Published online October 27, 2005

A Markov Chain-Based Probability Vector Approach for Modeling Spatial Uncertainties of Soil Classes

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

ABSTRACT

Due to our imperfect knowledge of soil distributions acquired from field surveys, spatial uncertainties inevitably arise in mapping soils at unobserved locations. Providing spatial uncertainty information along with survey maps is crucial for risk assessment and decision-making. This paper introduces a novel probability vector approach for spatial uncertainty modeling of soil classes based on an existing two-dimensional Markov chain model for conditional simulation. The objective is to find an accurate and efficient way to represent spatial uncertainties that arise in mapping soil classes. Joint conditional probability distribution (JCPD) represented by a set of occurrence probability vectors (PVs) of soil classes is directly calculated from conditional Markov transition probabilities, rather than the conventional approximate estimation from a limited number of simulated realizations. By visualizing the calculated PVs, information reflecting spatial uncertainty of soil distribution can be quickly assessed. We hypothesize that these directly calculated PVs are equivalent to the PVs estimated from an infinite number of realizations and thus realizations visualized from the calculated PVs represent the spatial variation of soil distribution. This hypothesis is supported by simulation results showing that: (i) with increasing the number of realizations generated by the Markov chain model from 10 to 100 and to 1000, PVs estimated from these realizations gradually approach the calculated PVs; (ii) similar to simulated realizations, realizations visualized from calculated PVs also can reflect the spatial patterns of soil classes and approximately reproduce the complex indicator variograms of soil classes of the original soil map.

Soil mapping is crucial for natural resource evaluation and environmental protection. However, the knowledge of soil distribution acquired through field survey (and other ways) is always imperfect. Thus spatial uncertainties inevitably arise in soil mapping; for example, a prominent problem is the difficulty in accurately determining the boundaries of multinomial soil classes in making choropleth maps on unsurveyed locations. Given the same observed dataset for a survey area, different persons normally delineate similar but different area-class maps of soil distribution because of their different interpretations over the unobserved portion of the landscape. A human-delineated soil map based on a set of observed sparse data only represents one possibility or good guess of soil occurrence in the survey area. An interpolated map using standard interpolation techniques may represent an optimal guess based on the dataset and the interpolation method used, but does not reflect the real spatial variation characteristics because of uneven smoothing effects (Goovaerts, 1997, p. 369–370). As Journel (1997, p. viii) pointed out: "The very reason for geostatistics and the future of the discipline lie in the modeling of uncertainty, at each node through conditional distribution and globally through stochastic images (conditional simulations)." Therefore, soil survey maps should be accompanied by related spatial uncertainty information. Data reflecting spatial uncertainty usually include occurrence probability maps and realizations provided by random field models (Zhang and Goodchild, 2002; Zhang and Li, 2005). This is particularly useful for risk assessment and decision-making. In addition to informing users about the existence and degree of the spatial uncertainty in delineated maps, a significant utility of conditionally simulated data using random field models is that they can be introduced into application models (e.g., process-based ecological models or hydrological models) to further infer response distributions of variables of interest (e.g., water budgets) (Goovaerts, 1996; Kyriakidis and Dungan, 2001; Li et al., 2001).

There are several problems hindering spatial uncertainty modeling: (1) It is difficult to mathematically calculate the JCPD of a random variable at all unknown locations in a study area of even moderate size. So far we have not yet found any existing geostatistical method that realizes this goal. The normal way for spatial uncertainty modeling is through generating a set of alternative realizations and then approximately estimating the JCPD (represented as a series of probability maps) from a number of realizations (Zhang and Goodchild, 2002; Zhang and Li, 2005). Thus, the accuracy of probability maps is largely dependent on the number of realizations used. (2) Many random field models have difficulties generating a sufficiently large number of realizations within acceptable computation time and computer storage (Dubrule and Damsleth, 2001), particularly when the number of classes or thresholds is large. With the ongoing development of computer techniques, this problem has been relaxed in recent years. For example, the sequential indicator simulation is an efficient variogram-based simulation method; in recent years it has been used for

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-X. Zhu, State Key Lab. of Resources and Environmental Information System, Inst. of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China and Dep. of Geography, Univ. of Wisconsin, Madison, WI 53706. Received 29 July 2004. *Corresponding author (weidong6616@yahoo.com).

Published in Soil Sci. Soc. Am. J. 69:1931–1942 (2005).
Soil Physics
doi:10.2136/sssaj2004.0258

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: CCDF, conditional cumulative distribution function; CMC, coupled Markov chain; JCPD, joint conditional probability distribution; PV, occurrence probability vector; PV-realizations, visualized realizations from the calculated PVs; TMC, triplex Markov chain; TPM, transition probability matrix; TP-realizations, simulated realizations using the TMC model through conditional transition probabilities.

spatial uncertainty modeling of land-cover classes with a small number of classes (Kyriakidis and Dungan, 2001; Zhang and Goodchild, 2002). But for some heavily iterative methods that may need a long computation time (days or even months) to generate a realization, such as the recently emerged Bayesian-Markov random field method (using simulated annealing) for sparse data modeling (Norberg et al., 2002), computation load is still a big concern. (3) How to effectively incorporate the complex spatial variation of random variables into a simulation method is still a difficult issue. Given the same conditioning dataset, a simulation method that can incorporate more information of spatial heterogeneity of the targeted random variable will generate more realistic realizations of the unknown 'truth', and thus more effectively reflect the spatial pattern of the targeted random variable and decrease the spatial uncertainty. Variograms provide widely accepted measures of spatial continuity, but conventional variogram-based methods are not capable of reflecting the complex interdependence of multiple classes and reproduce complex large-scale and long-range features (Ortiz and Deutsch, 2004; Liu and Journel, 2004). The major reasons may be that class interdependences (including cross correlations) are normally ignored in variogram-based simulation algorithms because of the awkwardness in cokriging a number of classes (Goovaerts, 1996, p. 911–912) and auto-variograms are too limiting in capturing complex heterogeneity of real patterns of categorical geographical variables (Caers and Zhang, 2004). In the 1990s and thereafter, large effort has been devoted to this issue in geostatistics, and significant progress has been made in recent several years. Major work has mainly focused on two general approaches: (i) incorporating multiple-point statistics into indicator simulation from various data sources, such as training images (Guardiano and Srivastava, 1993; Caers and Zhang, 2004), blasthole data (Ortiz and Deutsch, 2004), and structured paths (Liu and Journel, 2004); and (ii) using Markov chains in multi-dimensions to generate conditional realizations (Luo, 1996; Elfeki and Dekking, 2001; Li et al., 2004; Zhang and Li, 2005) or using transition probabilities to replace variograms in indicator simulation (Carle and Fogg, 1996).

Markov chain models have been used in soil science for characterizing spatial distribution of soil classes and soil layers. For one-dimensional applications, see Li et al. (1997, 1999). For two-dimensional applications, see Li et al. (2004) and Wu et al. (2004). The triplex Markov chain (TMC) model proposed by Li et al. (2004) for conditional simulation of soil classes in two-dimensions uses the method of coupled Markov chains (CMC) (Elfeki and Dekking, 2001) in its calculation of conditional transition probabilities. Only four nearest known neighbors along the orthogonal directions (i.e., x and y axes) are considered in determining the conditional transition probability at each point to be estimated and the conditional transition probability is explicitly calculated by directly conditioning to known data. Therefore the method is highly efficient. Recently, Wu et al. (2004) proposed an efficient Markov mesh model for reconstruction of binary images of heterogeneous soil structures. Through scanning the real image for simultaneous local parameter estimation, the model can well capture the spatial patterns and reproduce the variogram. The Markov mesh models (Abend et al., 1965; Qian and Titterington, 1991; Gray et al., 1994) represent a special subclass of Markov random fields (Besag, 1986), also outside conventional geostatistics. The Markov mesh models are cliqued-based and have been widely used for image processing. Differing from conventional Markov random field methods that use iterative updating approaches, Markov mesh models can be used to conduct efficient simulation by a one-pass way and are particularly used for image structure analysis (mainly for binary images) through unconditional simulation. Directly using Markov mesh models for conditional simulation on sampled data seems infeasible. Norberg et al. (2002) recently used Markov random fields for the geostatistical modeling purpose by using simulated annealing for iterative updating.

