Transactions in GIS, 2006, 10(4): 651–669

Research Article

A Generalized Markov Chain Approach for Conditional Simulation of Categorical Variables from Grid Samples

Weidong Li
Department of Geography
Kent State University

Chuanrong Zhang
Department of Geography
Kent State University

Abstract
Complex categorical variables are usually classified into many classes with interclass dependencies, which conventional geostatistical methods have difficulties to incorporate. A two-dimensional Markov chain approach has emerged recently for conditional simulation of categorical variables on line data, with the advantage of incorporating interclass dependencies. This paper extends the approach into a generalized method so that conditional simulation can be performed on grid point samples. Distant data interaction is accounted for through the transiogram – a transition probability-based spatial measure. Experimental transiograms are estimated from samples and further fitted by mathematical models, which provide transition probabilities with continuous lags for Markov chain simulation. Simulated results conducted on two datasets of soil types show that when sufficient sample data are conditioned complex patterns of soil types can be captured and simulated realizations can reproduce transiograms with reasonable fluctuations; when data are sparse, a general pattern of major soil types still can be captured, with minor types being relatively underestimated. Therefore, at this stage the method is more suitable for cases where relatively dense samples are available. The computer algorithm can potentially deal with irregular point data with further development.

1 Introduction

Spatial heterogeneity is a typical feature of categorical spatial soil variables, such as soil types. Here categorical spatial soil variables mean mutually exclusive classes, as delineated in area-class maps. Acquiring information about these kinds of variables is a basic requirement for humans to manage natural resources and study environmental problems

Address for correspondence: Chuanrong Zhang, Department of Geography, Kent State University, Kent, OH 44242, USA. Email: zhangc@uww.edu

at regional scales. Due to the effect of spatial heterogeneity of categorical soil variables on ecological and hydrological processes, their quantification is also necessary to assess effectively these processes. Note that the term *class* is used below to generally represent a category.

Soil area-class maps are normally interpreted from field survey data. Spatial uncertainty inevitably arises in the processes of data interpretation and boundary delineation because of the scarcity of observed data, as discussed by Mark and Csillag (1989), Goodchild et al. (1992), and Shi et al. (1999), for example. Conditional (to observed data) simulation using random field models has been recognized in the geosciences as a feasible approach, not only for spatial distribution prediction but also for spatial uncertainty assessment of discrete variables (Goodchild et al. 1992, Deutsch and Journel 1998). However, simulating categorical variables is not an easy task. Variogram-based sequential indicator kriging methods have been used widely as a practical approach for conditional simulation of thresholds (or cutoffs) of continuous variables. With an analogy between a class and a threshold (i.e. coding a class as indicator 1 for occurrence and 0 for non-occurrence at every location), indicator kriging methods have also been applied to categorical geographical variables in some case studies; examples, though remaining scarce, can be found in Bierkens and Burrough (1993), Goovaerts (1996), Miller and Franklin (2002), and Zhang and Goodchild (2002). However, as discussed by many geostatistical experts (e.g. Goovaerts (1996), Deutsch and Journel (1998, p. 86), and Atkinson (2001)), conventional geostatistical methods have difficulties accounting for interclass dependencies (including cross-correlations); therefore, interclass dependencies are normally ignored in conventional geostatistical algorithms. While interdependencies between cutoffs of continuous variables or classes of simple categorical variables may be trivial or not a concern, interdependencies between classes of some complex categorical variables may be prominent (Zhang and Li 2005). For example, some soil types are geographically associated. Occurrence of one class will inevitably affect the occurrence of associated classes in its vicinity. Therefore, incorporating interclass dependencies is crucial to capture effectively the complex pattern of soil classes and because it also makes better use of the spatial variation information conveyed by the sampled data. Without incorporating interclass dependencies in a simulation, this characteristic of categorical soil variables is difficult to reveal in simulated results. Reflecting on geostatistical measures, ignoring interclass dependencies leads to poor reproduction of cross-variograms (see Goovaerts 1996).

Interclass dependencies not only include cross-correlations that can be measured by indicator cross-variograms, but also include the juxtaposition relationships and directional asymmetries of spatial distribution of multinomial classes, which are not effectively captured by indicator cross-variograms because of their intrinsic symmetric property. Markov cross transition probabilities have the capability of representing interdependencies of multinomial classes (Carle and Fogg 1997, Zhang and Li 2005). Markov chains were traditionally used in one-dimensional (1-D) simulations in geology (Harbaugh and Bonham-Carter 1970), ecology (e.g. Balzter 2000), soil science (e.g. Li et al. 1999), and other fields. Multidimensional (multi-D) applications of Markov chains for unconditional simulations can be traced to Krumbein (1968). Recently, multi-D nonlinear Markov chain models for conditional simulation have emerged in the geosciences (Elfeki and Dekking 2001, Li et al. 2004). Existing models, despite having limitations and lacking wide practicality, offer some special features that are desirable for simulating categorical spatial variables with the incorporation of interclass dependencies.

When sufficient survey line data are available, the 2-D Markov chain model described by Li et al. (2004) can effectively reproduce both auto-variograms and cross-variograms, and generate imitative large-scale patterns of soil types and land cover classes (Zhang and Li 2005).

Categorical spatial soil variables such as soil types are usually composed of many nominal classes. Major characteristics of soil types may include:

- **Complex cross-correlations.** As shown in Li et al. (2004), the cross-variograms between soil types are complex and difficult to describe using classical variogram models.
- **Juxtaposition tendencies.** For example, class *A* and class *B* may frequently occur as neighbors, and class *A* and class *C* may never occur as neighbors. To respect these juxtaposition relationships, Markov transition probabilities may be a better choice than covariances because transition probabilities are normally asymmetric.
- **Directional asymmetry.** For example, classes *A*, *B*, and *C* may occur as a sequence of *ABC* along a direction (e.g. west-to-east). This asymmetry can be captured by unidirectional Markov transition probabilities along that direction and thus reflected in realizations (Carle and Fogg 1997, Zhang and Li 2005).
- **Abundance of classes.** For example, there may be dozens of different soil types (e.g. soil series) occurring in a watershed stretching over dozens of square kilometers (USDA 1962). Using iterative simulation methods (e.g. simulated annealing) with consideration of cross-correlations, or solving a large cokriging equation system to deal with a large number of soil classes may be impractical in terms of computation time, numerical stability, and order relations. However, Markov chain models have no obvious computation limitation on the number of involved classes (Li et al. 2004) and also do not suffer order relation problems.
- **Large-scale or long-range patterns.** Soil types normally exhibit large-scale or long-range patterns. As discussed by Gray et al. (1994), Tjelmeland and Besag (1998), and Wu et al. (2004), conventional Markov random field models that use small neighborhoods and cliques have difficulties generating large-scale patterns and accounting for anisotropies (note that a clique means a configuration of adjacent neighbors). The spatial continuity of large-scale patterns is also not well represented in realizations generated by conventional indicator methods, as discussed in Ortiz and Deutsch (2004). However, multi-D Markov chain models are more capable in this instance (see Zhang and Li 2005).

