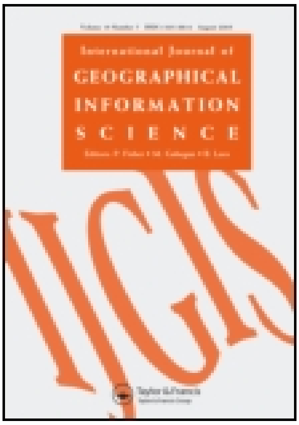

International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Combining spatial transition probabilities for stochastic simulation of categorical fields

Guofeng Cao [a], Phaedon C. Kyriakidis [a] & Michael F. Goodchild [a]

[a] Department of Geography, University of California, Santa Barbara, CA, USA

Available online: 01 Dec 2011

To cite this article: Guofeng Cao, Phaedon C. Kyriakidis & Michael F. Goodchild (2011): Combining spatial transition probabilities for stochastic simulation of categorical fields, International Journal of Geographical Information Science, 25:11, 1773-1791

To link to this article: http://dx.doi.org/10.1080/13658816.2010.528421

International Journal of Geographical Information Science
Vol. 25, No. 11, November 2011, 1773–1791

Combining spatial transition probabilities for stochastic simulation of categorical fields

Guofeng Cao*, Phaedon C. Kyriakidis and Michael F. Goodchild

Department of Geography, University of California, Santa Barbara, CA, USA

(*Received 7 May 2010; final version received 26 September 2010*)

Categorical spatial data, such as land use classes and socioeconomic statistics data, are important data sources in geographical information science (GIS). The investigation of spatial patterns implied in these data can benefit many aspects of GIS research, such as classification of spatial data, spatial data mining, and spatial uncertainty modeling. However, the discrete nature of categorical data limits the application of traditional kriging methods widely used in Gaussian random fields. In this article, we present a new probabilistic method for modeling the posterior probability of class occurrence at any target location in space-given known class labels at source data locations within a neighborhood around that prediction location. In the proposed method, transition probabilities rather than indicator covariances or variograms are used as measures of spatial structure and the conditional or posterior (multi-point) probability is approximated by a weighted combination of preposterior (two-point) transition probabilities, while accounting for spatial interdependencies often ignored by existing approaches. In addition, the connections of the proposed method with probabilistic graphical models (Bayesian networks) and weights of evidence method are also discussed. The advantages of this new proposed approach are analyzed and highlighted through a case study involving the generation of spatial patterns via sequential indicator simulation.

Keywords: categorical data; indicator kriging; conditional independence; Tau model

1. Introduction

Categorical spatial data are very common in geographical information science (GIS) research. They are typically represented by mutually exclusive and collectively exhaustive (MECE) classes and visualized as area-class maps. The class labels of such variables may be distinguished only by their attributes (nominal, e.g., land use and land cover classes) or by ranked orders indicating whether an observation has a higher or lower value than another one (ordinal, e.g., best, better). Many aspects of GIS, such as spatial uncertainty modeling (Goodchild *et al.* 1992) and remote sensing image classification (Boucher and Kyriakidis 2006), require a good understanding of the spatial statistical properties of these data, such as the shape of objects of a certain class and the spatial relationships between classes. A key task in modeling categorical spatial data statistically is to estimate the joint probability mass function (PMF) of a set of georeferenced categorical variables or more intuitively the posterior probability of class occurrence at a target location (where the actual class is unknown) conditioned jointly to all known source data (class labels available at

*Corresponding author. Email: cao@geog.ucsb.edu

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2010.528421
http://www.tandfonline.com

sample locations). In this article, we present such a new statistical method by fusing spatial transition probabilities in a multiplicative way, while accounting for class dependencies in a spatial context, an all important issue that current methods tend to ignore.

In traditional spatial analysis and geostatistics in particular, indicator kriging (IK) (Solow 1986) and indicator cokriging (ICK) (Deutsch and Journel 1998) are the most frequently used methods for estimating the posterior (conditional) probability of class occurrence at any target location. Both IK and ICK rely on two-point statistics, indicator (cross)covariance or (cross)variogram models, for characterizing the spatial class structure implied in categorical spatial data. In the dual kriging form of IK and ICK (Carle and Fogg 1996), the posterior probability can be interpreted as a weighted linear combination of the influence of a particular neighboring class through indicator (cross)covariance or (cross)variogram values. Although covariances and variograms are suitable for Gaussian random fields, the discrete response variables, as well as sharp boundaries and complex spatial patterns found in categorical data, render the interpretation of such covariances and variograms less intuitive, with associated implications for kriging-derived probabilities of class occurrence. Research on spatial interaction in lattice systems abounded in the statistics literature and eventually led to the celebrated Markov random fields (MRFs) theory (Besag 1974; Geman *et al.* 1993). However, the global nature of these models, the computational cost of parameter estimation, and the tendency to ignore directional effects that are common in complex spatial patterns often hinder the application of MRFs in categorical data modeling.

A recent approach addressing this problem involves modeling the conditional multivariate probability by directly scanning carefully designed training images, that is, multipoint geostatistics (MPG) (Guardiano and Srivastava 1993; Strebelle 2002). The training images are deemed representive of the expected spatial patterns in the study area and can be built by experts based on their prior knowledge regarding the phenomenon under study or from pertinent classified remotely sensed imagery. The central idea of the MPG paradigm is to capture multipoint statistics or spatial patterns by scanning training images using a predefined template and to generate pattern realizations using such multi-point statistics in stochastic simulation. Most recently, a pattern-based geostatistical approach was developed based on MPG by extending multipoint statistics to a pattern histogram (Arpat *et al.* 2002; Arpat 2005). MPG-based methods, which can be regarded as nonparametric MRF models, avoid the assumption of an underlying random field and alleviate the computational cost associated with large scanning templates (necessary to capture large-scale features) through a pattern search tree. On the one hand, the lack of specification of an underlying random field improves flexibility and efficiency; on the other hand, it impedes further inference in categorical fields since the underlying model (training image) is hard to parameterize efficiently and may be difficult to acquire.

In another approach, the multivariate posterior probability is expressed as a weighted combination of elementary probability distributions (e.g., bivariate or trivariate probability distributions) conditioned to each individual observed datum or data event if any. Similar to indicator covariograms or variograms used in traditional geostatistics, two-point transition probabilities are often used to quantify spatial continuity or structure in this approach (Carle and Fogg 1996; Li 2006). The critical and most challenging step here is to model the data redundancy between these elementary probabilities or interactions among multiple variables. Most often, this step is either avoided by assuming some form of independence or simplified by only considering dependencies in parts of data rather than all data together. Recently, a spatial Markov chain (SMC) model based on Markov chain random fields (MCRFs) was developed (Li 2007) for combining transition probabilities into posterior probabilities of class occurrence. MCRFs can be regarded as a very special case of MRFs, where only the nearest neighbors in cardinal directions are considered for estimating conditional probabilities of class occurrence under the conditional independence assumption. Most recently, a simplified variant of the Bayesian maximum entropy (BME) approach (Bogaert 2002), namely, multinomial regression (D'Or and Allard 2008), was proposed to address the same problem by combining bivariate joint probabilities, but the assumption of conditional independence was also implicitly invoked. This unrealistic assumption and simplification may be inadequate due to the overwhelming dependencies in a spatial context and the consequences of this simplification should be investigated carefully (Journel 2002). In this article, an approach is proposed to synthesize the elements of the above methods to account for dependencies among given data and thus relax the conditional independence assumption.