One typical feature of multinomial categorical variables in soil science such as soil layers and soil types is that they normally exhibit strong interdependence between multinomial classes. This interdependence includes strong cross correlations, juxtaposition relationships, and directional asymmetry in spatial occurrence of multinomial classes (Li et al., 1997; Li et al., 2004). Although many random field models can be used to simulate categorical variables (Chiles and Delfiner, 1999), Markov chain-based conditional simulation methods can better incorporate these special features because of the special characteristics of transition probabilities. For example, if Class A frequently occurs as a neighbor of Class B and seldom occurs as a neighbor of Class C, this juxtaposition relationship can be reflected in Markov transition probabilities and therefore respected in realizations. Similarly, if Classes A, B, C often occur as a sequence of ABC along a direction (e.g., west-to-east), this asymmetry also can be reflected in Markov transition probabilities along that direction and thus also respected in realizations. But such behaviors may be difficult to be captured with other spatial measures (Zhang and Li, 2005). The second typical feature of categorical soil variables is that the number of classes may be very large. For example, there may be dozens of soil series occurring in a watershed stretching over dozens of square kilometers (USDA, 1962). To deal with a large number of soil classes with due consideration of cross correlations, approaches based on iterative simulation methods (e.g., simulated annealing) or solving large cokriging equation systems may be unpractical in both computation time and numerical stability (Goovaerts, 1996). Unlike many conventional geostatistical methods that describe spatial correlations of classes by indicator covariance, Markov chain methods use transition probabilities of classes for the same purpose. In a Markov transition probability matrix (TPM), the diagonal elements (auto-transitions) represent autocorrelations of individual classes and the off-diagonal elements (cross-transitions) represent cross correlations between different classes. Because TPMs are normally estimated unidirectionally (e.g., from north to south), directional asymmetries can be included natu-

rally. Thus, class interdependence is implicitly incorporated in Markov chain models by cross-transition probabilities. Since simulation by Markov chain models does not involve complex computation, they have the advantage in dealing with a large number of classes for higher resolution simulation. This capability is especially important for simulation of soil classes, where normally many classes (or types) may be involved (Li et al., 2004).

In this paper, we proposed an innovative probability vector approach to directly calculating the JCPD of multinomial classes through the TMC model as an alternative to the conventional "brute force" method that estimates the same from a number of realizations. We represent the JCPD of multinomial classes as a set of PVs. The objective is to find a more accurate and efficient way to represent the spatial uncertainties that arise in mapping soil classes. We hypothesize that calculated PVs represent the PVs estimated from an infinite number of realizations and thus the visualized realizations from calculated PVs represent the spatial variation of multinomial classes. A simplified soil map is used as a case study for providing conditioning data and for evaluating this hypothesis. The probability vector approach proposed here is also applicable to conditional simulation with other existing Markov chain models based on single pass algorithms and explicit conditional transition probability expressions.

MATERIALS AND METHODS

On Modeling Spatial Uncertainty

It is difficult to calculate the JCPD function of all unknown points in a study area of any significant size. Single-point conditional probabilities based on only observed data merely provide a measure of *local uncertainty* (Goovaerts, 1997, p. 259–367). To represent *spatial (or joint) uncertainty*, the JCPD of all unknown points in the study area is required. The sequential simulation algorithms for spatial uncertainty assessment (Goovaerts, 1997, p. 369–436) rely on a set of realizations. First, a one-point conditional cumulative distribution function (CCDF) is modeled and sampled at each of the unknown locations visited along a random sequence, and each one-point CCDF is made conditional not only to the original observed data but also to all values simulated at previously visited locations. Occurrence probability maps (or PVs) representing the JCPD may be approximately estimated from a large number of realizations, and used to model the joint uncertainty of the random variable in a study area.

The Markov chain conditional simulation methods such as the TMC model (Li et al., 2004) follow the same simulation technique as sequential simulation algorithms except for using smaller (and changeable) neighborhoods of conditioning data and a fixed sequence (path). Although probability maps can be estimated from multiple realizations in the TMC model (Zhang and Li, 2004; Zhang and Li, 2005), the accuracy of estimated occurrence probabilities depends on the number of realizations. When the simulation area is large and the number of involving classes is large, estimating accurate PVs by generating a large number of realizations is computationally burdensome. The facts that the TMC model (a) is a single-pass simulation method, (b) has an explicit conditional transition probability expression, and (c) conditions the estimate of each unknown point to a few (four) nearest known points on fixed axis directions, suggests another way. In particular, it may be possible to calculate the JCPD (i.e., PVs) from the Markov transition probabilities by conditioning to both observed data and previously estimated locations, rather than approximately estimate them from a large number of simulated realizations. Thus, single realizations may be directly obtained from the calculated PVs by Monte Carlo sampling. Below we explain how to make this calculation.

For a categorical variable, a JCPD of all unknown points in a study area can be expressed as

$$p[z_1(u_1),...,z_N(u_N)|(n)] \quad [1]$$

where $z_1(u_1),...,z_N(u_N)$ represent all of the N unknown points (i.e., grid cells or pixels) in a study area with z standing for state and u for location, and n represents all of the observed data points. Here $n + N$ is equal to the total number of pixels, known plus unknown, of the study area.

By using the Bayes' Theorem (i.e., the definition of conditional probability), Expression [1] can be factored as

$$p[z_1(u_1),...,z_N(u_N)|(n)] = p[z_N(u_N)|(n + N - 1)]\times...\times p[z_i(u_i)|(n + i - 1)]\times...\times p[z_2(u_2)|(n + 1)] \times p[z_1(u_1)|(n)] \quad [2]$$

where $z_1(u_1),...,z_N(u_N)$ follow the visiting sequence of a simulation, that is, $z_1(u_1)$ is the first unknown point visited, and $z_N(u_N)$ is the last unknown point visited in a single-pass simulation process. The later-visited point is conditioned to both observed data and previously visited locations so that all the estimates of unknown points are spatially related. Equation [2] is the JCPD function for categorical variables. See Goovaerts (1997, p. 377, Eq. [8.6]) for the similar expression for thresholds of continuous variables.

To solve Eq. [2], our task is to solve every one-point conditional probability distribution on the right-hand side, for example, the conditional probability distribution of the ith visited point

$$p[z_i(u_i)|(n + i - 1)] \quad [3]$$

It is clear that the conditional probability distribution of a later-visited point is dependent on the solutions of conditional probability distributions of all previously visited points, that is, solving Eq. [3] needs first solving the conditional probability distributions of the first $i - 1$ unknown points in the visiting sequence.

It is difficult to directly solve Eq. [3] in sequential simulation algorithms without all of the $(n + i - 1)$ points being known. For example, kriging in sequential simulation algorithms deals only with single indicator values of all of the $(n + i - 1)$ points, not their conditional probability distributions (i.e., vectors of conditional probability values). That means we have to estimate the CCDF of one unknown point first, allocate a value to the point by Monte Carlo sampling, then estimate the next unknown point. Thus, by following a (random or fixed) visiting sequence, sequential simulation algorithms can generate a set of alternative realizations to represent the JCPD (i.e., spatial uncertainty). However, with the simple neighborhood structure and the simple explicit solution of conditional transition probability in the TMC model, Eq. [3] can be directly solved as long as the one-point conditional probability distributions of the $(n + i - 1)$ points are known.