Considering all of these issues, Markov chain methods that are both efficient in computation time and effective in accounting for complex interclass dependencies would be desirable for simulating soil classes. In fact, these are exactly the driving forces of the development of Markov chain-based multi-D conditional simulation models in recent years, despite the fact that accounting for all of these issues in one approach is never easy. Although significant progress has been made, this approach still has limited functions and can only work with line data (borehole logs or survey lines). This is caused mainly by two reasons: the lack of suitable simulation algorithms and the difficulty of estimation of transition probabilities from point samples. Making the multi-D Markov chain approach more versatile, e.g. making it work with point data, needs further exploration. That is the purpose of this study.

In this paper, we extend the work of Li et al. (2004) into a generalized Markov chain model to enable conditional multi-D Markov chain simulation on point data. We use transiograms (i.e. 1-D transition probability diagrams; see Li (2006a)) to estimate

transition probabilities with continuous lags from point data and provide transition probability input to the simulation model. This paper is mainly focused on the following aspects: (1) using transiograms to generalize the 2-D Markov chain model, which previously used transition probability matrices as parameter input and only worked with survey line data; (2) applying transiogram models to Markov chain simulation; and (3) designing a simulation algorithm (i.e. procedure) for conditional Markov chain simulation on grid point data. With further development, this algorithm potentially may accept irregular point data by rotating the so-called cardinal directions. Such a Markov chain-based geostatistical approach is also applicable to categorical (or discrete) variables in other fields such as land-cover classes and provides an alternative to indicator kriging for simulating categorical variables with incorporation of interclass dependencies. Section 2 introduces the generalized Markov chain approach, including transiogram modeling. In section 3, case studies based on two regular point datasets of soil types are provided to demonstrate the potential use in soil type simulation and remaining constraints of the extended approach, and transiogram analyses of simulated realizations are conducted to show whether the approach reproduces the spatial variation structures of sample data described by transiograms. Section 4 recaps the major arguments and notes several ideas in terms of future work.

2 Methods

2.1 The Transiogram

Multi-D transition probabilities that involve more than two adjacent points are difficult to estimate from sample data (e.g. lines or points). Therefore, multi-D transition probabilities that are estimated from original images and used in Markov mesh models for image analysis (see Wu et al. 2004) have not been used in conditional simulation in the geosciences. 1-D one-step transition probabilities can be easily estimated from line data (e.g. borehole logs, survey lines). Based on the first-order Markovian assumption, 1-D multi-step transition probabilities can be calculated from one-step transition probabilities. Both Elfeki and Dekking (2001) and Li et al. (2004) use only one-step transition probabilities (i.e. transition probability matrix) as model input and calculate multi-step transition probabilities from one-step transition probabilities in the simulation process. Carle and Fogg (1997) suggested using the transition rate matrix method (Krumbein 1968) to infer continuous transition probability models from one-step transition probabilities and estimated mean boundary spacings (e.g. mean lengths of hydrofacies) for hydrofacies modeling. These methods are simple but have some shortcomings. One problem is that directly estimating one-step transition probabilities from point data is not feasible. Unlike subsurface survey conducted by drilling boreholes, point sampling is used more widely on the surface (or in the horizontal dimensions). The second problem is that they are based on the first-order Markovian assumption and thus have an intrinsic constraint that assumes boundary spacing (e.g. class polygon lengths, layer thickness) to be exponentially distributed. Boundary spacings of categorical variables may not always be exponentially distributed. For example, the boundary spacings of lithofacies and soil layers usually tend to be lognormally distributed (Krumbein 1968, Li et al. 1997). The third problem is that the non-Markovian property of sample data cannot be reflected on transition probabilities derived from methods based on the first-order Markovian assumption.

Considering the aforementioned three reasons, as long as some samples are available, it is better to estimate continuous (i.e. one-step to *N*-step) transition probabilities directly from data. Estimation of 1-D two-point continuous transition probabilities from observed data with model fitting is feasible. It is just not as well established as the estimation of variograms. The transiogram concept is therefore suggested as a spatial continuity measure to provide the flexibility for transition probability estimation from data and the direct input to multi-D Markov chain models (Li 2006a, b). As a transition probability-based spatial continuity measure, transiograms differ from indicator variograms. A transiogram is defined as a 1-D two-point transition probability diagram over the lag h:

$$p_{ij}(h) = \Pr(Z(x + h) = j \mid Z(x) = i) \tag{1}$$

Here the Markov chain (i.e. the random variable) Z is assumed spatially stationary, that is, $p_{ij}(h)$ is dependent only on the lag h and not dependent on the specific locations x. An auto-transiogram $p_{ii}(h)$ represents the self-dependence (i.e. auto-correlation) of single class i and a cross-transiogram $p_{ij}(h)$ $(i \neq j)$ represents the cross-dependence of class j on class i. Here class i is called a head class and class j is called a tail class. Because of the asymmetry of transition probabilities, cross-transiogram $p_{ij}(h)$ is not equal to $p_{ji}(h)$. A specific study on the properties and features of transiograms is given in Li (2006b).

2.2 The Generalized Markov Chain Model

The 2-D triplex Markov chain (TMC) model described by Li et al. (2004) is based on an idea of conditioning on four nearest known neighbors along four cardinal directions, e.g. north, south, east and west. The coupled Markov chain idea, particularly the idea of conditioning a 1-D Markov chain on a known state ahead, suggested by Elfeki and Dekking (2001), is used to implement the model. Instead of re-addressing the coupled Markov chain model and the triplex Markov chain model here, we refer readers who have the interest of tracing the model background to the published papers.

The TMC model can be simply generalized as the following equation:

$$p[z(u) \mid (m)] = p[z(u) \mid z_{west}, z_{north}, z_{east}, z_{south}] \tag{2}$$

where m represents all of the observed and previously simulated data in the study area; z_{west}, z_{north}, z_{east} and z_{south} represent the four nearest known neighbors in the four cardinal directions of the unknown location $z(u)$ to be estimated (Figure 1). These four nearest known neighbors, one in each direction, may be located distantly from the unknown one to be estimated. It is not necessary that they are adjacent. They may be sampled data or previously simulated points.

Based on Equation (2) and the transiogram concept given in Equation (1), we can rewrite the Markov chain model in Li et al. (2004, p. 1481, Equation [3]) into:

$$\begin{aligned} p_{k|lmqo} &= p[z(u) = k \mid z_{west} = l, z_{north} = m, z_{east} = q, z_{south} = o] \\ &= \frac{p^x_{lk}(h_1) \cdot p^x_{kq}(h_2) \cdot p^y_{mk}(h_3) \cdot p^y_{ko}(h_4)}{\sum_f [p^x_{lf}(h_1) \cdot p^x_{fq}(h_2) \cdot p^y_{mf}(h_3) \cdot p^y_{fo}(h_4)]} \end{aligned} \tag{3}$$

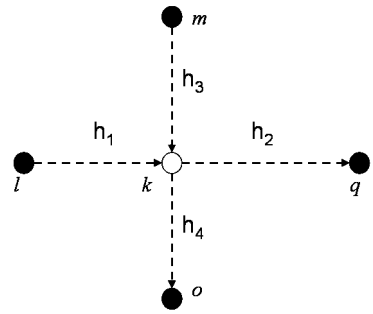

Figure 1 The generalized two-dimensional Markov chain model. Solid cells represent known locations in cardinal directions. The empty cell represents the unknown location to be estimated. h_1, h_2, h_3, and h_4 stand for the distances from the current unknown location to the four nearest known neighbors in the cardinal directions. *k, l, m, q,* and *o* denote states of the five locations (as shown in Equation 3). The dashed arrows represent interactions and transition probability directions

where superscripts x and y represent the axis directions; labels k, l, m, q, o and f all represent the states of the Markov chain at related locations; and h_1, h_2, h_3, and h_4 represent the distances (or numbers of spatial steps) from the four nearest known neighbors to the current unknown location, respectively (see Figure 1). One-step and multi-step transition probabilities are not used here anymore. On the contrary, any $p_{ij}(h)$ represents a continuous transition probability diagram from class i to class j over a lag h, i.e. a transiogram. Here h can be an exact distance (e.g. meters), or for raster data the number of pixels. We will use numbers of pixels as h later in this paper and thus we need not be concerned with the pixel size. For non-raster data, it is better to use exact distance measures.