Stochastic simulation is a broadly accepted tool for uncertainty propagation and for studying the properties of a statistical model through the generation of alternative realizations (simulated attribute images in a spatial context) from that model. The simulations are termed globally accurate when they reproduce essential statistics, such as the class proportions or spatial structure, of the model under study. However, different simulation algorithms usually impart different global statistics and spatial features on each realization (Deutsch and Journel 1998). In categorical fields, a simulation approach based on a system of simultaneous linear equations was proposed to study the error model of categorical data in spatial databases (Goodchild *et al.* 1992). In this approach, however, a homogeneous (global or spatially invariant) parameter ρ rather than a correlation function of distance is considered. In addition, the method requires the inversion of an $N \times N$ matrix, where N is the number of nodes in the grid to be simulated. Sequential simulation is one of the most popular and most computationally efficient stochastic simulation methods in Gaussian fields due to its conceptual simplicity and straightforwardness (Johnson 1987). In categorical fields, the sequential indicator simulation algorithm that relies on ICK is widely used (Deutsch and Journel 1998). The properties and limitations of this approach have been discussed most recently in Emery (2004). As with the SMC and BME approaches (Li 2007; D'Or and Allard 2008), the sequential simulation paradigm is also used to investigate the spatial patterns implied by the method proposed in this article. In each realization of an area-class map, class labels are simulated at different locations or nodes in a random sequence, conditionally to the previously simulated values and original data (if the latter exist). When sample data are available, one is talking about conditional sequential simulation; otherwise, the term unconditional sequential simulation is used.

In Section 2, the key concepts underpinning current probabilistic data integration methods are reviewed. In Section 3, the proposed approach for fusing spatial transition probabilities into posterior probabilities of class occurrence is presented. In Section 4, the advantages of the proposed approach are highlighted through sequential simulation examples. In Section 5 discussion and conclusions are given.

2. Methods

Consider a categorical random variable (RV) $C(\boldsymbol{x}_0)$, which can take 1 out of K MECE states $c(x_0) \in 1, \ldots, K$, at any arbitrary location with coordinate vector $\boldsymbol{x}_0$. In the absence of any other information, the PMF of RV $C(\boldsymbol{x}_0)$ can be assumed stationary and populated by the K global class proportions $\pi_1, \ldots, \pi_K$. A central task in prediction and simulation of categorical fields is the estimation of the conditional PMF of $C(\boldsymbol{x}_0)$ in the presence of observed or previously simulated class labels available at N locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. For

K classes, one can define the $(NK \times 1)$ indicator vector $\boldsymbol{d} = vec(\boldsymbol{i}_k, k = 1, \ldots, K)$ where $i_k = [\boldsymbol{i}_k(x_n), n = 1, \ldots, N]^T$ is the $(N \times 1)$ indicator vector for the kth class with $i_k(\boldsymbol{x}_n) = 1$ if $c(\boldsymbol{x}_n) = k$, 0 if not; superscript T is used to indicate the transpose of matrix, and $vec()$ denotes the operator that stacks the columns of a matrix one below the other. Note that alternatively $\boldsymbol{d}$ can also be written as $\boldsymbol{c} = [c(x_n), n = 1, \ldots, N]^T$, that is, a $(N \times 1)$ vector of known class labels. Our task can now be re-stated as that of estimating the conditional PMF $P\{C(x_0) = k|\boldsymbol{d}; \pi\}, k = 1, \ldots, K$, where $\pi = [\pi_k, k = 1, \ldots, K]^T$, that is, the $(K \times 1)$ vector of global class proportions. For notational convenience, π is dropped in the following equations.

2.1. Transiogram: spatial structure measure

Transition probability is not a new concept, but it is until recently that transition probability has been regarded as a spatial structure measure and its relationship with the indicator cross-variogram/covariance was discussed (Carle and Fogg 1996, 1997). More specifically, the k to k' class transition probability $\pi_{k'|k}(h)$ for lag $\boldsymbol{h}$ is defined as

$$\pi_{k'|k}(h) = P\{I_{k'}(x + h) = 1|I_k(x) = 1\} \tag{1}$$

where $I_k(\boldsymbol{x})$ is the binary indicator RV at location $\boldsymbol{x}$. The above transition probability is linked to the indicator cross-covariance $\sigma_{kk'}$ $(\boldsymbol{h})$ as

$$\sigma_{kk'}(\boldsymbol{h}) = \pi_k[\pi_{k'|k}(\boldsymbol{h}) - \pi_{k'}] \tag{2}$$

Compared to indicator cross-covariances, transition probabilities have the following advantages as spatial structure measures (Carle and Fogg 1996):

- Easy to integrate with subjective information
- Asymmetry: $\pi_{k'|k}(\boldsymbol{h}) \neq \pi_{k'|k}(-\boldsymbol{h})$
- Satisfy fundamental probability constraints naturally, that is,

$$0 \leq \pi_{k'|k}(\boldsymbol{h}) \leq 1, \forall k, k'$$

and

$$\sum_{k'=1}^{K} \pi_{k'|k}(\boldsymbol{h}) = 1$$

The transiogram is typically defined as a parametric model of transition probabilities as a function of the lag vector $\boldsymbol{h}$, that is, $\pi_{k'|k}(\boldsymbol{h}; \boldsymbol{\theta}_{kk'})$, where $\boldsymbol{\theta}_{kk'}$ is a parameter vector specific to the pair of classes k and k' (Li 2006). Transiograms can be obtained by direct computation (no parametric model involved) from exhaustive sample data or from a probabilistic model of an underlying random field, for example, a truncated multivariate Gaussian field.

Similar with mathematical forms of variograms in Gaussian random fields, an exponential form of transiograms was also discussed in Li (2006), in the absence, however, of a clear underlying theory. Most recently, the bi-probagram, a model of bivariate (two-point) joint probabilities as a function of distance, was also proposed as a spatial structure measure in categorical fields (D'Or and Allard 2008). The connection between the bi-probagram and the transiogram is straightforward (Carle and Fogg 1996), and the former does not offer any additional advantage in spatial analysis over the latter.

Same as indicator cross-covariances, transition probabilities are also two-point statistics, and an efficient method that can integrate them into a multipoint statistics needs to be developed to capture complex spatial structure in categorical fields. Some pertinent concepts and existing approaches are introduced briefly in Section 2.2.