Note that in practical use, the n in Eq. [3] need not be all the observed data and also i need not be all previously simulated values in the whole study area, because the data closest to the location being estimated tend to screen the influence of distant data. In the practice of sequential simulation algorithms, only the original data and those previously simulated

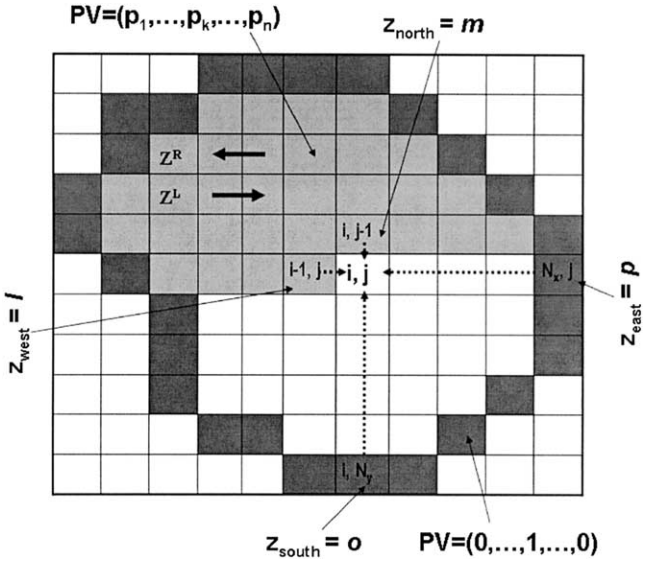

Fig. 1. Illustration of the simulation algorithm of the TMC model. The state of the current cell (i,j) depends on its four nearest known neighbors in four axis directions (west, east, south, and north), that is, grid cells ($i,j-1$), ($i-1,j$), (N_x,j), and (i,N_y), their states being denoted as m, l, p, and o, respectively. The dark gray cells stand for observed data (i.e., parts of survey lines, survey lines may be regular or irregular). The light gray cells stand for cells already visited. The thick arrows illustrate the alternate proceeding directions of the TMC model. The dashed arrows illustrate the interactions between the current unknown cell and its four nearest known neighbors in axis directions.

values closest to the location u_i are retained. Including all observed data and previously simulated data is not only unnecessary but also impossible from a practical standpoint.

The TMC Model

The TMC model uses a simulation algorithm similar to sequential simulation algorithms in kriging geostatistics, but rather than estimating CCDFs, it estimates the conditional transition probabilities. Unlike the sequential simulation algorithms, which need to define a search neighborhood to limit the number of conditioning data, the TMC model only conditions on four nearest known locations along the axes (Li et al., 2004). The conditional probability distribution of the ith unknown point can be simplified as

$$p[z_i(u_i)|(n+i-1)] = p[z_i(u_i)|z_{\text{west}},z_{\text{north}},z_{\text{east}},z_{\text{south}}] \quad [4]$$

under the neighborhood structure of the TMC model (Fig. 1), where z_{west}, z_{north}, z_{east}, and z_{south} represent the four nearest known neighbors along the axes. These four nearest known neighbors, each in one axis direction, include previously simulated points. Therefore, in the TMC model, all observed data and previously simulated data are used in the simulation process and the simulated data are spatially correlated.

There are various ways to decompose the right-hand side of Eq. [4]. Elfeki and Dekking (2001) provide a simple solution with a full independent assumption of two single Markov chains, each in one direction. Such a simple solution permits efficient conditional simulation of subsurface vertical sections with borehole data (through conditioning to a future state in a one-dimensional Markov chain). The TMC algorithm introduced in Li et al. (2004) is based on an extension of the solution and therefore has an explicit solution of the right-hand side of Eq. [4]. It conducts simulations by conditioning on survey line data. Although other kinds of data may be used after the development of a parameter estimation strategy from point data, survey line data are preferred for observable categorical variables because of their advantage in representing spatial continuity of class parcels (polygons). Survey line data may be acquired by observing class boundary changes along a line (see Zhang and Li, 2005). The TMC model conducts simulation by following a fixed path row by row from top to bottom (Fig. 1). Therefore, among the four nearest known neighbors two are adjacent predecessors (one side pixel z_{west} or z_{east} and one upper pixel z_{north}) and two are '*future*' known states (observed data) located on survey lines (one pixel may be z_{west} or z_{east} and one is z_{south}). Here '*future*' means that the Markov chain has not proceeded to these locations.

A TMC can be explained as two extended CMCs in the opposite directions—the CMC Z^L and the CMC Z^R. The two CMCs proceed alternately in opposite directions. The alternate paths in the TMC model are a necessity, not only for avoiding directional artifacts (i.e., the directional trend effect along the simulation direction; see demonstrations in Gray et al. (1994)), but also for effectively imposing influences of known data (simulated or observed) at both (left and right) sides on the estimate of the current unknown location. Therefore, the TMC model is composed of a conditional transition probability pairs ($p^L_{lm,k|qo}$, $P^R_{lm,k|qo}$), which represent the left-to-right and right-to-left CMCs, respectively. Here, l and m represent the specific states of the two adjacent predecessors, k represents the state of the current pixel, q and o represent the specific states of the two future known pixels. For more detailed explanation of the expressions of the conditional transition probability pairs, see Li et al. (2004).

Occurrence Probability Vectors

The probability vector approach calculates the JCPD by following the visiting sequence (simulation path) and using the conditional transition probability expressions in the TMC model. The JCPD is represented as a set of PVs, and each PV actually represents a one-point conditional probability distribution in Eq. [3]. The calculation of the PV of each unknown point is conditioned to the observed data and previously estimated values—not values of classes, but the calculated PVs of classes at those locations.

A PV consists of a set of probability values representing the likelihood that each class occurs at a particular point (i.e., grid cell or pixel). Mark and Csillag (1989) and Goodchild et al. (1992) discussed the feasibility of using probability vectors to represent the transition between two classes if there were cartographic (or locational) errors. If each class is represented as a probability distribution, categories (or classes) have transition zones varying gradually between the maximum and minimum class likelihood with 0.5 at the location of class 'boundary'. Zhang and Li (2005) estimated PVs from a large number of realizations to represent the spatial uncertainty of multinomial land-cover classes and demonstrated the transition zones between classes. The PVs consist of the occurrence likelihoods of multiple classes that possibly occur in a study area. They can be 'hardened' into an area class map (i.e., the prediction map) by choosing a standard such as the maximum occurrence probabilities and assigning corresponding class values (or labels). Such PVs may be used to describe the locational uncertainty of multiple classes arising in the mapping process of hand-delineated area class maps. The PV approach developed here provides more accurate PVs in a more efficient way.

The PV for a pixel or grid cell (i, j) can be expressed as

$$\text{PV}(i,j) = [p_1(i,j),..., p_k(i,j),..., p_n(i,j)] \quad [5]$$

where n represents the number of classes, the i and j are used to represent the cell location on a grid, and $p_k(i,j)$ is an element of a PV, representing the conditional occurrence probability of class k in the cell (i, j). Here $p_1(i, j)$, …, $p_n(i, j)$ are actually the n specific values of the right-hand side of Eq. [4] for the n classes of a categorical variable.

Using the TMC model, PVs for every grid cell can be estimated. If a cell is located on the survey lines (i.e., an observed point), its PV is already known with its certain element $p_k(i,j)$ being 1 if class k occurs in the cell and other elements being 0, that is,

$$\mathrm{PV}(i,j) = (0, \ldots, 0, 1, 0, \ldots, 0) \quad [6]$$

For an unobserved cell, we calculate any element $p_k(i,j)$ of its PV as

$$p_k(i,j) = \sum_{q=1}^{n}\sum_{o=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n}(p_{lm,k|qo} \times p_l \times p_m \times p_q \times p_o) \quad [7]$$

where p_l and p_m represent elements of PVs of the two adjacent predecessors—one side cell and the upper cell, respectively, of which PVs are already known (they are visited points or survey data), but of which states (l and m) are not decided if they are not on survey lines. Because the two future states q and o are known states as survey data (see Fig. 1), their PVs can be cancelled in Eq. [7]. So we have

$$p_k(i,j) = \sum_{l=1}^{n}\sum_{m=1}^{n}(p_{lm,k|qo} \times p_l \times p_m) \quad [8]$$

Considering that the TMC model is composed of a conditional probability pair and the two CMCs proceed in opposite directions, we have

$$p_k(i,j) = \sum_{l=1}^{n}\sum_{m=1}^{n}[p^L_{lm,k|qo} \times p_l(i-1,j) \times p_m(i,j-1)] \quad [9]$$

for the left-to-right CMC Z^L, and

$$p_k(i,j) = \sum_{l=1}^{n}\sum_{m=1}^{n}[p^R_{lm,k|qo} \times p_l(i+1,j) \times p_m(i,j-1)] \quad [10]$$

for the right-to-left CMC Z^R.