Equation (3) is the generalized Markov chain model for conditional simulation. It is generalized because: (1) any conditioning point is not required to be adjacent, and (2) it directly uses transiograms as parameters. Such approaches are conventionally called "nonparametric", because they are free of statistical assumptions such as the normal distribution and the first-order Markovian property. Apparently, the model is a nonlinear combination of many transiograms in different directions. Because Equation (3) provides the explicit solution of the conditional probability distribution (CPD) of an unsampled location, the model is efficient in computation cost.

1-D conditional Markov chain models may be useful in some situations such as simulating an outer boundary. A 1-D model conditioned on two nearest known neighbours can be obtained by simplifying Equation (3), i.e. deleting the transition probabilities in one direction, as:

$$p_{k|lq} = p[z(u) = k \mid z_{left} = l, z_{right} = q] = \frac{p_{lk}(h_1) \cdot p_{kq}(h_2)}{\sum_f [p_{lf}(h_1) \cdot p_{fq}(h_2)]} \tag{4}$$

where z_{left} and z_{right} represent the two nearest known neighbors in opposite directions of the unknown z(u) to be estimated. In addition, the simplest 1-D Markov chain model, i.e. the one-step transition probability p_{lk}, may also be used in simulating outer boundaries.

2.3 Transiogram Inference

Transiograms can be estimated from data through two steps: (1) first estimating transition probabilities with different lag h (i.e. estimating experimental transiograms); and (2) then fitting the experimental transiograms with mathematical models and expert knowledge. Thus, transition probabilities at any distances can be acquired from fitted transiogram models.

To guarantee that at any lag h all transition probabilities involving the same head class (i.e. class i in $p_{ij}(h)$) sum to 1, we need always leave one transiogram (e.g. $p_{ik}(h)$) not model-fitted and infer it by:

$$p_{ik}(h) = 1 - \sum_{\substack{j=1 \\ j \neq k}}^{n} p_{ij}(h) \tag{5}$$

where n is the number of classes (Li 2006b). Otherwise, the constraint condition of summing to 1 may be easily violated in the processes of model fitting. To guarantee that $p_{ik}(h)$ be non-negative and well-fitted with the experimental transiogram, the model fitting process of other transiograms may need repetitive tuning.

There are quantitative relations between $p_{ij}(h)$ and proportions of individual classes. The sill of $p_{ij}(h)$ is theoretically equal to the proportion p_j of the tail class j (see Carle and Fogg 1997). The equality between sills of transiograms and proportions of tail classes provides a guide to model fitting, because proportions can be approximately estimated from observed data and expert experience about the study area. The exponential model and the spherical model are often used to model auto-variograms in geostatistics (Deutsch and Journel 1998). Here we also use these two basic models to model transiograms. For modeling auto-transiograms, the exponential model is adapted to:

$$p_{ii}(h) = 1 - (1 - p_i)[1 - \exp(-3h/a_i)] \tag{6}$$

where p_i is the proportion of class i, serving as the sill, and a_i is the practical auto-correlation range (Ritzi 2000, Li 2006b). For modeling cross-transiograms, the exponential model can be adapted to:

$$p_{ij}(h) = p_j[1 - \exp(-3h/a_{ij})] \tag{7}$$

where p_j is the proportion of class j, serving as the sill, and a_{ij} is the practical cross-correlation range (Li 2006b). In Equations (6) and (7), the sill is explicitly set to the proportion of the tail class. Other mathematical models, such as spherical and Gaussian models may follow the same way in defining sills.

If anisotropies and asymmetries of spatial distribution of classes are not considered, transition frequencies in different directions may be pooled together to get omni-directional transiograms. Otherwise, transiograms need be estimated unidirectionally, for example, west-to-east, or south-to-north, which of course needs more data to get reliable experimental transiograms.

2.4 Simulation Algorithm

Monte-Carlo simulation is used to generate realizations. The simulation procedure consists of the following steps (Figure 2):

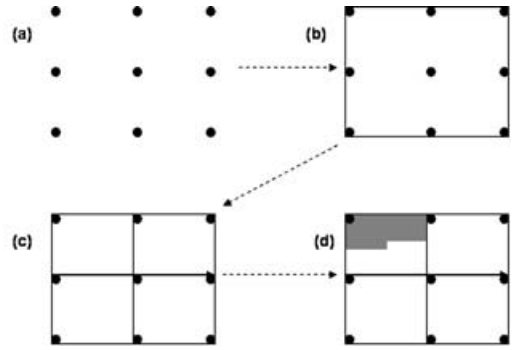

Figure 2 The simulation procedure: (a) sample points, (b) simulating outer boundaries, (c) connecting internal sample points, and (d) filling in each mesh

- **Step 1:** Simulate outer boundaries using 1-D Markov chain models. When there are known data ahead, use Equation (4). When there are no known data ahead, use one-step transition probabilities.
- **Step 2:** Connect all neighboring observed data points that are not connected by simulation in Step 1 using Equation (3), so that simulated lines form a network.
- **Step 3:** Within each mesh formed by simulated lines, perform simulation row by row from top to bottom still using Equation (3).

The above procedure ensures that when estimating one unsampled location in a mesh formed by simulated lines, there are no close known neighbors in off-cardinal directions. Here the so-called cardinal directions in the model refer to just four orthogonal directions and that they may be rotated. So the above algorithm is potentially applicable to randomly distributed point data.

Simulation path is important in 2-D Markov chain simulation because it is related to the directional effect (i.e. simulated patterns are inclined along the simulation direction). When simulating large-scale patterns, directional effect not only occurred in the coupled Markov chain model suggested by Elfeki and Dekking (2001), but also occurred in Markov mesh models that were developed for image analysis, as demonstrated by Gray et al. (1994). Note that Markov mesh models are inappropriate for conditional simulation on sample data. To deal with the directional effect caused by asymmetric neighborhoods, the triplex Markov chain model is composed of two extended coupled Markov chains with an alternate advancing (AA) path (see Li et al. 2004). Along the AA path, the two coupled Markov chains move alternatively in opposite directions row by row (or column by column) (see Figure 3) and thus overcome the directional effect; at the same time, the model imposes balanced influences of known data (simulated or observed) at both (left and right) sides to the estimate of the unsampled location and therefore increases the simulation effectiveness. The AA path essentially also represents a solution to Markov mesh models for overcoming the directional effect. In the above procedure, the AA path is used in both making the network and filling meshes.