2.2. Probabilistic integration method

Direct modeling of the conditional PMF $P\{C(x_0) = k|d\}$ is difficult as the size of $\boldsymbol{d}$ increases. As mentioned earlier, one approach to address this problem is MPG, which models the joint spatial variability of classes at three or more points simultaneously by directly scanning training images (Strebelle 2002). Alternatively, one could attempt to approximate the complex conditional multipoint PMF by a function of well-defined two-point probabilities. This is the problem addressed in this article, which can be stated as follows: Location $\boldsymbol{x}_0$ with unknown state $c(\boldsymbol{x}_0)$ is surrounded by sample locations, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ with states $c(\boldsymbol{x}_1), \ldots, c(\boldsymbol{x}_N)$, respectively. The individual pair-wise spatial interaction between $\boldsymbol{x}_0$ and each of its neighbors has been individually evaluated through two-point transition probabilities $P\{C(x_0) = k|c(x_n)\}$. Modeling the complex conditional probability of $C(\boldsymbol{x}_0)$ given all the neighboring data by combining these transition probabilities, while accounting for information redundancy, becomes the problem this article addresses. In other words, we want to estimate a function, f, such that

$$\begin{aligned} P\{C(x_0) &= k|c(x_1), \ldots, c(x_N)\} \\ &= f(P\{C(x_0) = k|c(x_1)\}, \ldots, P\{C(x_0) = k|c(x_n)\}) \end{aligned} \tag{3}$$

Figure 1 gives a graphical example, where there are two labels, white and gray, to be decided for location $\boldsymbol{x}_0$, and this assignment depends only on its five nearest neighboring observed states at locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5$. Each arrow indicates a pair-wise spatial interaction, quantified by a transition probability $P\{C(x_0)|C(x_n)\}, n = 1, \ldots, 5$.

One of the simplest functions f is the weighted sum of transition probabilities, that is,

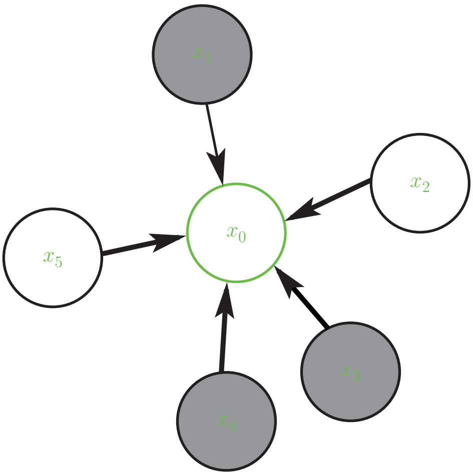

Figure 1. Example of an unknown state (class label) at location $\boldsymbol{x}_0$ depending on its five nearest neighboring states at locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5$.

$$\begin{aligned} P\{C(x_0) &= k|c(x_1), \ldots, c(x_N); \} \\ &= \sum_{n=1}^{N} \lambda_n P\{C(x_0) = k|c(x_n)\} \end{aligned} \tag{4}$$

where λ_n is the weight for transition probability $P\{C(x_0 = k)|c(x_n)\}$ to account for the redundancy information between $\boldsymbol{x}_n$ and its neighbors.

This simple model of Equation (4) can be seen as the dual form of ICK reformulated as function of transiograms instead of indicator (cross) variograms (Carle and Fogg 1997). The problems of the ICK system and its associated spatial structure measures, indicator (cross)variograms, however, have been well documented (Journel and Posa 1990, Cao and Kyriakidis 2008). From another point of view, the model of Equation (4) can also be regarded as an extension of the mixture transition distribution (MTD) model in a spatial context, which was originally proposed for modeling higher order Markov chains in one-dimensional spaces (Raftery 1985, Berchtold and Raftery 2002).

2.3. Conditional independence

An alternative approach to the additive model of Equation (4) is Bayes expansion:

$$\begin{aligned} &P\{C(x_0) = k|c(x_1), \ldots, c(x_N)\} \\ &\quad = \frac{P\{C(x_0) = k, c(x_1), \ldots, c(x_N)\}}{P\{c(x_1), \ldots, c(x_N)\}} \\ &\quad = \frac{P\{C(x_0) = k\}P\{c(x_1)|C(x_0) = k\}, \ldots, P\{c(x_n)|C(x_0) = k, c(x_1), \ldots, c(x_{N-1})\}}{P\{c(x_1), \ldots, c(x_N)\}} \end{aligned} \tag{5}$$

The conditional independence assumption among neighboring data has been adopted extensively in statistics to simplify the computation of the numerator and denominator in Equation (5). $C(\boldsymbol{x}_N)$ is independent of $C(\boldsymbol{x}_1)$, . . ., $C(\boldsymbol{x}_{N-1})$ given $c(\boldsymbol{x}_0)$ if the following condition holds:

$$P\{c(x_N)|C(x_0) = k, \ldots, c(x_{N-1})\} = P\{c(x_n)|C(x_0) = k\} \tag{6}$$

The conditional independence assumption has been widely used in many Bayes-related fields, such as Bayesian networks classifiers (Friedman *et al.* 1997). Practice has shown that this assumption performs quite well, although it is clearly unrealistic. But there should be room for improvement if this assumption could be relaxed somehow, especially in a spatial context where dependencies are complex and common. In the following sections, a new approach for modeling spatial dependencies is proposed after introducing a recently developed method which is based on the above assumption of conditional independence.

2.4. SMC model

Based on the concept of a transiogram, Li (2007) proposed a method to model categorical data using a single Markov Chain (SMC model) moving randomly within a stationary random field, the so-called MCRF, where the unknown state at an arbitrary location depends only on the states at its nearest neighbors (Markovian property). Capitalizing on the notion of a transiogram, the conditional probability of class occurrence in an SMC model, is given as

$$\begin{aligned} P\{C(x_0) = k|c(x_1), \ldots, c(x_N)\}_{\text{SMC}} \\ &= P\{I_k(x_0) = 1|I_{k_L}(x_0^L) = 1, I_{k_U}(x_0^U) = 1, I_{k_R}(x_0^R) = 1, I_{k_B}(x_0^B) = 1\}_{\text{SMC}} \\ &= \frac{\pi_{k|k_L}(h_1^x)\pi_{k_U|k}(h_2^x)\pi_{k|k_R}(h_3^y)\pi_{k_B|k}(h_4^y)}{\sum\limits_l [\pi_{l|k_L}(h_1^x)\pi_{k_U|l}(h_2^x)\pi_{l|k_R}(h_3^y)\pi_{k_B|l}(h_4^y)]} \end{aligned} \tag{7}$$

where x and y represent the axis directions; k_L, k_U, k_R, and k_B all represent the states of the Markov Chain (class indicators) at neighboring locations; x_0^L, x_0^U, x_0^R, and x_0^B, in four cardinal directions. h_1^x, h_2^x, h_3^y, and h_4^y represent the distances between these four nearest neighbors and the target location $\boldsymbol{x}_0$.