The PVs for the study area hold all information about the JCPD of all classes in the area. Corresponding to this probability vector approach, we refer to the realization generation algorithm through conditional transition probabilities of the TMC model (Li et al., 2004) as the simulation approach. The calculation of PVs is once for all for a dataset, and the time needed for this calculation is equivalent to that used for generating one realization using the simulation approach.

Visualizing the Probability Vectors

From the calculated PVs of all cells in a study area, we can get the following information: (1) a series of occurrence probability maps of individual classes, which represent where and with how much certainty (or uncertainty) a class will occur in the study area; (2) the maximum occurrence probability map, that is, the map of greatest occurrence probabilities among occurrence probabilities of all classes at every location, which represents how much certainty (or uncertainty) exists with each point in the prediction map; (3) the prediction map based on the maximum occurrence probabilities, which represents the optimal prediction; and (4) single realizations, each of which represents one possible configuration of soil classes in the study area based on the survey data. The generation of such realizations is accomplished by Monte Carlo simulation based on the PVs. We refer to this as a "visualization" process. This visualization process can use any path, fixed or random, because it does not involve calculation of conditional transition probabilities that depend on predecessor cells. Therefore, the predictive mapping process using this probability vector approach actually consists of the calculation of PVs and the visualization of the PVs. The visualization of PVs is very quick (normally within seconds) because no complex computation is needed.

In the following sections, for clarity we will refer the realizations visualized from calculated PVs as PV-realizations, and the realizations generated by the simulation approach of the TMC model through conditional transition probabilities as TP-realizations.

Table 1. Input parameters (one-step transition probability matrices [TPMs] and grid information) estimated from survey lines with an interval of about 500 m (corresponds to Fig. 5a)†.

Soil class	1	2	3	4	5	6	7
			TPM in the x-direction				
1	.835	.073	.055	.000	.009	.018	.009
2	.167	.750	.000	.000	.056	.000	.028
3	.075	.015	.791	.015	.015	.075	.015
4	.043	.000	.000	.696	.130	.087	.043
5	.044	.000	.000	.074	.794	.059	.029
6	.069	.000	.085	.000	.085	.660	.106
7	.044	.000	.000	.044	.111	.067	.733
			TPM in the x'-direction				
1	.820	.054	.045	.009	.027	.027	.018
2	.222	.750	.028	.000	.000	.000	.000
3	.095	.000	.841	.000	.000	.063	.000
4	.000	.000	.042	.667	.208	.000	.083
5	.014	.029	.014	.043	.771	.057	.071
6	.043	.000	.106	.042	.085	.660	.064
7	.023	.023	.023	.023	.045	.114	.750
			TPM in the y-direction				
1	.833	.037	.093	.000	.000	.019	.018
2	.083	.833	.000	.000	.083	.000	.000
3	.078	.016	.797	.000	.000	.109	.000
4	.040	.000	.000	.600	.120	.120	.120
5	.058	.038	.000	.019	.712	.077	.096
6	.020	.000	.102	.082	.102	.633	.061
7	.000	.000	.049	.098	.097	.049	.707

† States: 7; grid columns: 80; grid rows: 34.

Case Study

A simple case study is used to verify the probability vector approach. We calculated the PVs of soil classes on a small area of 4×1.7 km^2. The soils in the area were classified into 7 soil classes (or types). The specific soil types are themselves of no particular interests in this study because our method does not involve any physical processes. They are just used here to show that spatial heterogeneity of soil types or classes can be characterized (Li et al., 2004). The study area is discretized into an 80×34 grid with a cell size of 50×50 m. Survey lines are distributed in the study area (i.e., the map) with an interval of about 500 m. Survey line data may be acquired by observing soil class boundary changes along a line. It is not necessary to observe every point along a line within class parcels (polygons) since their labels are the same within a parcel. Although the model itself does not limit the shape of survey lines, here we mainly use regular survey lines for the convenience of parameter estimation in the demonstration of the probability vector approach. One single-step TPM would be enough for a simulation without considering anisotropies and asymmetry. Here in our case study we used three one-step TPMs, all of which were estimated from these regular survey lines (Table 1). We first used the probability vector approach to calculate the PVs of all grid cells and visualized them. For the purpose of a comparison, we then used the simulation approach to generate some realizations and esti-

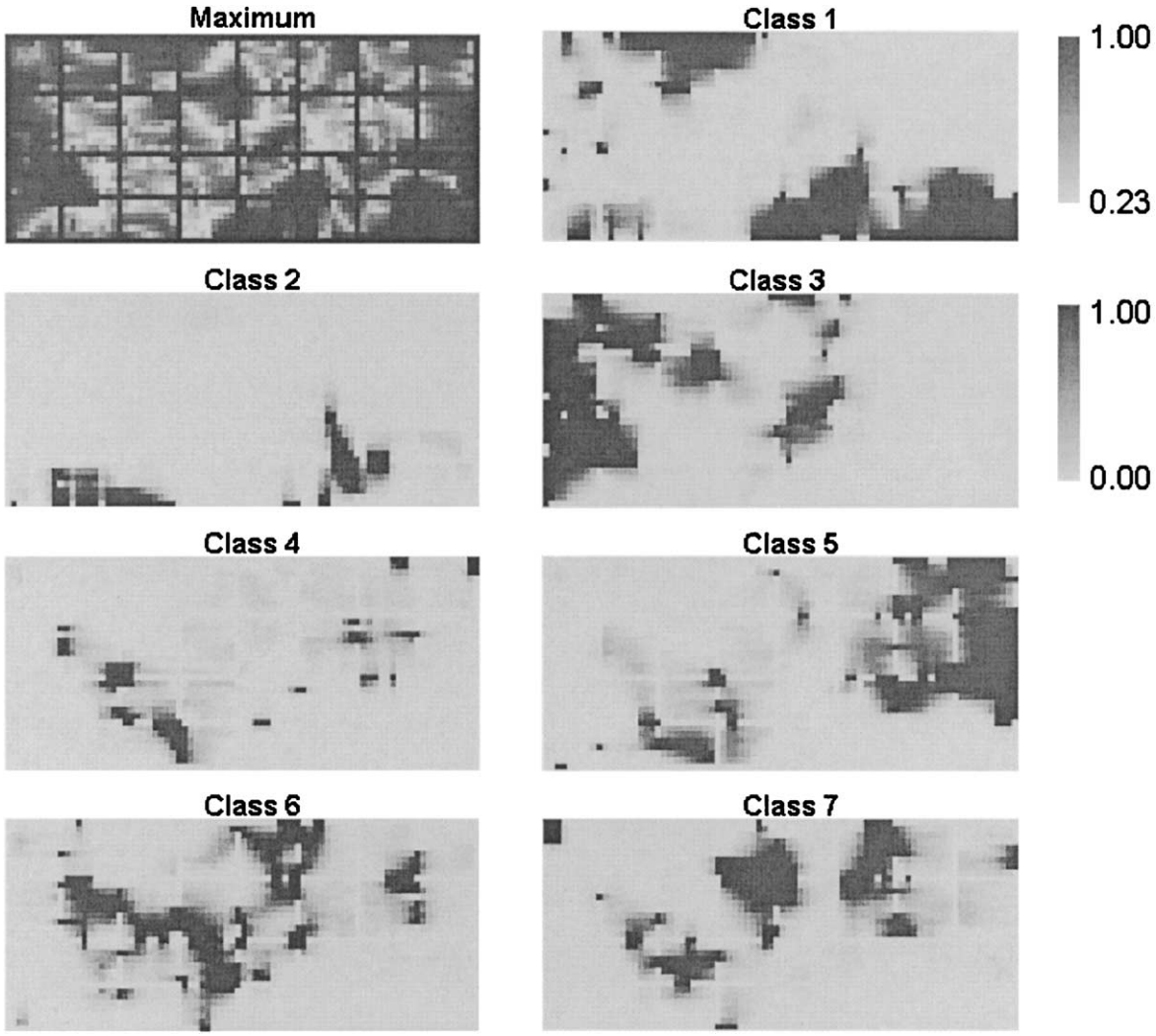

Fig. 2. The maximum occurrence probability map of soil classes and the occurrence probability maps of individual soil classes visualized from the calculated PVs using the probability vector approach.

mated probability maps from these simulated TP-realizations. The original soil map is used to represent the 'truth' (unknown in a real world application) against which we can evaluate our results.