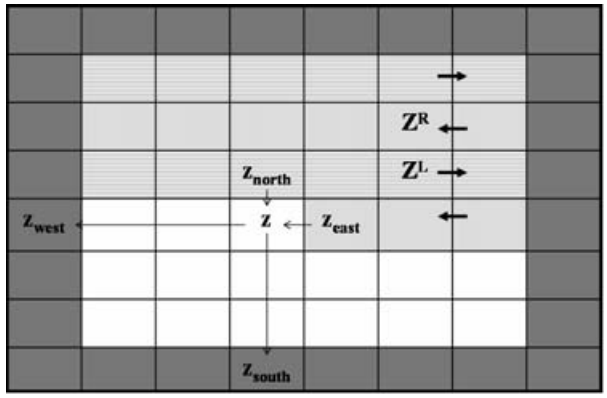

Figure 3 The alternate advancing simulation path: thick arrows represent simulation proceeding directions; and thin arrows represent interactions of the unknown location with its four nearest known neighbors in cardinal directions

3 Simulation Examples and Discussions

3.1 Datasets

We used two small sampled datasets – one dense dataset and one sparse dataset, to conduct simulations for demonstrating the potential use and possible constraints of the method. The dense dataset is composed of 136 points and the sparse one contains only 45 points. Both are regularly distributed in the same rectangular study area, which is located in a river basin of Belgium with an extent of 4×1.7 km. The soil is classified into seven soil types (or classes) by merging similar soil series for clarity of presentation (see Li et al. 2004, p. 1482). Simulations are performed by conditioning on each set of data and using the AA path.

The study area is discretized into an 80×34 grid with a pixel size of 50 m, a coarse spatial resolution, so that each pixel (including observed pixels) can be seen clearly. All of the images were prepared using ESRI's ArcMap so that pixels and polygons are displayed exactly without boundary smoothing.

3.2 Experimental Transiograms and Fitted Models

Experimental transiograms were estimated from the dense dataset. Note that regular data were more efficient for estimating transiograms in the cardinal directions. Considering that the dense dataset (136 regular points) was small for seven soil types, we pooled transition frequencies in the four cardinal directions together to get only one set of experimental transiograms – seven auto-transiograms and 42 cross-transiograms. This means we did not consider anisotropies and directional asymmetry of class sequences. But cross-transiograms are normally asymmetric, i.e. $p_{ij}(h) \neq p_{ji}(h)$. Figure 4 shows the seven experimental transiograms headed by soil type 1 and fitted models. Apparently, experimental transiograms have complex shapes. However, it is clear that these experimental transiograms indeed reflect corresponding proportions of all classes (the sill of each fitted model provided here is set to the proportion of the corresponding

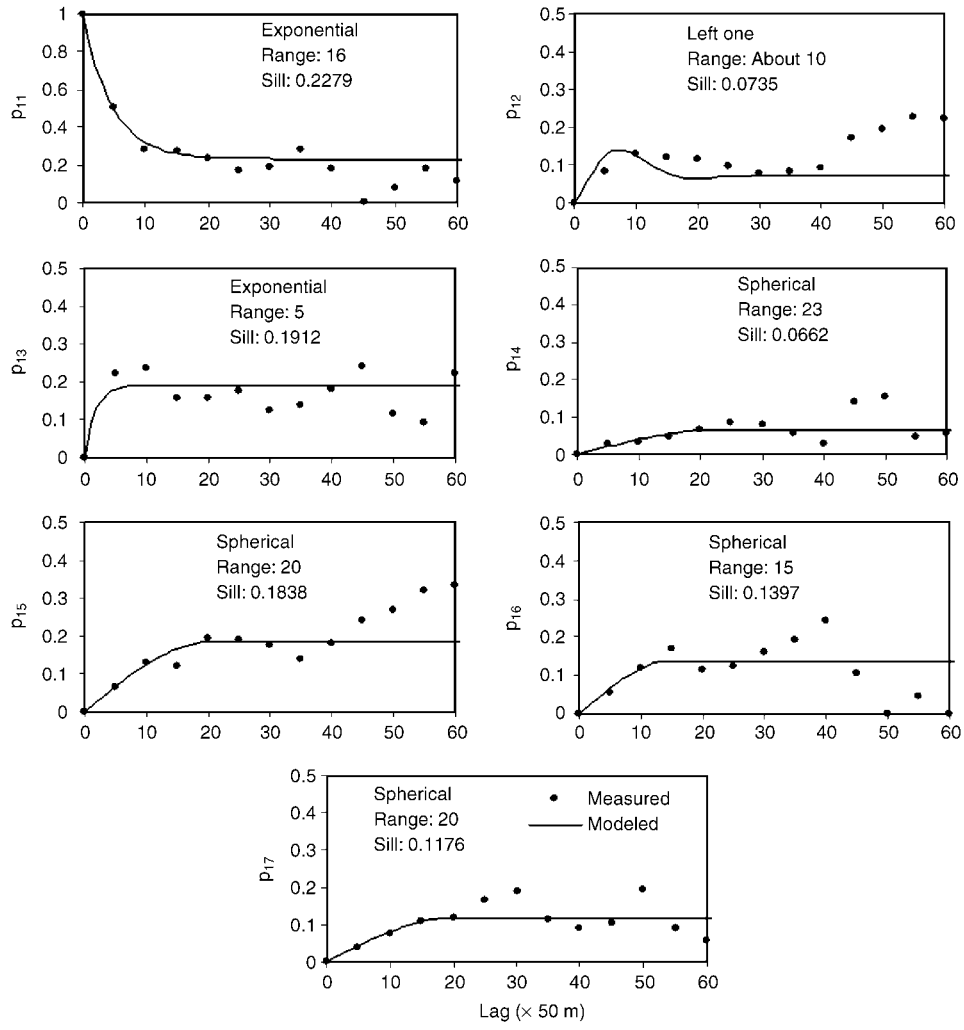

Figure 4 Experimental transiograms headed by soil type 1 and their fitted models, estimated from the dense dataset (i.e. 136 points). Sills are set to the proportions of corresponding tail classes in the dense dataset

tail class in the sampled dataset). We used two basic models – the exponential model and the spherical model to fit all of the experimental transiograms, except for the "left one" in each subset headed by one common class, which was calculated using Equation (5). Correlation ranges and curve shapes of models were approximately interpreted from corresponding experimental transiograms, and sills were all set to the proportions of corresponding soil types in the dense dataset. Apparently, the fitted models only approximately capture some general trends and ignore many details. To capture more details of experimental transiograms, particularly the peaks and troughs, complex models are required, which need further studies. However, overfitting to details of experimental transiograms may not always be necessary or preferable. Except for a large workload,

some details may be just noise resulting from the deficiency of sample data. All 49 fitted transiogram models were used as transition probability inputs to simulations.