A simplified variant of BME, the so-called multinomial regression model, was also recently proposed to address this problem. This model does not assume an underlying random field as SMC does and can be applied to arbitrary neighborhood settings instead of only neighbors along cardinal directions. By the same token, Equation (7) can be generalized to the following equation in likelihood form:

$$P\{C(x_0) = k|c(x_1), \ldots, c(x_N)\} = \frac{\pi(k) \prod\limits_{n=1}^{N} P\{c(x_n)|C(x_0) = k\}}{\sum\limits_{k=1}^{k=K} \pi(k) \prod\limits_{n=1}^{N} P\{c(x_n)|C(x_0) = k\}} \tag{8}$$

Both the SMC model (Equation (7)) and multinomial regression (Equation (8)) provide extremely simple ways for merging preposterior (two-point) transition probabilities of class occurrence into posterior probabilities. The simplicity is due to the assumption of conditional independence, which is difficult to corroborate in real-world situations as discussed before. In Section 3, a recently developed information redundancy model (the Tau model) is introduced and applied for spatial dependence modeling in categorical fields, thus relaxing the stringent conditional independence assumption.

3. A spatial dependence model

The task of accounting for spatial interdependencies when merging two-point transition probabilities can be regarded as similar to that of accounting for information redundancy in information fusion, which is a challenging problem in group decision theory. Bordley (1982) proposed a multiplicative approach for combining expert assessments of an event's probability of occurrence in the context of group probability assessment, based on the reliability of each expert's opinion or on dependencies between experts. This approach has been applied in the classification of remote sensing imagery (Benediktsson and Swain 1992) and was recently developed into a general solution (the so-called Tau model) for the problem of combining prior probabilities accounting for information redundancy between data sources (Journel 2002; Krishnan 2008).

3.1. The Tau model for probabilistic fusion of information

The assumption of permanence of ratios is another way to approximate the conditional probability of Equation (5). To condense notation, we use A and $D_1, \ldots, D_N$ to represent the events in sample spaces of $C(\boldsymbol{x}_0)$ and $C(\boldsymbol{x}_1), \ldots, C(\boldsymbol{x}_N)$, respectively. For two events D_1 and D_2, considering the following logistic-type probability ratios, $r_0 = \frac{1-P(A)}{P(A)}$, $r_1 = \frac{1-P(A|D_1)}{P(A|D_1)}$, $r_2 = \frac{1-P(A|D_2)}{P(A|D_2)}$, and $r = \frac{1-P(A|D_1,D_2)}{P(A|D_1,D_2)}$, the permanence of ratios amounts to assuming

$$\frac{r}{r_1} \approx \frac{r_2}{r_0} \tag{9}$$

The idea behind this assumption is that ratios of information increments are typically more stable than increments themselves. Compared to the assumption of conditional independence, this assumption avoids the calculation of the marginal probability in Bayesian expansion (denominator of Equation (5)). Actually, in practice, the summation in the denominator of Equations (7) and (8) does not necessarily equal the marginal probability. It can be easily demonstrated that Equation (9) implies conditional independence (Equation (6)) but the reverse is not necessarily true.

This approximation actually also assumes a certain form of independence between D_1 and D_2. To relax this assumption, Journel (2002) introduced an exponent factor, τ_n, to Equation (9) to account for information redundancy between D_1 and D_2.

$$\frac{r}{r_1} = \left(\frac{r_2}{r_0}\right)^{\tau(D_1,D_2)} \tag{10}$$

Equation (10) can be generalized to N data events (Journel 2002, Krishnan 2008). Denoting $r_n = \frac{1-P(A|D_n)}{P(A|D_n)}, n = 1, \ldots, N$, and reexpressing r as $r = \frac{1-P(A|D_1,\ldots,D_N)}{P(A|D_1,\ldots,D_N)}$, one gets

$$\frac{r}{r_0} = \prod_{n=1}^{N} \left(\frac{r_n}{r_0}\right)^{\tau_n} \tag{11}$$

and thus

$$P(A|D_1, \ldots, D_N) = \frac{1}{1+r} \in [0, 1] \tag{12}$$

The main problem with this model is the determination of the exponent factor τ_n, which actually quantifies the information redundancy between D_n and D_{n-1} (Krishnan 2008). Recently, Chugunova and Hu (2008) showed that the Tau model with constant weights is inapplicable in some cases and suggested the necessity of inference of τ_n in each case and at each simulation point.

In this article, the following procedure is applied to obtain τ_n. First the nearest neighbor $\boldsymbol{x}_1$ of the target location $\boldsymbol{x}_0$ is selected and we let $\tau_1 = 1$. Then we assume the value $c(\boldsymbol{x}_1)$ of this selected location $\boldsymbol{x}_1$ is unknown and perform ordinary kriging (OK) to estimate it using the remaining neighbors as known data taking the OK weights as τ_n, $n > 1$. Equation (11) can be reformulated as

$$r = r_1 \left(\frac{r_2}{r_0}\right)^{\tau_2} \cdots \left(\frac{r_N}{r_0}\right)^{\tau_N} \tag{13}$$

where $\tau_n, n = 2, \ldots, N$, are the OK weights.

This procedure can be interpreted using consensus theory (Benediktsson and Swain 1992) as follows: First, the nearest neighbor $\boldsymbol{x}_1$ of the unknown event location $\boldsymbol{x}_0$ is selected and its 'opinion' on what the unknown event should be is assumed completely credible. Then the degree of agreement between the remaining $N - 1$ neighbors and the first selected nearest neighbor $\boldsymbol{x}_1$ is quantified. The more the class label (or attribute value in general) at $\boldsymbol{x}_n$ agree with that at $\boldsymbol{x}_1$, the larger the OK weights for $\boldsymbol{x}_n$ will be; this implies more redundant information between states at $\boldsymbol{x}_n$ and $\boldsymbol{x}_1$, and thus the 'opinion' of $\boldsymbol{x}_n$ should be suppressed. In kriging, all those weights depend (through the variogram model) on the distances between the sample data locations. For example, if the distance between $\boldsymbol{x}_n$ and $\boldsymbol{x}_0$ is much larger than the variogram range, the OK weight for $\boldsymbol{x}_n$ is 0, that is, $\tau_n = 0$, and its corresponding component in Equation (11) is $\left(\frac{r_n}{r_0}\right)^{\tau_n} = 1$; this means that the observed state at location $\boldsymbol{x}_n$ has no influence on the unknown state at location $\boldsymbol{x}_0$. On the other hand, if the OK weight $\tau_n = 1$, $\left(\frac{r_n}{r_0}\right)^{\tau_n} = \frac{r_n}{r_0}$, which means the 'opinion' of $\boldsymbol{x}_n$ is entirely credible. A nonnegativity constraint is imposed on the OK weights (Deutsch 1996) to ensure each $\tau_n \in [0, 1]$ and the sum of these exponents is 1.

3.2. *Connections with other methods*

3.2.1. Weights of evidence

Taking logarithms in Equation (11) gives the log-linear expression:

$$\ln r - \ln r_0 = \sum_{n=1}^{N} \tau_n (\ln r_n - \ln r_0) \tag{14}$$

Denoting $\tau_n(\ln r_n - \ln r_0)$ as w_n, this log-linear expression can be reformulated as

$$\ln r - \ln r_0 = \sum_{n=1}^{N} w_n \tag{15}$$

Equation (15) is actually the so-called *weights of evidence* method (Bonham-Carter 1994), which was originally developed for mineral potential assessment and has been applied for combining information from multiple datasets in GIS modeling. It should be stressed that conditional independence is assumed for computing the weights w_n in practice. From this point of view, the Tau model (Equation 11) is a more general weight of evidence model.