Indicator variograms are widely accepted spatial continuity measures for categorical variables. To evaluate whether or not the PV-realizations reproduced this spatial statistical property as the TP-realizations did (Li et al., 2004), we calculated related indicator (cross) variograms.

RESULTS AND DISCUSSION

Figure 2 shows the probability maps visualized from the PVs directly calculated using the probability vector approach. Figure 3 gives the probability maps estimated from 100 TP-realizations using the simulation approach. It can be seen that there were no apparent differences between the results from both approaches except for minor details. We also estimated the PVs from 1000 TP-realizations (not shown) and found that they were essentially the same as the calculated PVs. No visual difference could be seen between the probability maps visualized from the calculated PVs and those estimated from 1000 TP-realizations. However, the probability maps estimated from 10 TP-realizations (Fig. 4) showed obvious deviations from those based on the calculated PVs (see the scattered gray patches in Fig. 4). This shows that the PVs estimated from TP-realizations would approach the calculated PVs (i.e., the JCPD) with increasing the number of realizations. In other words, the calculated PVs represented the PVs estimated from the ensembles of TP-realizations that the TMC model could generate. This also verifies that the calculated PVs do capture the observed joint uncertainty.

The maximum occurrence probability map is also called "purity map" in Bierkens and Burrough (1993a, 1993b), who used simple indicator kriging to estimate the occurrence probabilities of water table classes. In the maximum occurrence probability map, it is clear that the smallest probabilities occurred on the boundaries between more homogeneous areas (i.e., large soil parcels). The light-gray stripes are so-called transition zones, which are appropriate to represent the spatial uncertainty of class boundaries in multinomial area-class maps (Mark and Csillag, 1989; Goodchild et al., 1992; Zhang and Li,

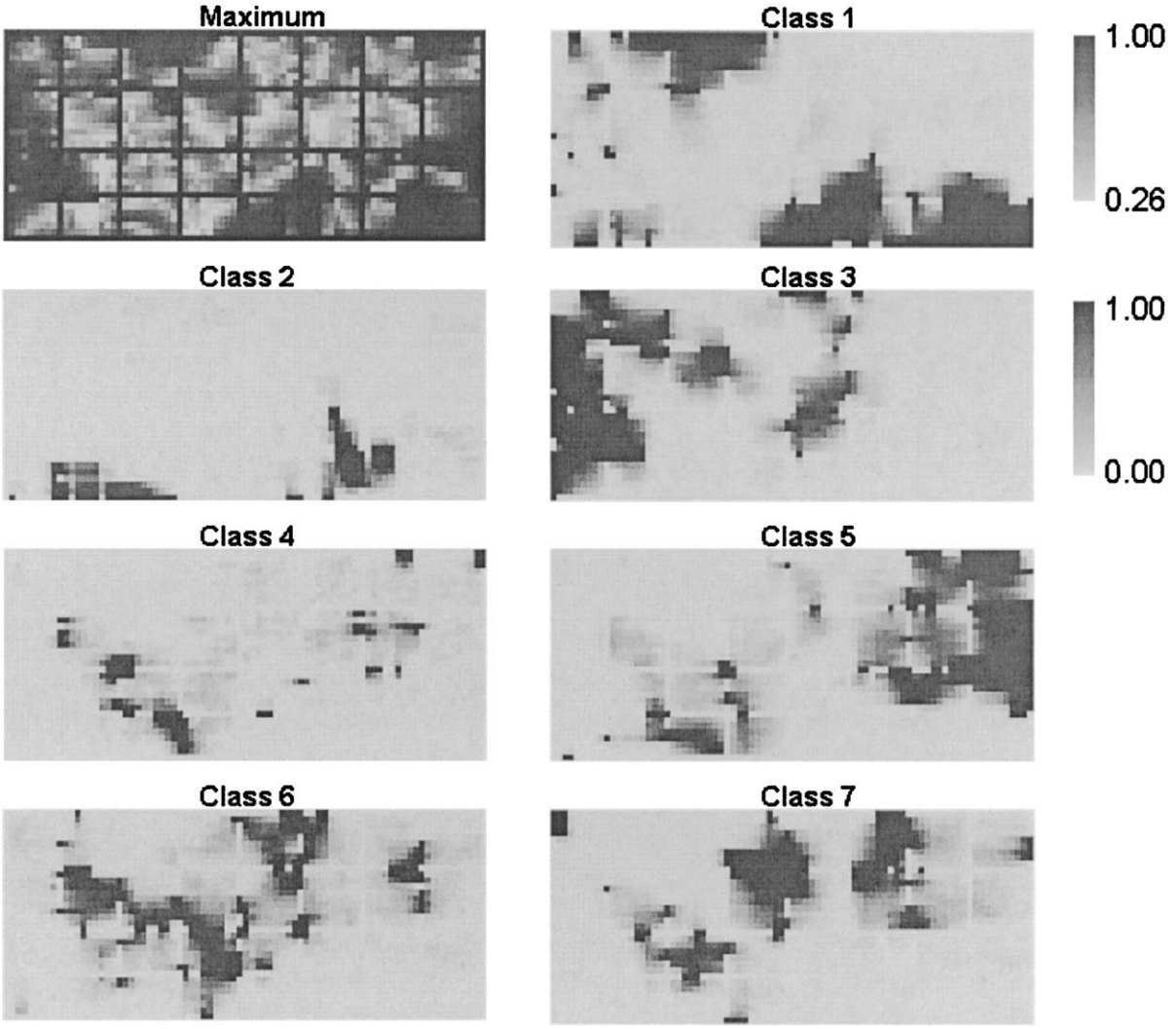

Fig. 3. The maximum occurrence probability map of soil classes and the occurrence probability maps of individual soil classes visualized from the PVs estimated from 100 TP-realizations.

2005). They also revealed places where more observations were needed to accurately predict soil class.

For a large area or high-resolution simulation, even when using efficient random field models, generating a large number (e.g., 100–1000) of realizations is time-consuming. But if the number of simulated realizations is small, the probability maps estimated from a few realizations may be very inaccurate. Figure 4 indicates that the results estimated from 10 TP-realizations deviated significantly from the results estimated from 100 TP-realizations. Therefore the uncertainty information estimated from a small number of simulated realizations may not be reliable.

Such uncertainty information is crucial for users to understand the possible distribution of soils in their study areas, and the positional uncertainty of soil classes existing in the predicted soil map. More importantly, these uncertainty data may serve as direct input data of risk assessment and decision-making models so that decision makers can make more reasonable decisions with the awareness of spatial uncertainties existing in their reference maps—usually hand-delineated maps or model-interpolated maps (Zhang and Goodchild, 2002).

A prediction map visualized from PVs based on maximum occurrence probabilities represents the optimal prediction. From Fig. 5, it can be seen that the prediction maps from the probability vector approach were similar to those obtained from the TMC simulation approach. This is not surprising, because both PVs (the calculated PVs and the PVs estimated from 100 TP-realizations) had similar values.