The sparse dataset of 45 points was too small to provide reliable experimental transiograms of seven soil types for model fitting. Considering both datasets came from the same area, they should contain similar spatial variation information. Therefore, the auto-transiogram and cross-transiogram models fitted from the dense dataset were used for the conditional simulation on the sparse dataset.

3.3 Simulated Results

One hundred realizations were generated in each of the two simulations and probability maps for each simulation were estimated from those realizations. Figure 5 shows some simulated results conditioned on the dense dataset. Both the simulated realizations (Figures 5c, d) and the prediction map based on maximum occurrence probabilities (Figure 5b) effectively capture all of the seven soil types, which are also reflected in the dataset (Figure 5a). Occurrence probability maps of single soil types (e.g. Figure 5f) indicate that the occurrence location of a soil type is not certain, and the probability values provide information on the trend of occurrence of the soil type in the study area. The maximum occurrence probability map (Figure 5e) clearly demonstrates the predicted transition zones between different soil types. The capture of transition zones in the maximum probability map is remarkable in spatial uncertainty representation. It provides useful information about the spatial uncertainty of soil polygons in the predicted map. These

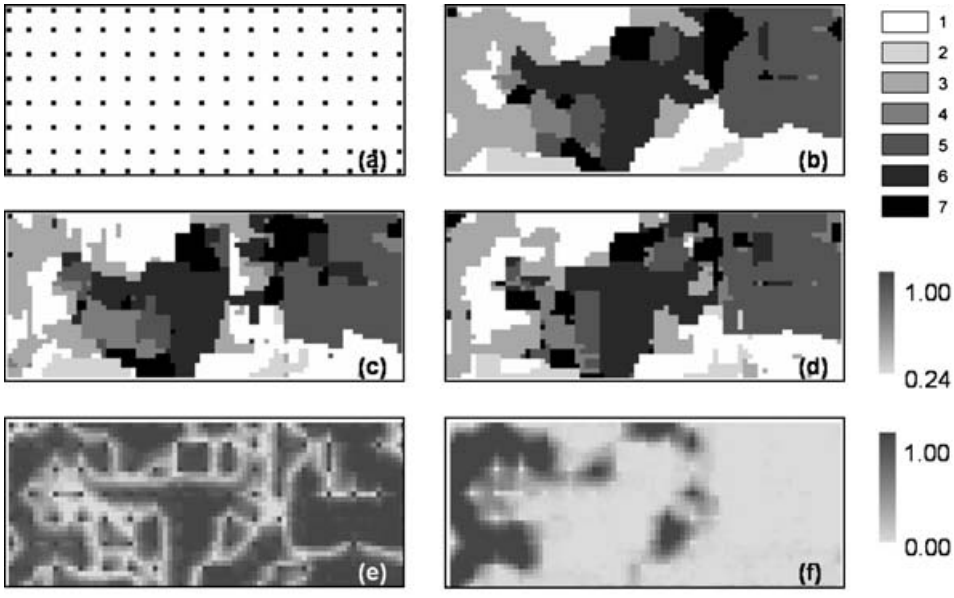

Figure 5 Simulated results based on the dense dataset: (a) the 136 points; (b) the prediction map based on maximum occurrence probabilities; (c) and (d) two realizations; (e) the maximum occurrence probability map estimated from 100 realizations; and (f) the occurrence probability map of soil type 3

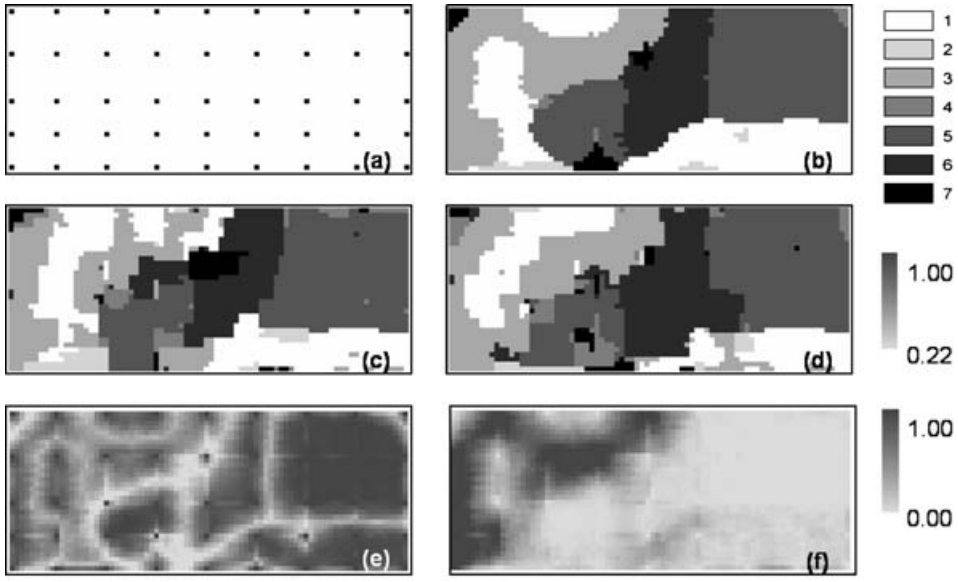

Figure 6 Simulated results based on the sparse dataset: (a) the 45 points; (b) the prediction map based on maximum occurrence probabilities; (c) and (d) two realizations; (e) the maximum occurrence probability map estimated from 100 realizations; and (f) the occurrence probability map of soil type 3

features shown in simulated results should be largely attributed to the incorporation of interclass dependencies.

Figure 6 shows some simulated results based on the sparse dataset. Obviously, only a general pattern is captured in the prediction map and realizations due to the sparseness of the conditioning data and some details displayed in Figure 5 are not shown here. However, the probability maps still provide attractive information – the possible occurrence locations of single classes and the transition zones between major soil types. It is clear that some minor soil types such as types 2, 4 and 7 are relatively underestimated in realizations.

Figure 7 provides the proportion data of the seven soil types, estimated from conditioning datasets and realizations. Simulated realizations conditioned on the dense dataset basically reproduce the proportions of all seven soil types, although the tendency of underestimation of minor types (corresponding to overestimation of some major soil types such as type 5) can still be seen. The problem of underestimation of minor types increases in the simulated realizations conditioned on the sparse dataset. Currently, this is the major tradeoff of the approach to its incorporation of interclass dependencies and its ability to deal with a large number of classes (though for clarity of presentation the number of classes used here is not large). This issue is not unique to this approach. It also occurs in simulated realizations using the Markov random field method (see Norberg et al. 2002). Further study is necessary to theoretically overcome this problem in the model.