3.2.2. Bayesian networks

Bayesian networks, also known as probabilistic graphical models, provide another means for expressing joint probabilities in a set of RVs by exploiting conditional probabilities in a directed acyclic graph (DAG), in which nodes represent random variables and edges represent direct probabilistic dependencies between them. Bayesian networks are closely related with the methods mentioned in this article. For example, one of simplest and most widely used Bayesian networks is the Naive Bayes (NB) network, which is based on conditional independence and is tantamount to the SMC model. In Figure 2, the SMC model is represented using a probabilistic graph. Many approaches have been proposed to improve on NB networks. One of the simple and efficient way is the super-parent Bayesian network (Keogh and Pazzani 1999), which selects one child node acting as common parent (super-parent)

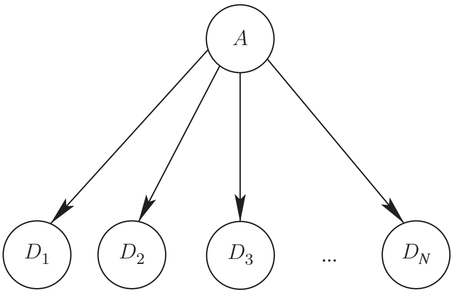

Figure 2. Graphical representation of the SMC model: unknown state A at any location depends on its N nearest neighboring states $D_1, \ldots, D_N$.

for all the remaining children nodes. Applying this paradigm to the computation of the posterior probability, one gets

$$P\{C(x_0) = k|c(x_1), \ldots, c(x_N)\} \\ = \frac{P\{C(x_0) = k\}P\{c(x_i)|C(x_0) = k\} \prod_{n=1}^{N,n\neq i} P\{c(x_n)|C(x_0) = k, c(x_i)\}}{\sum_{k=1}^{k=K} P\{C(x_0) = k\}P\{c(x_i)|C(x_0) = k\} \prod_{n=1}^{N,n\neq i} P\{c(x_n)|C(x_0) = k, c(x_i)\}} \tag{16}$$

where $\boldsymbol{x}_i$ is the selected super-parent from the neighbors of $\boldsymbol{x}_0$ by a certain criterion, for example, mutual information between $\boldsymbol{x}_i$ and $\boldsymbol{x}_0$. The idea behind this augmented Bayesian Network (ABN) is similar to the Tau method proposed in this article, which is represented in Figure 3 using a probabilistic graph. The difference is that Equation (16) uses three-point probabilities to take information dependence into account.

Note that the result of Equation (11) depends on the sequence of expansion. There are $N!$ possible ways to expand the right-hand side of Equation (11). Finding the optimal sequence is an NP-hard (non deterministic polynomial time hard) problem. One solution to this problem was proposed by Friedman *et al.* (1997) based on the concept of *maximum weighted spanning tree* and *mutual information* function. A super-parent Bayesian network

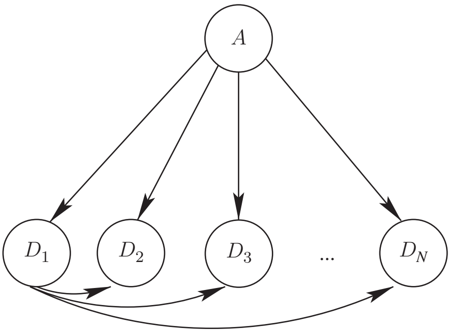

Figure 3. Graphical representation of the Tau model as adopted in this article: all neighboring events depend on each other (are redundant) due to their dependence on the nearest neighbor event D_1.

can be regarded as a simplified version of this approach, but the details of this connection are beyond the scope of this work (Sahami 1996, Friedman *et al*. 1997) .

3.3. Summary

In this section, the Tau model was introduced to combine two-point transition probabilities into multipoint posterior probabilities, while accounting for spatial dependencies through the exponents of the Tau model. An approach based on OK weights for computing the unintuitive exponents was proposed. Its connections with the weights of evidence model and efforts in the Bayesian network field to relax the assumption of conditional independence were discussed. In Section 4, conditional and unconditional simulations of categorical fields are conducted to showcase the advantages of the proposed approach over existing spatial simulation models.

4. Results

To investigate the statistical properties of the proposed model and demonstrate its abilities in reproducing spatial patterns, sequential indicator simulation, one of the most frequently used simulation algorithms in categorical fields, is performed with and without conditioning data. As discussed in the previous sections, a set of auto- and cross-transiogram models is used for quantifying spatial pattern in class labels. The performance of the proposed fusion algorithm is then evaluated in terms of how well the above transiogram models are reproduced from simulated realizations of categorical fields. In other words, the objective here is not to reproduce spatial patterns in real data, but instead to reproduce abstractions (models) of spatial patterns encapsulated in a set of auto- and cross-transiogram models. To this respect, we use a range of parametric transiogram models and evaluate their reproduction from simulated realizations using the proposed transition probability fusion algorithm. In the first case, unconditional simulation is conducted to reveal the true patterns implied in the proposed model. In the second case, the performance of the proposed method to incorporate auxiliary information is investigated using conditional simulation.

In what follows, a reference (true) spatial distribution of class labels is generated through multivariate truncated Gaussian (TG) simulation. The spatial patterns of the TG-generated reference field are then reproduced via unconditional and conditional sequential simulation. The reason for selecting a TG-generated reference field is that, in this case, the conditional probability of $P\{C(x_0) = k|c(x_1), \ldots, c(x_n)\}$ can be obtained analytically through the multivariate Gaussian integral without any approximation. Transiograms for the TG-generated reference are then compared with those obtained via unconditional and conditional simulation using various transition probability fusion models.

We consider $K = 3$ categories with labels $k = 1, 2, 3$ and global proportions $\pi_1 = 0.35$, $\pi_2 = 0.40$, and $\pi_3 = 0.25$; simulation is performed at the grid nodes of a 100×100 regular raster with unit spacing. The variogram model for the underlying Gaussian field, which is truncated to produce the TG-generated reference field is a nugget effect with sill 0.001 plus an isotropic spherical model with range 5 and sill 0.999; this is a favorable situation for sequential simulation since the range is smaller than one-twentieth of the simulation domain. Sets of 50 conditional and unconditional simulations are generated, using a neighborhood containing a maximum of 15 previously simulated class labels and conditioning data, if there are any. Due to space limitations, results from other parameter settings are provided in the following URL: http://www.ncgia.ucsb.edu/programs/ijgis2010

4.1. TG-generated reference categorical field

The procedure of unconditional sequential simulation with conditional probabilities computed directly from a multivariate TG field can be summarized as follows:

- A random simulation path is defined for visiting the nodes of the simulation grid.
- For each node along that path
  - Search neighboring nodes to extract simulated class labels to be used as data and compute the probability of occurrence for each class label using Equation (5) via a multivariate Gaussian integral. If there are no nearby informed nodes, or if this is the first node in the simulation path, use the global proportions as the estimated conditional probabilities. A class label is simulated from the computed conditional probabilities and assigned to this node.
- Proceed to the next node along the random path and repeat the above step until all nodes of the random path are visited once.