But looking at the PV-realizations and the TP-realizations (Fig. 5), surprisingly we found differences. The TP-realizations had bigger patches and clear boundaries, and they closely resembled the prediction map; however, the soil class parcels in the PV-realizations were obviously more fragmentary, particularly at the boundary zones between classes. The reason for this discrepancy may be related with probabilities used for determining the state of a cell in the realization generation process. In the simulation approach, the cumulative conditional transition probability function was used, where the max-

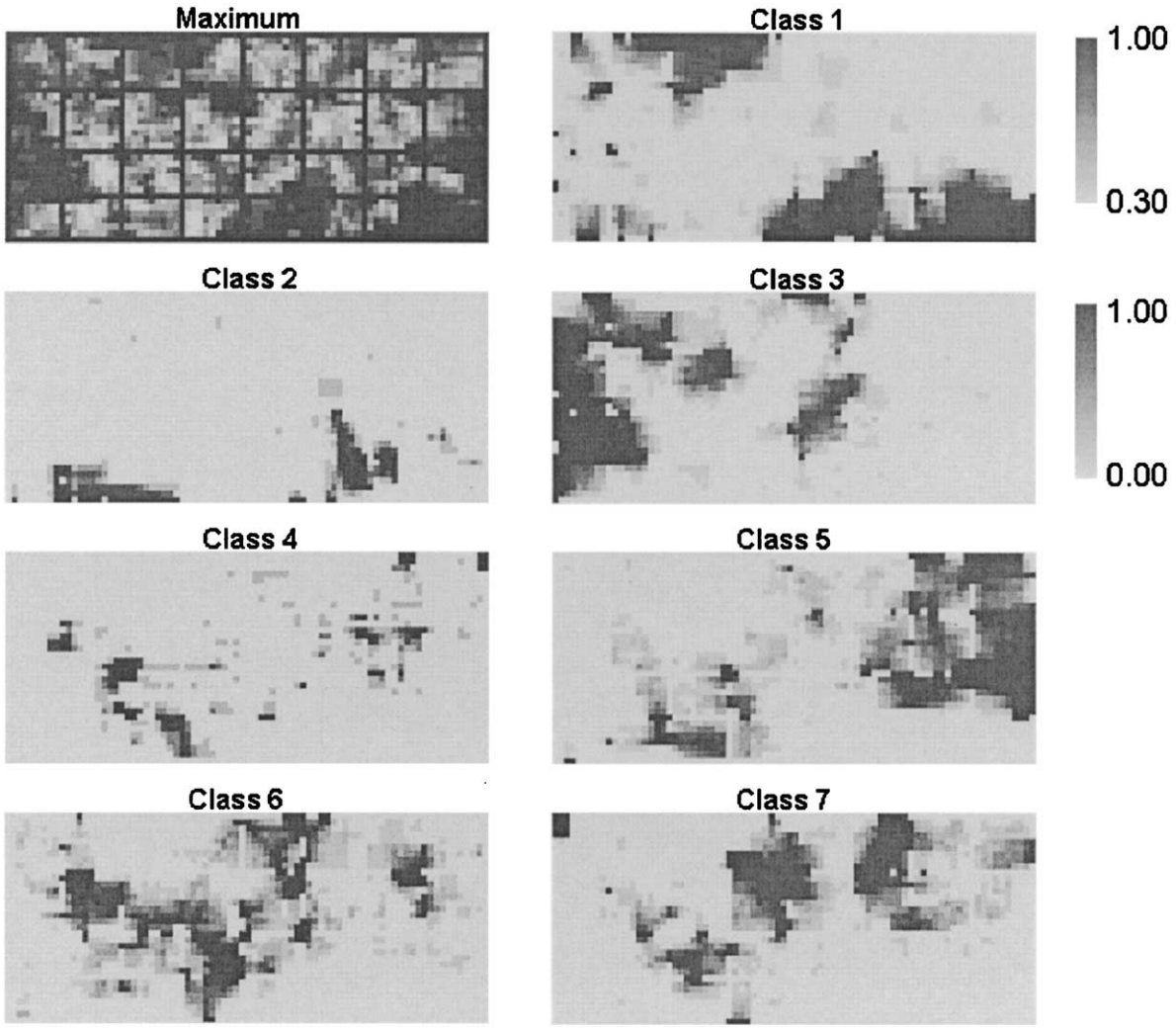

Fig. 4. The maximum occurrence probability map of soil classes and the occurrence probability maps of individual soil classes visualized from the PVs estimated from 10 TP-realizations.

imum conditional transition probability to a preferential state for the current cell was more prominent and other states had little chance to occur. But in the probability vector approach, the CCDF was used, where the maximum occurrence probability of the preferential state at the current cell might not be so prominent and other states also had some chance to be drawn in Monte Carlo sampling. Although the PV-realizations did not have abrupt boundaries, this should not be interpreted as a defect. Rather, it might reflect a situation where the transition between different soil classes was not so abrupt in the field, and thus there should be some interlacing of adjacent classes. For example, in the transition zones from grassland to forest, the two land cover types may be interlaced with small patches, and this should also be true for the corresponding soil types.

Figure 6 gives only a subset of omnidirectional (cross) semivariograms from the original soil map, the first PV-realization (Fig. 5c) and the first TP-realization (Fig. 5f). It can be seen that both kinds of realizations approximately reproduced the (cross) variograms. The small deviations between observed and simulated results were expected for single realizations. That is, variograms from multiple realizations normally should fluctuate around the observed results (Goovaerts, 1997). Of course, if conditioning data were more dense, such discrepancies should largely decrease. From cross variograms in Fig. 6, it also can be seen that the spatial cross-correlation between soil classes was very complex; this kind of spatial dependence is difficult to model using conventional cross variogram models.

For the purpose of visual comparison and method verification, we provided the original soil map in Fig. 5. (Of course in a real world application we would not have an "original" soil distribution map; the only data available would be an observed dataset that normally only accounts for a small portion of the study area.). From Fig. 5 it can be seen that the visualized PV-realizations, simulated TP-realizations, and the prediction maps all imitated the original soil map to some extent in spatial patterns of soil classes. The degree of similarity increases with increasing the number of conditioning data and vice versa. This means that the calculated PVs and the visualized realizations all effectively captured the spatial certainty provided by the observations.