A typical characteristic of the Markov chain approach is that continuous large-scale soil patterns (i.e. soil polygons with abrupt boundaries) are directly generated in

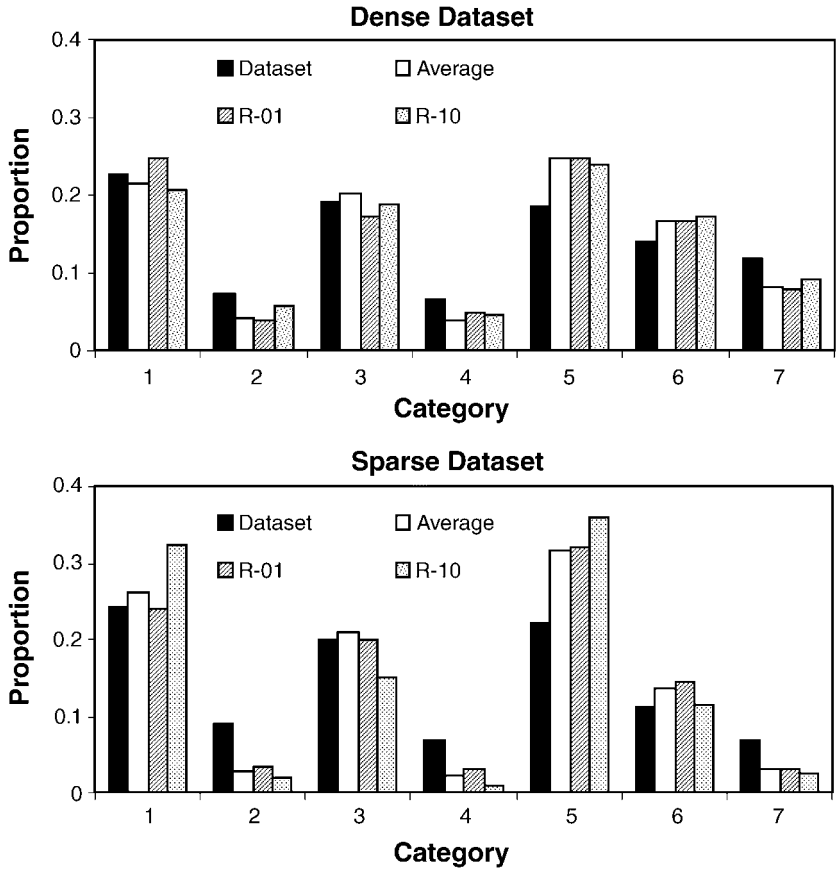

Figure 7 Proportions of different soil types in the datasets and simulated realizations: R-01 means the first realization; R-10 means the tenth realization; and the average values are computed from 100 realizations

simulated realizations. This differs from indicator geostatistical approaches such as the sequential indicator simulation, which normally generate dispersed patterns (see Deutsch and Journel, 1998, p. 307). Recently, based on the triplex Markov chain model, a probability vector approach was suggested by Li et al. (2005). The probability vector approach can also generate dispersed patterns by visualizing probability vectors calculated using the Markov chain model (or estimated from a large number of realizations simulated by the Markov chain model), similar to those generated from the sequential indicator simulation approach. Figure 8 shows some realizations visualized from probability vectors estimated from 100 simulated realizations generated by the generalized Markov chain model presented in this paper. These realizations do not have clear polygons. However, realizations visualized from probability vectors and those directly generated by the Markov chain model have no statistical difference, as shown in Li et al. (2005).

This approach may find its usefulness for predictive soil mapping in flatter areas, where soil-landscape models that infer soil types from environmental factors (e.g. Zhu

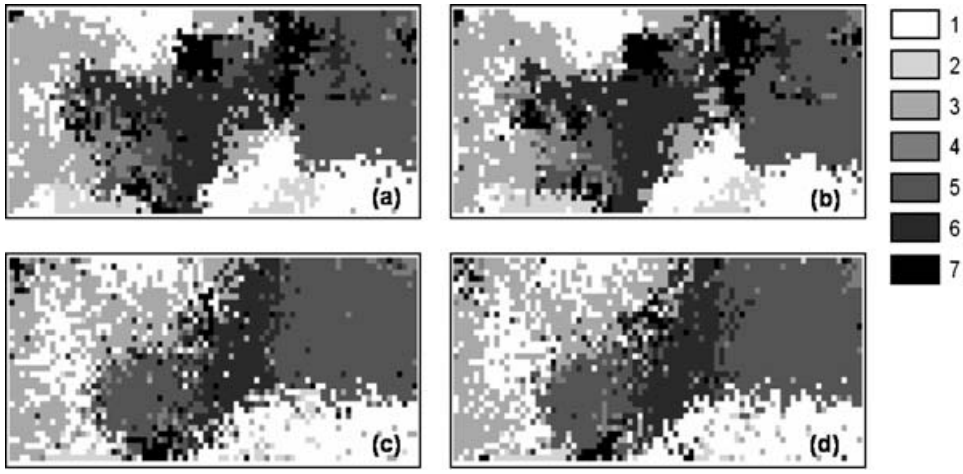

Figure 8 Realizations visualized from probability vectors: (a) and (b) conditioned on the dense dataset (136 points); (c) and (d) conditioned on the sparse dataset (45 points)

et al. 1996) may not work effectively but observing a relatively dense dataset of soil types in the field is usually feasible. The sparse dataset is intentionally used here to demonstrate the constraint of the approach at the current stage. In general, when using this method to simulate the distribution of soil types, users should anticipate the data density issue and try to observe more data, because more data also means more details are captured in simulated realizations. In addition, post-processing methods such as simulated annealing may be used to improve the reproduction of proportions of classes in single realizations, as demonstrated by Goovaerts (1997, pp. 427–9). Therefore, post-processing may serve as a choice to improve the realizations generated by this Markov chain approach, if a single realization is needed as data input to other studies.

3.4 Transiogram Analysis

Li et al. (2004) and Zhang and Li (2005) have conducted indicator variogram analyses on the simulated realizations of soil types and land cover classes generated by the previous Markov chain model that uses survey line data for simulation. Considering that transiograms have been used in this study, we conduct transiogram analysis to check whether simulated realizations can reproduce the spatial structure of soil types described by transiograms.

Figure 9 shows transiograms headed by soil type 1, estimated from the first 10 realizations conditioned on the dense dataset. Interestingly, these transiograms not only approximately match the input models at short lags, but also have an obvious tendency to follow the shapes of experimental transiograms. This is understandable because the conditioning data are also inputs to the simulation. This also implies that conditioning data plays a significant role in simulations using the Markov chain approach. The 10 simulated transiograms fluctuate around the experimental transiograms. The fluctuations are reasonable as "ergodic fluctuations" (see Deutsch and Journel 1998, pp. 128–32, for further discussion of ergodic fluctuations in kriging simulation). Figure 10 shows