The entire procedure is repeated to generate another realization, possibly with a different random path associated with each such realization. Figure 4a displays one such realization generated with TG-based simulation, from which 1000 locations are randomly sampled (Figure 5a) and used as conditioning data for conditional simulation. The ensemble average (simulated mean) of sample transiograms of TG-based simulations, which are taken as references, are displayed in Figures 6 and 7 with green solid lines.

4.2. SMC simulation

Different from the SMC simulation presented in Li (2007), all the neighbors instead of only four neighbors along cardinal directions and a random path rather than the alternating advancing path are adopted in this experiment. Our SMC simulation procedure is same as the TG-based simulation with the only difference that Equation (7) is used to compute the conditional probability for each class label, and the bivariate Gaussian integal is applied to compute transition probabilities. In the conditional simulation scenario, both conditioning data and previously simulated nodes are taken into account. Figures 4b and 5b display realizations of unconditional and conditional SMC simulation, respectively. From these realizations one can detect (at least visually) that the SMC approach tends to generate larger and more symmetric clusters than the TG-based simulation. This could be a consequence of conditional independence, which ignores the information redundancy among neighbors. This is also reflected in the reproduced transiograms of Figures 6 and 7 (blue solid lines), in which the discrepancy between the reproduced transiograms of SMC and the reproduced transiograms of TG-based simulation suggests that the SMC model does not reproduce spatial structure very well.

4.3. Simulation with the Tau model

The same simulation procedure is also applied with the Tau model. The procedure discussed in Sections 4.1 and 4.2 is used to compute the occurrence probability for each class label. We first let all the weights in Equation (11) equal to 1, thus the only difference from the SMC model is that the permanence of ratios is applied in Equation (11). Figures 4c and 5c display the realizations of unconditional and conditional simulation, respectively, where smaller class patches are generated than with the SMC approach; the

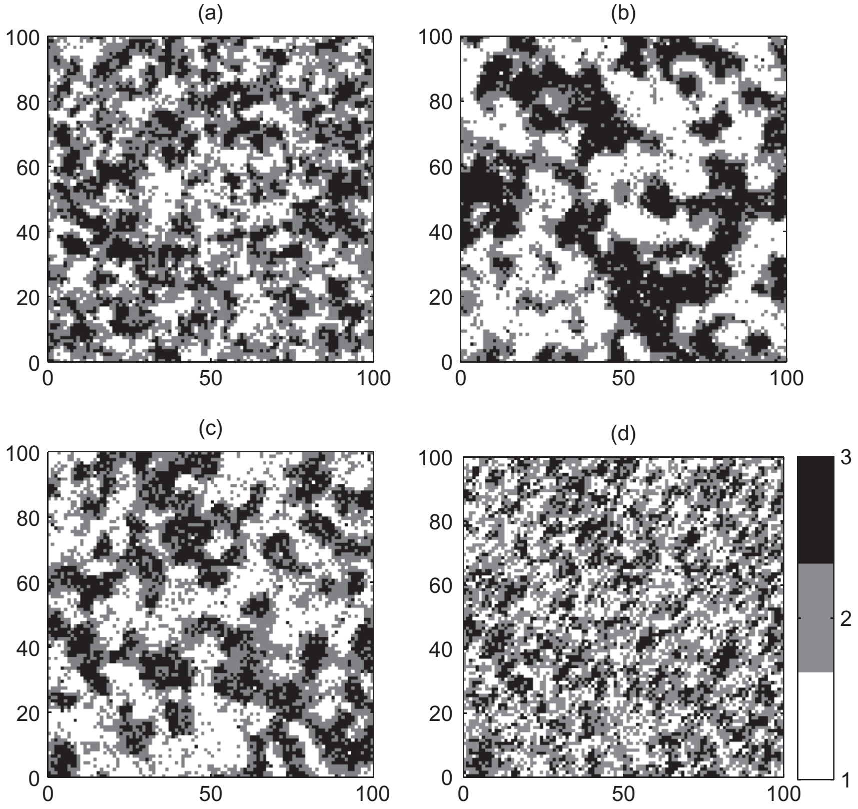

Figure 4. (a) A realization from TG-based simulation (used as a reference map for extracting conditioning data in conditional simulation); (b) a realization of unconditional SMC simulation; (c) a realization of unconditional Tau simulation with constant $\tau = 1$; and (d) a realization of unconditional Tau simulation with OK weights.

Note: TG, truncated Gaussian; SMC, spatial Markov chain; OK, ordinary kriging.

mean (ensemble average) simulated transiograms (cyan lines in Figures 6 and 7) are closer to the reproduced mean transiograms of TG-based simulation. This also suggests that permanence of ratios is a better approximation than conditional independence, at least for this particular example.

Subsequently, the OK weights are used as exponents in the Tau model. Figures 4d and 5d display the realizations of unconditional and conditional simulation. One can appreciate (at least visually) that the regenerated patterns are closer to those of the reference map (Figure 4a) and that the mean reproduced transiograms (red lines in Figures 6 and 7) almost overlap with the reproduced transiogram of TG-based simulation. Note that in the conditional simulation case, the reproduced transiograms under the Tau model are much more stable compared to the transiograms of other methods. This implies that the proposed approach captures and conveys the information in the truncated multivariate Gaussian fields correctly and provides an excellent approximation of the posterior multipoint probabilities. The transiograms of each unconditional and conditional realization are displayed in Figures 8 and 9, respectively. From Figure 8, we can see that the transiograms on the diagonal plots (class 1–1, class 2–2, and class 3–3)