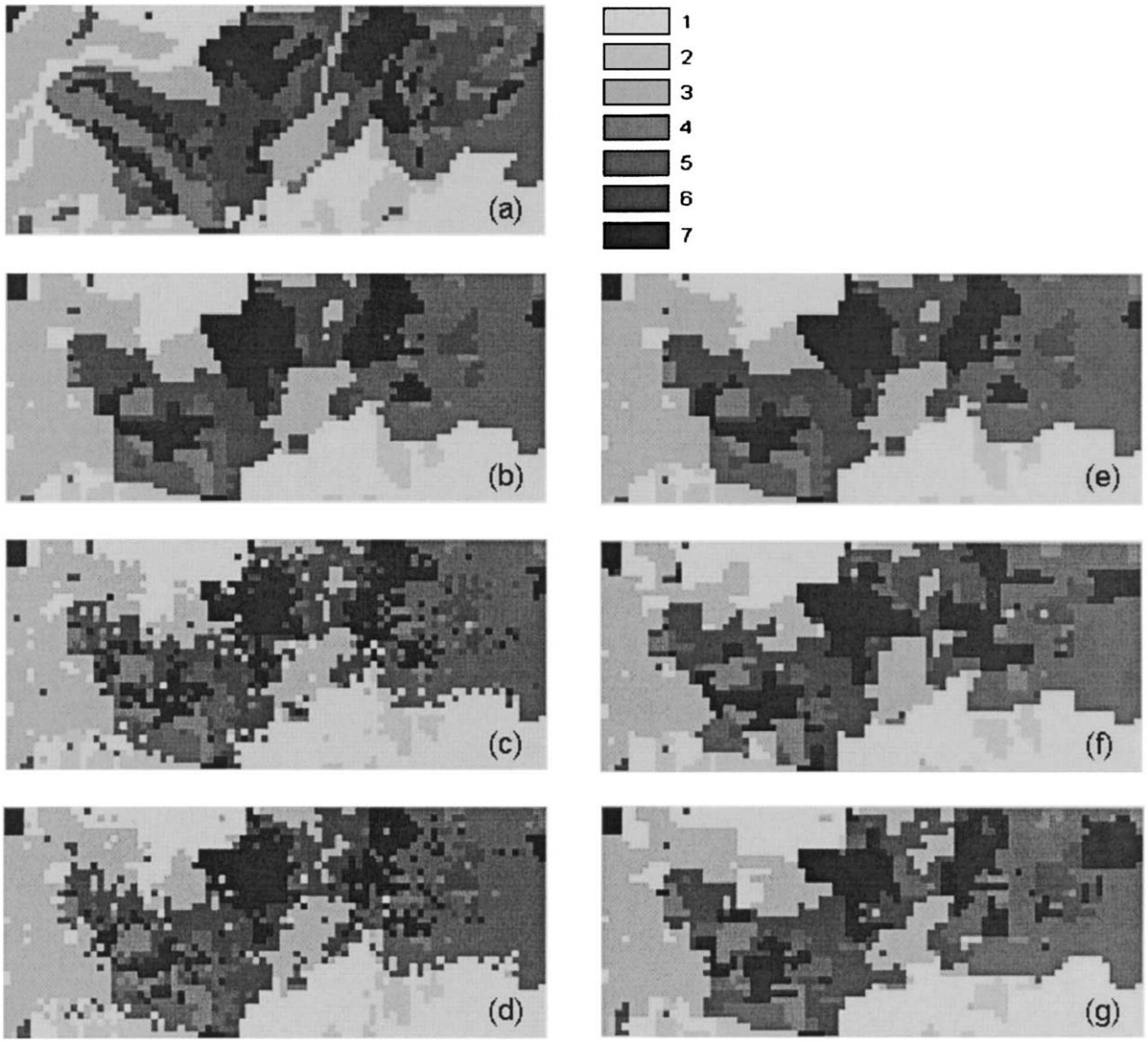

Fig. 5. Prediction maps and single realizations using the probability vector approach and the simulation approach of the TMC model. (a) The reference soil map. (b) The prediction map from the probability vector approach. (c) The first PV-realization. (d) The second PV-realization. (e) The prediction map estimated from 100 TP-realizations. (f) The first TP-realization. (g) The second TP-realization.

REMARKS

The purpose of geostatistical simulation, for our understanding, is to predict the unknown from the known with as little uncertainty as possible, and at the same time to reflect the inevitable spatial uncertainty. A good simulation conditioned to an observed dataset should reflect the uncertainty contained in the dataset as much as possible. That means that while spatial uncertainty is inevitable because of the limited survey data, it would be desirable for realizations conditioned on that limited data to imitate the "real" map (assuming we had it in model testing) as much as possible, since structure-imitating is also one simulation purpose (Koltermann and Gorelick, 1996). Conventional spatial continuity measure such as indicator autovariograms only measures a part of spatial variation information (i.e., autocorrelations) contained in sampled data or the target variables (Goovaerts, 2002). For example, various strongly different types of heterogeneities may produce similar autovariograms (see Caers and Zhang, 2004). Therefore, approximately reproducing indicator autovariograms is necessary but may not be sufficient for a successful simulation. The efforts in geostatistics in recent years, whether incorporating multi-point statistics from training images (or dense datasets) into indicator simulation approaches (e.g., Ortiz and Deutsch, 2004; Caers and Zhang, 2004; Liu and Journel, 2004), or using Markov chains (i.e., auto/cross transition probabilities) to incorporate interclass dependences into simulation (Elfeki and Dekking, 2001; Li et al., 2004; Zhang and Li, 2005), all had the same objective—to incorporate as much available spatial variation information into a random field model as possible. Therefore, to rigorously test a method (not an application), it may be helpful to compare simulated realizations with the "real" one so that it can be visually seen that whether realizations mimic the real spatial patterns. The apparent similarity between conditional realizations from Markov chain meth-

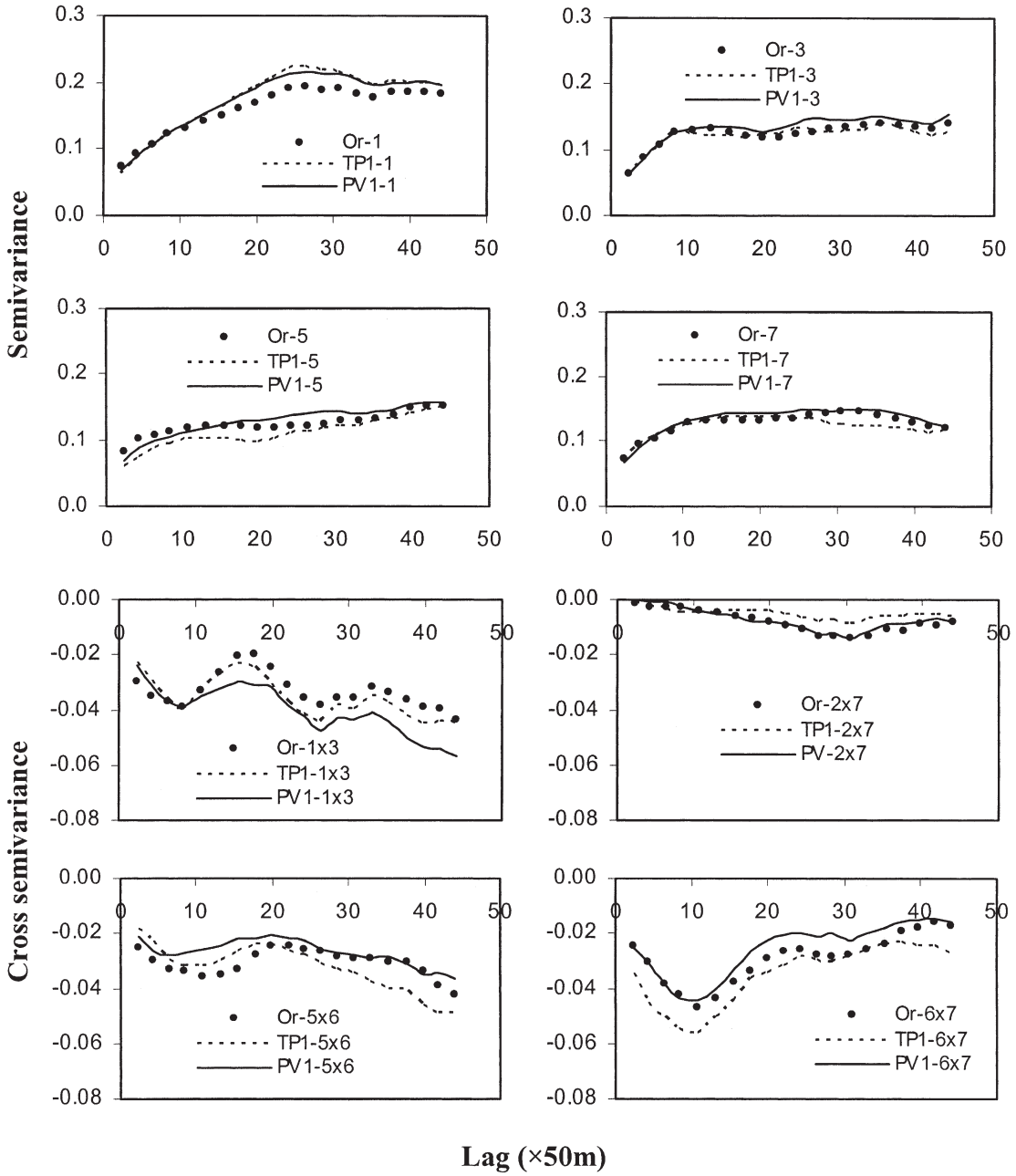

Fig. 6. Indicator (direct and cross) variograms of soil classes in the original soil map and realization maps. The first four are direct variograms. The last four are cross variograms. Or–The original soil map (corresponds to Fig. 5a). PV1–The first PV-realization (corresponds to Fig. 5c). TP1–The first TP-realizations (corresponds to Fig. 5f).

ods and the original image should be mainly attributed to the special characteristics of Markov transition probabilities in accounting for the interdependence of multinomial classes.