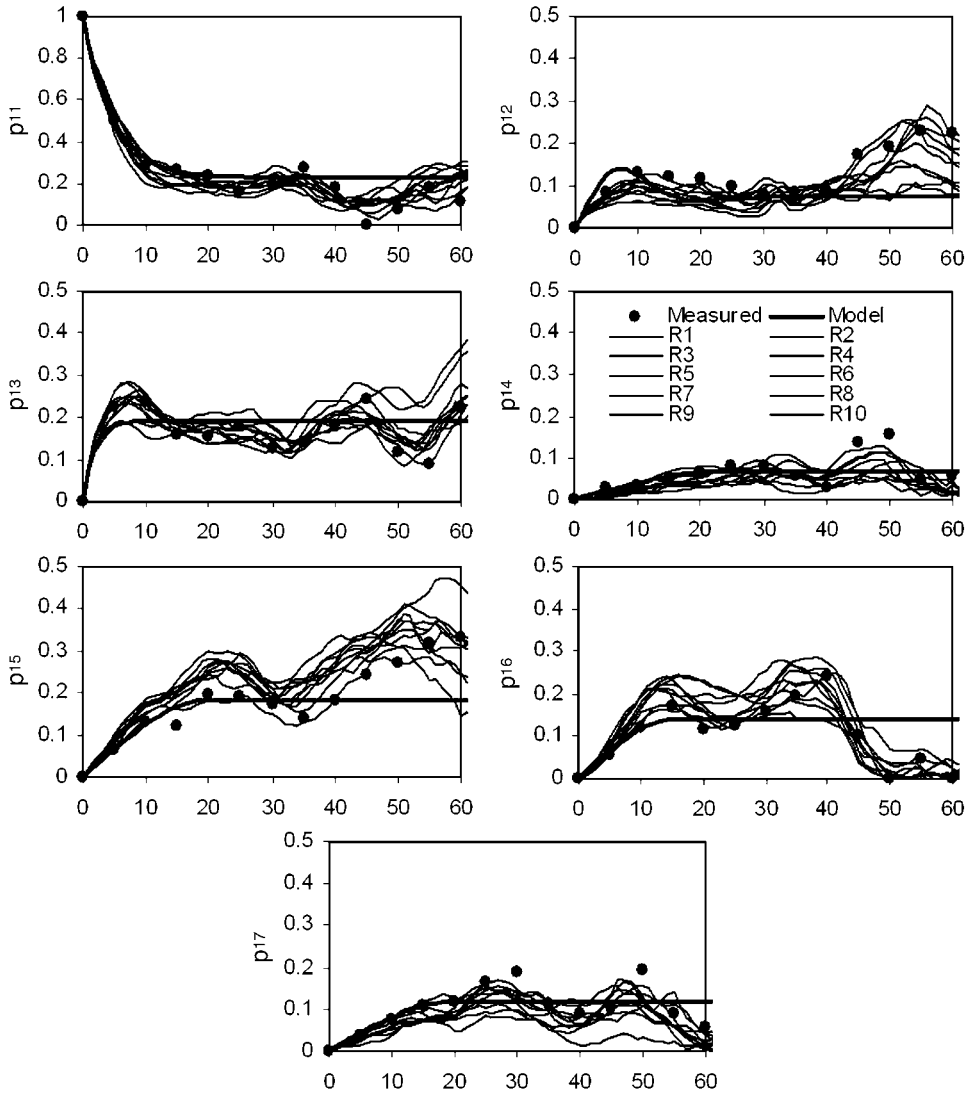

Figure 9 Simulated transiograms headed by soil type 1, estimated from the first 10 simulated realizations conditioned on the dense dataset

simulated transiograms with the head soil type 1, estimated from the first 10 realizations conditioned on the sparse dataset. Similar fluctuations are demonstrated on these simulated transiograms.

Although transiograms estimated from realizations follow the shapes of corresponding experimental transiograms, it is apparent that their sills may be different. In the transiograms based on the sparse dataset (Figure 10), simulated transiograms related to tail classes 2, 4, and 7 have lower sills than the corresponding experimental ones, and the opposite situation occurs on the simulated transiograms tailed by class 5. This is because sills of transiograms are the direct reflection of tail class proportions (Li 2006b).

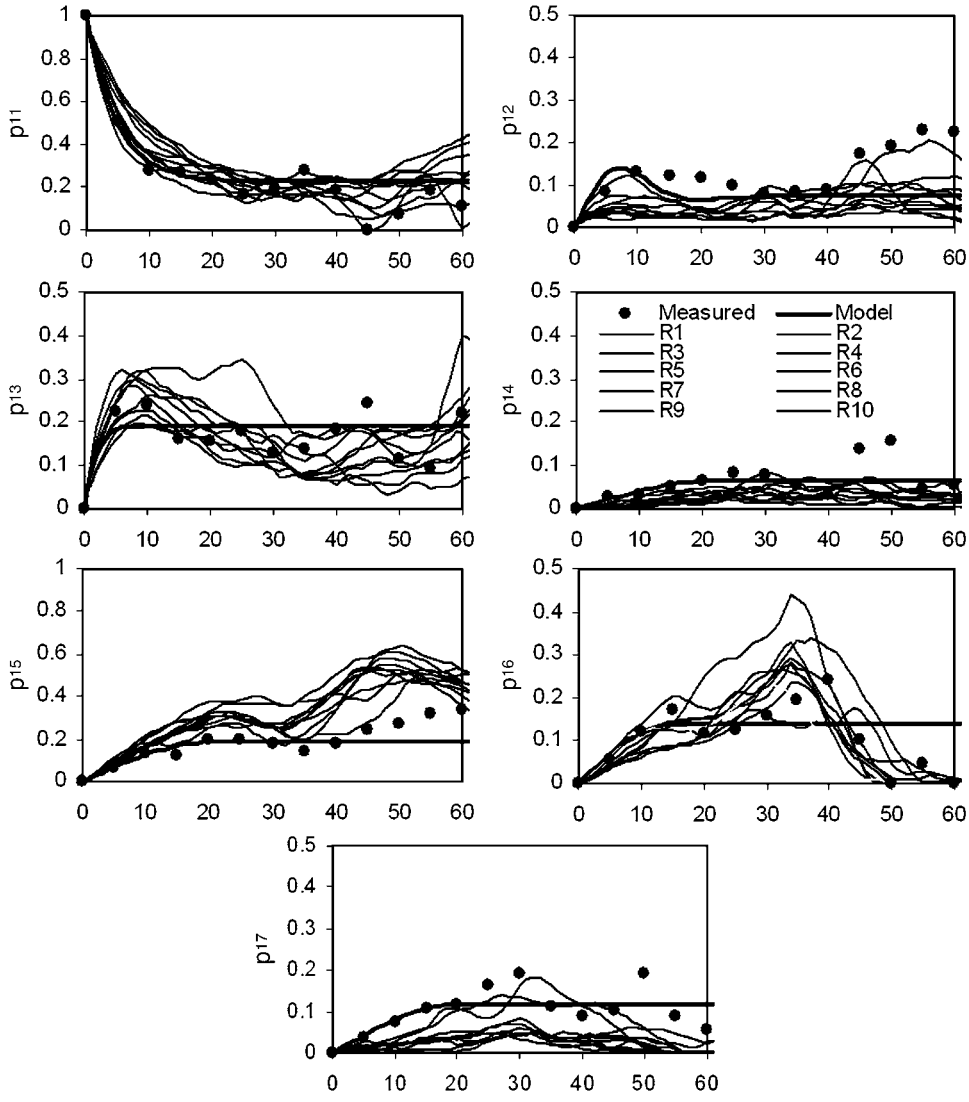

Figure 10 Simulated transiograms headed by soil type 1, estimated from the first 10 simulated realizations conditioned on the sparse dataset

Therefore, whether a class is underestimated or overestimated in a realization and the extent of anomalous estimation can be directly revealed by the sills of simulated transiograms tailed by that class. This direct relationship between sills and tail class proportions is unique for transiograms (Li 2006b). It seems that this mismatch on sills does not obviously affect other features of simulated transiograms, such as correlation ranges and general curve shapes. Of course, except for sills, detailed differences exist between the simulated transiograms shown in Figures 9 and 10 because they were conditioned on different datasets.