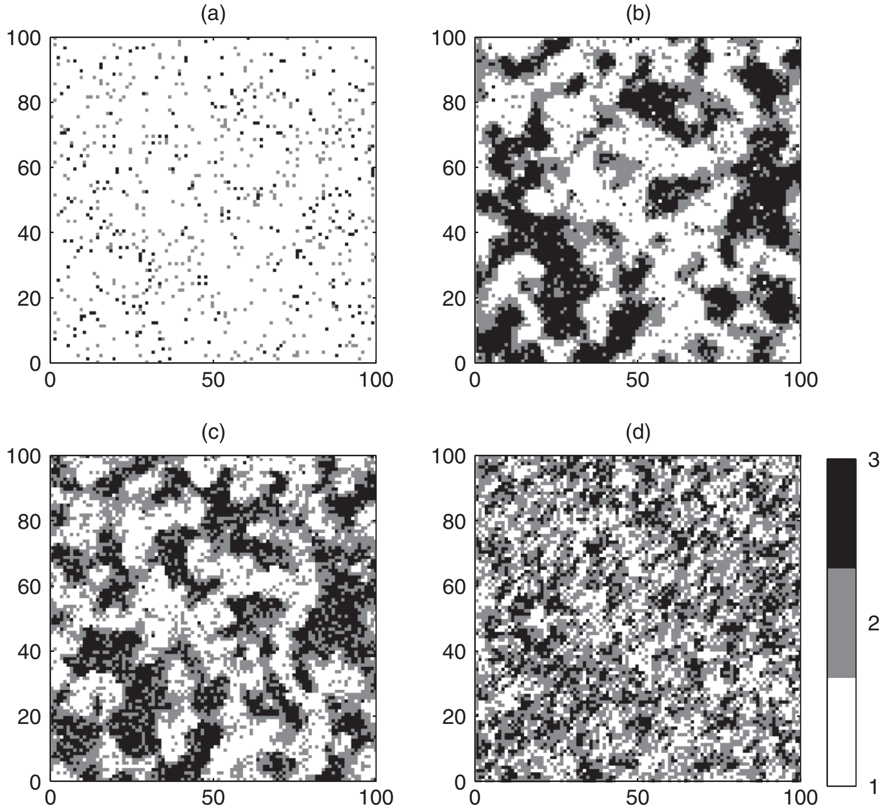

Figure 5. (a) Conditional data, locations of 1000 points sampled from (Figure 4a); (b) a realization of conditional SMC simulation; (c) a realization of conditional Tau simulation with constant $\tau = 1$; (d) a realization of conditional Tau simulation with OK weights.

Note: SMC, spatial Markov chain; OK, ordinary kriging.

are reproduced better than the cross-transiograms (off-diagonal plots) and that the simulated cross-transiograms between classes 1–2, 2–1, 2–3, and 3–2 are reproduced better than those for classes 1–3 and 3–1. This makes sense because in TG field, classes 1 and 3 should not be adjacent, and thus distances between pixels occupied by these two classes tend to be larger than the distances between pixels occupied by classes 1 and 2 (class 1–2 transitions) or by classes 2 and 3 (class 2–3 transitions). Therefore, the accuracy of the reproduced transiograms decreases when the distances increase. This is also true in the conditional simulation case (Figure 9), but the variations of these reproduced transiograms become smaller due to the information contained in the conditioning data.

Goodchild (2008) listed a series of requirements that a stochastic model of uncertainty for categorical data or area-class maps should satisfy. Similar to the SMC model, the simulated realizations of this proposed approach satisfy Requirement 1 (probability map can be generated for each class); Requirement 2 (realizations vary in the counts of faces, edges, and nodes); Requirement 3 (realizations exhibit spatial autocorrelation); Requirement 4 (the effects of generalization, such as changes in the mapping unit can be handled by changes in cell size); and Requirement 5 (results are invariant under changes of cell size). However, because the class label for each point is assigned by class proportions,

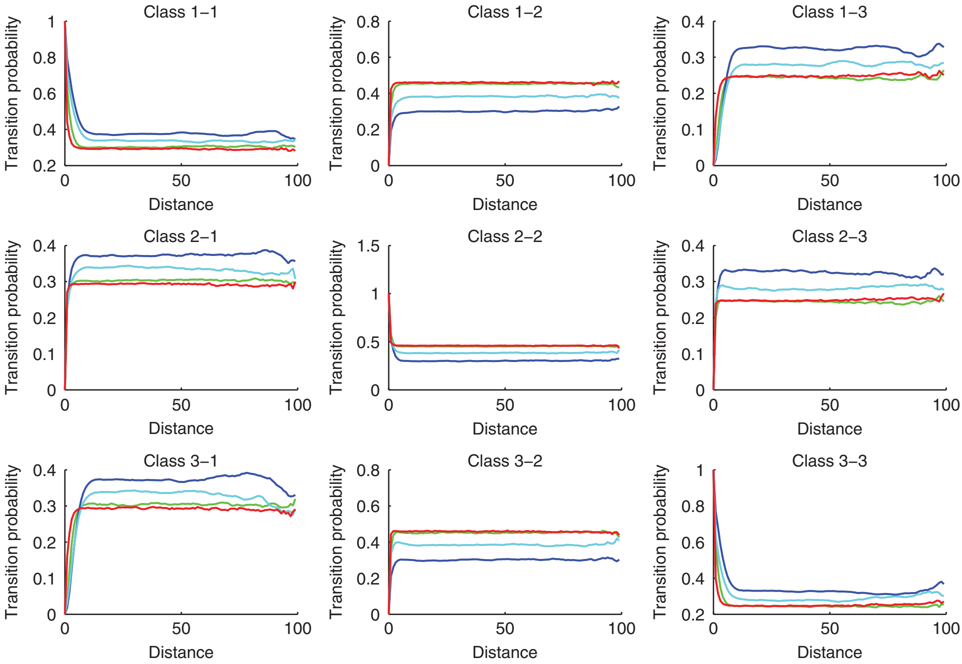

Figure 6. Ensemble averages of simulated auto- and cross-transiograms from different unconditional simulation methods. Green solid lines correspond to the reproduced transiograms of the TG-based model; red solid lines to the Tau model with OK weights; blue solid lines to the SMC model; and cyan solid lines to the Tau model with weights equal to 1.

Note: TG, truncated Gaussian; OK, ordinary kriging; SMC, spatial Markov chain.

the outcomes are not invariant under reordering of classes, and thus the Requirement 6 listed in Goodchild (2008) is not satisfied. Note, however, that this last requirement applies to nonordinal data only.

5. Conclusions and discussion

In this article, a new stochastic simulation method is proposed for modeling and reproducing spatial patterns in categorical fields. The proposed method uses the transiogram as an intuitive measure of spatial structure in categorical data and accounts for information redundancy between neighboring class labels during the fusion of two-point transition probabilities into posterior (multipoint) probabilities of class occurrence. A Matlab toobox based on the proposed method was developed and is available at the following URL: http://www.geog.ucsb.edu/~cao/research.html. The advantages of the proposed approach over the previous ones were showcased via an example case study, where transiograms of a reference image were better reproduced (in both unconditional and conditional simulation scenarios) than with other existing approaches.

It should be noted here that the Tau model is a general paradigm for combining information from diverse sources, in which, weight exponents τ_n are used to account for dependencies among all these sources. The connections of the Tau model with the weights of evidence model and Bayesian networks were also discussed in this article. As a special case of the Tau model (when $\tau_n = 1$), the permanence of ratios is a general form of

