation.**

of input parameters but an obvious increase in the importance of survey lines in determining the spatial distribution of soil parcels. 2. Realizations can effectively represent the spatial patterns and abrupt boundaries of soil types as observed in nature. With a reasonable density of survey lines for conditioning and parameter estimation, realizations can represent the actual soil type distribution very well in terms of both patterns and distributional locations. 3. The simulation process is simple. The major input parameters are just three one-step TPMs in the three directions. 4. Simulations are highly efficient. Generating 100 realizations of the example soil map only needs 10 min or so on a personal computer (See the computer run times in Tables 2–5). 5. The model also provides a method for assessing the spatial uncertainty of soil distributions from survey line data.

One obvious constraint is the underestimation of minor soil types (or overestimation of major soil types) when survey lines are sparse (e.g., 1000-m interval). This constraint influences the direct use of simulated results when survey data are too sparse if the minor soil types are of serious importance. This constraint may be related to the independence assumption of one-dimensional Markov chains, and to the exclusion of transitions

**Table 4. Proportions of different soil types in the original soil map (the whole map), estimated from survey lines, and averaged from 100 simulated realizations for each simulation scheme (corresponds to Fig. 6).**

| Survey lines | | Proportions of different soil types | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Columns × rows | Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Run time |
| | m | | | | | | | | min |
| | | | | Original | | | | | |
| – | – | .2557 | .0296 | .1385 | .0452 | .3143 | .1335 | .0832 | – |
| | | | | Estimated from survey lines | | | | | |
| 9 × 3 | 1000 | .2868 | .0465 | .1512 | .0349 | .2752 | .1292 | .0762 | – |
| 17 × 5 | 500 | .2876 | .0398 | .1504 | .0383 | .2773 | .1379 | .0686 | – |
| 33 × 8 | 250 | .2690 | .0352 | .1440 | .0390 | .2944 | .1398 | .0788 | – |
| | | | | Simulated | | | | | |
| 9 × 3 | 1000 | .3813 | .0132 | .0986 | .0079 | .4039 | .0612 | .0317 | 18 |
| 17 × 5 | 500 | .2932 | .0178 | .1308 | .0172 | .3615 | .1228 | .0567 | 16 |
| 33 × 8 | 250 | .2696 | .0227 | .1396 | .0263 | .3266 | .1422 | .0730 | 12 |