Spatial uncertainty (also called locational or positional uncertainty) is a concept relative to spatial certainty represented by observed (or sampled) data. It may not be feasible to analyze spatial uncertainty using geostatistical simulation if we have no observed data for a study area, or if the simulation method cannot condition on the observed data. Unconditionally simulated results using global parameters are completely uncertain; that is, the estimated occurrence probabilities are identical at all locations. Because the available data is limited, we obviously can only guess at classes for unobserved locations. However, the limited observed data can provide us some clue of what they may be (i.e., their occurrence probabilities) and the most likely values (at least for unobserved locations close enough to the observed data that spatial correlations contain a significant signal). An optimally interpolated map only tells us which class has the highest likelihood of occurring at an unobserved location (i.e., the interpolated value). But it is much more useful to know the probabilities of

all classes at unobserved locations. Unlike unconditional methods, the conditional Markov chain simulation method provides this critically important extra dimension.

CONCLUSIONS

A probability vector approach based on the TMC model was presented for spatial uncertainty modeling of soil classes. The method performs direct calculation of JCPD represented by a set of PVs from transition probabilities. Conventionally such PVs were estimated approximately from a large number of realizations. Considering that by following the simulation path of the TMC model the calculation process of PVs needs only one pass, and the time needed is similar to that used for generating one realization using the simulation approach of the TMC model, the probability vector approach is highly efficient in computational time for acquirement of uncertainty information. It avoids the inaccuracy of PVs estimated from a limited number of realizations in spatial uncertainty modeling of multinomial classes.

By visualizing the PVs, spatial uncertainty maps including occurrence probability maps of single classes, single realizations, the maximum occurrence probability map, and the prediction map based on maximum occurrence probabilities can all be acquired. These data may provide important input information for decision-making and risk assessment in natural resource evaluation and environmental conservation. A test study showed that the PVs estimated from simulated realizations gradually approached the calculated PVs with increasing the number of simulated realizations. When the number of realizations for estimating the PVs was small, the estimated PVs were not reliable. Individual realizations could be visualized from the calculated PVs quickly by Monte Carlo sampling. The visualized PV-realizations showed different characteristics from the simulated ones (i.e., TP-realizations). However, indicator (cross) variograms calculated from these two kinds of realizations showed that they were similar and both of them could approximately reproduce the complex spatial dependence relationships in the reference soil map. The PV-realizations also mimicked the reference soil map in spatial patterns. These mean that the visualized realizations from the calculated PVs could effectively reflect the spatial variation of soil classes.

Although the proposed method uses survey line data, it is possible to apply it to other kinds of data. With a more intensively developed software tool and a parameter estimation strategy, point data, line data, patch data, or even mixture of them might be used for conditional simulations. However, for representing the spatial continuity, line data may be more advantageous, and such data are not difficult to acquire in field survey for area class soil mapping. (As we had mentioned, survey line data might be acquired by just recording class boundary changes along a line in field survey as a special kind of purposeive sampling).

We had shown that the special features exhibited by the probability vector approach computed from a TMC were promising. Such an approach is also applicable to other Markov chain conditional simulation methods that used a one-pass simulation algorithm with explicit conditional transition probability expressions. This paper focused on demonstrating the estimation of PVs and their characteristics. Further studies are necessary to address related issues such as parameter estimation from other kinds of data and incorporation of secondary information, and to further the ability of Markov chain approaches to effectively modeling multinomial categorical variables.

ACKNOWLEDGMENTS

We thank Dr. Laosheng Wu, Dr. Kejian Wu, and anonymous reviewers for their insightful comments and suggestions. The support of the Geography Department at UW-Madison and that from the "One Hundred Talents Program" of Chinese Academy of Sciences are greatly appreciated.

REFERENCES

Abend, K., T.J. Harley, and L.N. Kanal. 1965. Classification of binary random patterns. IEEE Trans. Inf. Theory 11:538–544.

Besag, J. 1986. On the statistical analysis of dirty pictures (with discussions). J. Royal Statist. Soc. Ser. B 48:259–302.

Bierkens, M.F.P., and P.A. Burrough. 1993a. The indicator approach to categorical soil data: I. Theory. J. Soil Sci. 44:361–368.

Bierkens, M.F.P., and P.A. Burrough. 1993b. The indicator approach to categorical soil data: II. Application to mapping and land use suitability analysis. J. Soil Sci. 44:369–381.

Caers, J., and T. Zhang. 2004. Multiple-point geostatistics: A quantitative vehicle for integrating geologic analogs into multiple reservoir models. p. 383–394. *In* G.M. Grammer et al (ed.) Integration of outcrop and modern analogs in reservoir modeling. AAPG Memoirs, AAPG, Tulsa, OK.

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

Chiles, J.-P., and P. Delfiner. 1999. Geostatistics—Modeling spatial uncertainty. John Wiley & Sons, New York.

Dubrule, O., and E. Damsleth. 2001. Achievements and challenges in petroleum geostatistics. Petroleum Geosci. 7:S1–S7.

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

Goodchild, M.F., G. Sun, and S. Yang. 1992. Development and test of an error model for categorical data. Inter. J. Geog. Inf. Syst. 6:87–104.

Goovaerts, P. 1996. Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. Math. Geol. 28:909–921.

Goovaerts, P. 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York.

Goovaerts, P. 2002. Geostatistical modeling of spatial uncertainty using *p*-field simulation with conditional probability fields. Int. J. Geog. Inf. Sci. 16:167–178.

Gray, A.J., I.W. Kay, and D.M. Titterington. 1994. An empirical study of the simulation of various models used for images. IEEE Trans. Pattern Analysis and Machine Intelligence 16:507–513.

Guardiano, F., and M. Srivastava. 1993. Multivariate geostatistics: Beyond bivariate moments. Vol. 1, p. 133–144. *In* A. Soares (ed.) Geostatistics Troia'92. Kluwer, Dordrecht, the Netherlands.

Journel, A. 1997. Foreword. p. vii–viii. *In* P. Goovaerts Geostatistics for natural resources evaluation. Oxford Univ. Press, New York.

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.

Kyriakidis, P.C., and J.L. Dungan. 2001. A geostatistical approach for mapping thematic classification accuracy and evaluating the impact

of inaccurate spatial data on ecological model prediction. Environ. Ecol. Stat. 8:311–330.
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.
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, D. Jacques, and J. Feyen. 2001. Effect of spatial variation of textural layers on regional field water balance. Water Resour. Res. 37:1209–1219.
Li, W., C. Zhang, J.E. Burt, A.-X. Zhu, and J. Feyen. 2004. Two-dimensional Markov chain simulation of soil type spatial distribution. Soil Sci. Soc. Am. J. 68:1479–1490.
Liu, Y., and A. Journel. 2004. Improving sequential simulation with a structured path guided by information content. Math. Geol. 36:945–964.
Luo, J. 1996. Transition probability approach to statistical analysis of spatial qualitative variables in geology. p. 281–299. *In* A. Foster and D.F. Marriam (ed.) Geologic modeling and mapping. Plenum Press, New York.
Mark, D.M., and F. Csillag. 1989. The nature of boundaries on the 'area-class' maps. Cartographica 26:65–78.
Norberg, T., L. Rosen, A. Baran, and S. Baran. 2002. On modeling discrete geological structure as Markov random fields. Math. Geol. 34:63–77.
Ortiz, J.M., and C.V. Deutsch. 2004. Indicator simulation accounting for multiple-point statistics. Math. Geol. 36:545–565.
Qian, W., and D.M. Titterington. 1991. Multidimensional Markov Chain Models for image textures. J. Royal Stat. Soc. Ser. B 53: 661–674.
USDA. 1962. Soil Survey—Iowa County Wisconsin. USDA, SCS. U.S. Gov. Print. Office, Washington, DC.
Wu, K., N. Nunan, J.W. Crawford, I.M. Young, and K. Ritz. 2004. An efficient Markov chain model for the simulation of heterogeneous soil structure. Soil Sci. Soc. Am. J. 68:346–351.
Zhang, J., and M. Goodchild. 2002. Uncertainty in geographic information. Taylor & Francis, New York.
Zhang, C., and W. Li. 2004. Predictive area class mapping of multinomial land-cover categories using Markov chains. p. 239–242. *In* Proceedings of the Third International Conference on Geographic Information Science. University of Maryland University College, Adelphi, MD.
Zhang, C., and W. Li. 2005. Markov chain modeling of multinomial land-cover classes. GIScience Remote Sens. 42:1–18.