In general, the simulation examples show that the generalized two-dimensional Markov chain approach introduced in this paper has the capability of simulating

complex patterns of categorical soil variables, while excessive sparseness of conditioning data impacts the reproduction of correct proportions of minor classes. For specific application cases, if important minor classes are involved and sparseness of data is inevitable, post-processing techniques may be necessary to correct the proportions of classes.

4 Conclusions

A generalized 2-D Markov chain approach is presented for spatial distribution prediction and spatial uncertainty analysis of categorical soil variables. The generalized Markov chain model extends the Markov chain model described in Li et al. (2004). Through applying transiograms to the model and a designed simulation procedure, conditional simulations can be performed on grid point data, which expands the application scope of the approach. The application of transiograms with model fitting provides a versatile approach for estimating transition probabilities with continuous lags from a variety of data types and potentially capturing complex spatial variation of multinomial classes. Simple simulation examples demonstrate the potential applications of this approach to complex categorical soil variables.

Simulated results conditioned on two point datasets demonstrate that the method captures effectively the complex spatial patterns of seven soil types, but when data are too sparse minor soil types are obviously under-represented in realizations. Probability maps reveal some interesting information – the transition zones between different soil types. Transiogram analyses are especially interesting: simulated transiograms apparently follow the shapes of experimental transiograms which are estimated from the conditioning dataset, not that of simplified fitting models; and underestimation (or overestimation) of any class in a realization is revealed with the sills of simulated transiograms tailed by that class.

Within the context of indicator kriging, incorporating cross-correlations may not increase the simulation effect substantially and the overhead involved may not be worthwhile (Journel 1983). However, it is known that effectively incorporating interclass dependencies makes better use of the spatial information contained in sampled data and it should be crucial to some complex categorical variables such as soil types that have strong interclass dependencies. The remarkable capture of transition zones between classes in maximum occurrence probability maps and the reproduction of cross-transiograms in simulated realizations using the non-linear multi-D Markov chain approach should be attributed to the incorporation of interclass dependencies in simulation. This was exactly the driving force of the long-term efforts in developing the multi-D Markov chain approach for simulating categorical variables.

In general, the method is efficient and capable in the following aspects:

1. Efficient computation. The CPD function for each unsampled point is explicit. Transition probabilities are directly drawn from transiogram models.
2. No apparent computation limitation on the number of classes in a simulation. This is because increasing the number of classes does not change the CPD function. This is desirable for dealing with a large number of classes that usually occur in soil classification. Of course, a large number of classes means heavy workloads in transiogram preparation (i.e. model fitting).

3. Incorporation of class interdependencies through cross-transiograms. Thus, the method captures more spatial correlation information conveyed by the same dataset than methods that do not incorporate class interdependencies. Valid cross-transiograms can be simply obtained without coregionalization.
4. The nonlinearity of the CPD function. A nonlinear approach may be preferable for dealing with categorical data.
5. No order relation problem with this approach.

Although our current computer algorithm works only with grid point data, potentially it can deal with irregular point data with further development, for example, by rotating the so-called cardinal directions. Future efforts will focus on solving remaining issues in methodology and developing practical software systems for various data types.

Acknowledgements

We thank Dr John P Wilson and the two anonymous reviewers for their insightful comments and suggestions for revision of the manuscript.

References

Atkinson P M 2001 Geographical information science: GeoComputation and nonstationarity. *Progress in Physical Geography* 25: 111–22

Balzter H 2000 Markov chain models for vegetation dynamics. *Ecological Modeling* 126: 139–54

Bierkens M F P and Burrough P A 1993 The indicator approach to categorical soil data: I, Theory. *Journal of Soil Science* 44: 361–8

Carle S F and Fogg G E 1997 Modeling spatial variability with one- and multi-dimensional continuous Markov chains. *Mathematical Geology* 29: 891–918

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

Elfeki A M and Dekking F M 2001 A Markov chain model for subsurface characterization: theory and applications. *Mathematical Geology* 33: 569–89

Goodchild M F, Sun G, and Yang S 1992 Development and test of an error model for categorical data. *International Journal of Geographical Information Systems* 6: 87–104

Goovaerts P 1996 Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. *Mathematical Geology* 28: 909–21

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

Gray A J, Kay I W, and Titterington D M 1994 An empirical study of the simulation of various models used for images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16: 507–13

Harbaugh J W and Bonham-Carter G F 1970 *Computer Simulation in Geology*. New York, Wiley-Interscience

Journel A G 1983 Nonparametric estimation of spatial distributions. *Mathematical Geology* 15: 445–68

Krumbein W C 1968 *FORTRAN IV Computer Program for Simulation of Transgression and Regression With Continuous Time Markov Models*. Lawrence, KS, Kansas Geological Survey Computer Contribution No. 26

Li W 2006a Transiogram: A spatial relationship measure for categorical data. *International Journal of Geographical Information Science* 20: in press

Li W 2006b Transiograms for characterizing spatial variability of soil classes. *Soil Science Society of America Journal* (in revision)

Li W, Li B, Shi Y, and Tang D 1997 Application of the Markov-chain theory to describe spatial distribution of textural layers. *Soil Science* 162: 672–83

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

Li W, Zhang C, Burt J E, and Zhu A X 2005 A Markov chain-based probability vector approach for modeling spatial uncertainty of soil classes. *Soil Science Society of America Journal* 69: 1931–42

Li W, Zhang C, Burt J E, Zhu A X, and Feyen J 2004 Two-dimensional Markov chain simulation of soil type spatial distribution. *Soil Science Society of America Journal* 68: 1479–90

Mark D M and Csillag F 1989 The nature of boundaries on the 'area-class' maps. *Cartographica* 26: 65–78

Miller J and Franklin J 2002 Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* 157: 227–47

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

Ortiz J M and Deutsch C V 2004 Indicator simulation accounting for multiple-point statistics. *Mathematical Geology* 36: 545–65

Ritzi R W 2000 Behavior of indicator variograms and transition probabilities in relation to the variance in lengths of hydrofacies. *Water Resources Research* 36: 3375–81

Shi W Z, Ehlers M, and Tempfli K 1999 Analytical modelling of positional and thematic uncertainties in the integration of remote sensing and geographical information systems. *Transactions in GIS* 3: 119–36

Tjelmeland H and Besag J 1998 Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics* 25: 415–533

USDA 1962 *Soil Survey of Iowa County, Wisconsin*. Washington, DC, United States Department of Agriculture, Soil Conservation Service

Wu K, Nunan N, Crawford J W, Young I M, and Ritz K 2004 An efficient Markov chain model for the simulation of heterogeneous soil structure. *Soil Science Society of America Journal* 68: 346–51

Zhang C and Li W 2005 Markov chain modeling of multinomial land-cover classes. *GIScience and Remote Sensing* 42: 1–18

Zhang J and Goodchild M 2002 *Uncertainty in Geographical Information*. New York, Taylor and Francis

Zhu A X, Band L E, Dutton B, and Nimlos T 1996 Automated soil inference under fuzzy logic. *Ecological Modeling* 90: 123–45