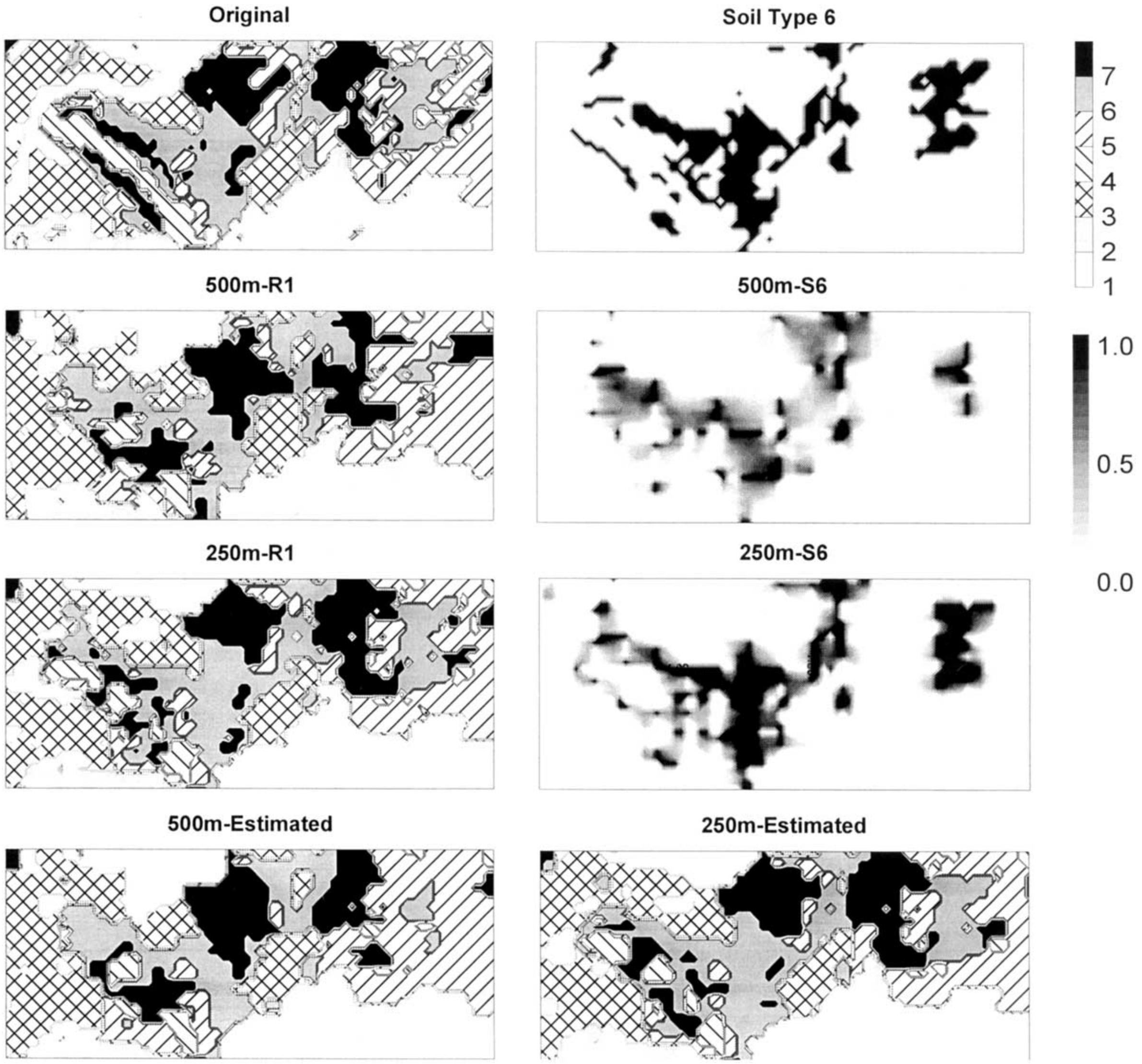

**Fig. 7. Simulated results of the soil type distribution in the left half of the study area under different conditioning schemes. Labels 1000m, 500m, and 250m represent conditioning schemes used, that is, survey line intervals. R1 means the first simulated realization based on the corresponding survey line interval. S6 means Soil Type 6. The bottom row gives the estimated soil map based on maximum occurrence probabilities. Parameters for each simulation are directly estimated from the survey lines used in the simulation.**

to different states from neighboring cells to the current cell in the CMCs. This problem also occurs in the Bayesian Markov random field model of Norberg et al. (2002), where it was suggested that Markov chain random fields might have the tendency of overestimating spatial dependencies of classes because of the possible existence of phase transitions. Further research is necessary to fully understand and overcome this constraint. A second limitation of this model is that currently the model only conditions simulations on survey line data. This may be

**Table 5. Proportions of different soil types in the original soil map (the left half), estimated from survey lines, and averaged from 100 simulated realizations for each simulation scheme (corresponds to Fig. 7).**

| Survey lines | | Proportions of different soil types | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Columns × rows | Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Run time |
| | m | | | | | | | | min |
| | | | | Original | | | | | |
| – | – | .2112 | .0522 | .1784 | .0764 | .1633 | .1733 | .1451 | – |
| | | | | Estimated from survey lines | | | | | |
| 9 × 5 | 500 | .2413 | .0737 | .1821 | .0694 | .1734 | .1373 | .1228 | – |
| 17 × 8 | 250 | .2205 | .0662 | .1852 | .0645 | .1601 | .1635 | .1400 | – |
| | | | | Simulated | | | | | |
| 9 × 5 | 500 | .2258 | .0410 | .2031 | .0383 | .2060 | .1491 | .1366 | 8 |
| 17 × 8 | 250 | .2108 | .0470 | .1967 | .0455 | .1580 | .1884 | .1536 | 6 |

advantageous for categorical soil data given the prominence of transect field soil survey. But to increase the model's application scope, it is necessary to further extend it for conditioning on point data.

Because a suitable density of survey lines is required for generating satisfactory simulated results, this method may find its usefulness in flatter areas where survey line data are relatively easier to acquire, and where terrain exerts only very subtle influence on soil patterns. In the regions with complex terrain, soil-landscape models (Zhu, 1997; Zhu et al., 2001), which infer soil types from environmental factors, may be more useful. Such models have been shown to do a good job of capturing topographic controls that are largely lacking in nearly flat terrain suggested as appropriate for TMC modeling.

Since this method is very flexible and efficient, it can be used for simulating over large areas at high resolution. However, when soils have markedly different spatial patterns in different subareas, TPMs should be estimated separately for each subarea.

## ACKNOWLEDGMENTS

Support by the Graduate School of University of Wisconsin–Madison and support by the Institute of Geographical Sciences and Natural Resources Research as part of the "One Hundred Person Plan" are greatly appreciated. We also thank the Institute for Land and Water Management, Catholic University of Leuven, Belgium for the early support to the first author's postdoctoral research on a related topic during 1998–1999.

## REFERENCES

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Stat. Soc. B 36:192–236.

Bouma, J., B.J. van Alphen, and J.J. Stoorvogel. 2002. Fine tuning water quality regulations in agriculture to soil differences. Environ. Sci. Policy. 5:113–120.

Brus, D.J., J.J. de Gruijter, D.J.J. Walvoot, F. de Vries, J.J.B. Brouswijk, P.F.A.M. Romkens, and W. de Vries. 2002. Mapping the probability of exceeding critical thresholds for cadmium concentrations in soils in the Netherlands. J. Environ. Qual. 31:1875–1884.

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

Burgess, T.M., and R. Webster. 1984a. Optimal sampling strategies for mapping soil types: I. Distribution of boundary spacings. J. Soil Sci. 35:641–654.

Burgess, T.M., and R. Webster. 1984b. Optimal sampling strategies for mapping soil types: II. Risk functions and sampling intervals. J. Soil Sci. 35:655–665.

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

Carle, S.F., and G.E. Fogg. 1997. Modeling spatial variability with one and multidimensional continuous-lag Markov chains. Math. Geol. 29:891–918.

Deutsch, C.V., and A.G. Journel. 1997. GSLIB: Geostatistics software library and user's guide. Oxford Univ. Press, New York.

Ehlschlaeger, C.R. 1998. The stochastic simulation approach: Tools for representing spatial application uncertainty. Ph.D. Diss. University of California, Santa Barbara.

Ehlschlaeger, C.R. 2000. Representing uncertainty of area class maps with a correlated inter- map cell swapping heuristic. Comput. Environ. Urban Syst. 24:451–469.

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

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

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

Goovaerts, P. 1999. Geostatistics in soil science: State-of-the-art and perspectives. Geoderma 89:1–45.

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

Heuvelink, G.B.M., and R. Webster. 2001. Modeling soil variation: Past, present, and future. Geoderma 100:269–301.

Journel, A.G. 1983. Nonparamtric estimation of spatial distributions. Math. Geol. 15:445–468.

Kite, G.W., and N. Kauwen. 1992. Watershed modeling using land classification. Water Resour. Res. 28:3193–3200.

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

Krumbein, W.C. 1968. Statistical models in sedimentology. Sedimentology 10:7–23.

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

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

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

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

McGwire, K.C., and P. Fisher. 2001. Spatially variable thematic accuracy: Beyond the confusion matrix. p. 308–329. *In* C.T. Hunsaker et al. (ed.) Spatial uncertainty in ecology. Springer-Verlag, New York.

Mowrer, H.T., and R.G. Congalton. (ed.) 2000. Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing. Ann Arbor Press, Chelsea, MI.

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

Rosen, L., and G. Gustafson. 1996. A Bayesian Markov geostatistical model for estimation of hydrogeological properties. Ground Water 34:865–875.

Scull, P., J. Franklin, O.A. Chadwick, and D. McArthur. 2003. Predictive soil mapping: A review. Prog. Phys. Geography 27:171–197.

Weissmann, G.S., and G.E. Fogg. 1999. Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework. J. Hydrol. (Amsterdam) 226:48–65.

Wingle, W.L., and E.P. Poeter. 1993. Uncertainty associated with semivarograms used for site simulation. Ground Water 31:725–734.

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

Zhu, A.X. 1997. A similarity model for representing soil spatial information. Geoderma 77:217–242.

Zhu, A.X., B. Hudson, J. Burt, K. Lubich, and D. Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Sci. Soc. Am. J. 65:1463–1472.

Zhu, A.X., and D.S. Mackay. 2001. Effects of spatial detail of soil information on watershed modeling. J. Hydrol. (Amsterdam) 248: 54–77.