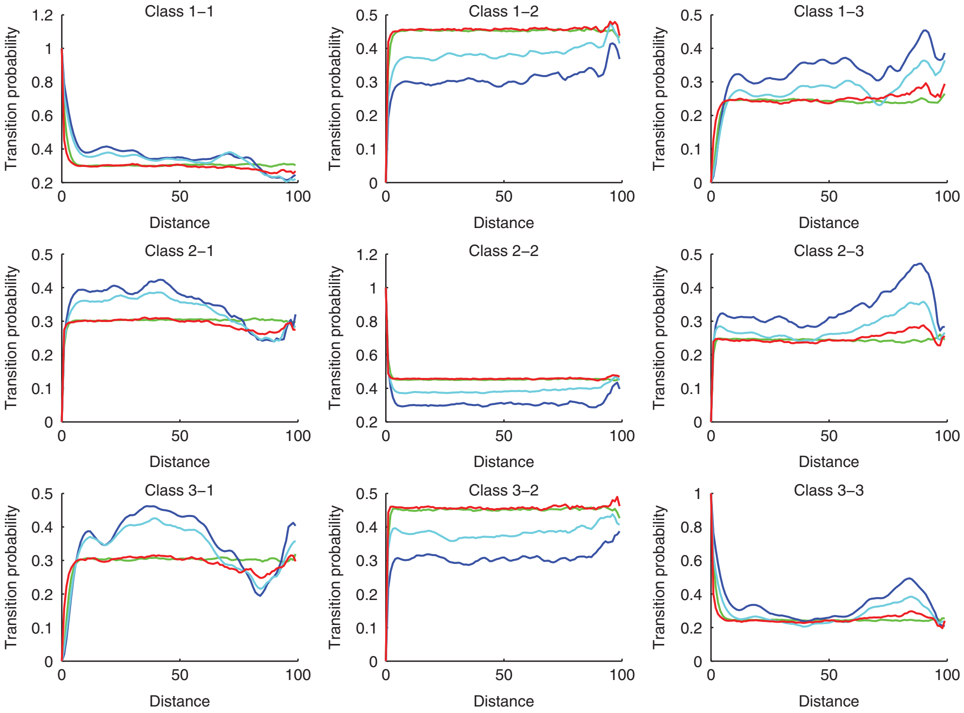

Figure 7. Ensemble averages of simulated auto- and cross-transiograms from different conditional simulation methods. Green solid lines correspond to the reproduced transiogram of TG-based model; red solid lines to the Tau with OK weights; blue solid lines to the SMC model; and cyan solid lines to the Tau with weights equal to 1.

Note: TG, truncated Gaussian; OK, ordinary kriging; SMC, spatial Markov chain.

conditional independence. It avoids the calculation of marginal probabilities and shows a significant performance improvement over conditional independence in approximating a multipoint posterior probability. One problem with the Tau model is the difficulty in obtaining these unintuitive weights. This article proposed an algorithm for computing these exponents using the OK weights, which carry the effects of spatial correlation between neighbors and their spatial configuration. It was demonstrated that the proposed approach performs very well, at least in the case of truncated multivariate Gaussian field. It generated transiograms closest to the target ones (red lines in Figures 6 and 7) and simulated class maps closest to the target reference map (Figures 4d and 5d).

In the proposed method, transiograms, which can be obtained by exhaustive enumeration of sample data or analytical computation from an underlying probabilistic model, were used for quantifying spatial continuity in categorical fields. Note that the sample data could be available at different spatial supports, such as points, polygons, and uniform grids. In addition, expert knowledge regarding the spatial distribution of categorical data in a study region could be incorporated into the task of transiogram modeling (not addressed in this article). As Carle and Fogg (1997) have shown, the proportion of class k' theoretically approaches the sill of the transiogram model $\pi_{k'|k}(h)$, the transition rate (slope) of the auto-transiogram $\pi_{k|k}(h)$ is an increasing function of the mean size of class k objects, and the transition rate of the cross-transiogram $\pi_{k'|k}(h)$ prescribes the juxtapositional tendencies

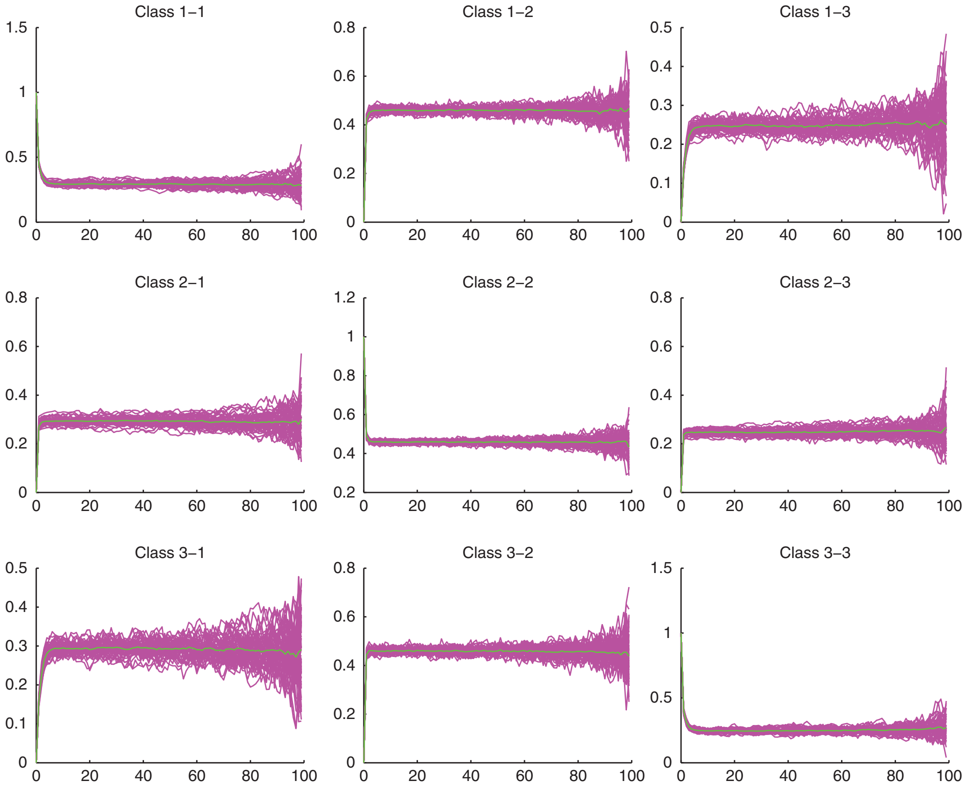

Figure 8. Auto- and cross-transiograms of simulated class labels generated from unconditional simulations with the Tau model with OK weights.

Note: OK, ordinary kriging.

between classes k and k'. Experts can thus adjust the shapes of transiograms according to their prior knowledge about the particular class distributions in a study region. The transiogram is actually a model for pair-wise transition probabilities, which can be generalized to quantify dependencies between different information sources, such as images from different sensors or time instants and prior knowledge from different experts. In the geospatial applications, data usually vary in sources, reliability, spatiotemporal scales, spatial supports, and attribute types (categorical, count, continuous). Based on the previous discussion, the proposed method (Equation (12)) could be extended to spatiotemporal modeling and aggregation of heterogeneous geospatial data, along with the methods to compute the corresponding fusion weights τ_n for those cases.

Finally, this article alluded to the fact that the Tau model can be regarded as an analogue to a Bayesian network, the graphical probabilistic model, which is often adopted to model dependencies among mutiple variables in the machine learning literature. The proposed approach for obtaining Tau weights is similar to an ABN, which is often used for relaxing the conditional independence assumption in Bayesian network theory. It has been pointed out that the sequence of the Bayes expansion will also influence the results and the search for an optimal sequence is a NP-hard problem. This issue is not discussed further in this article but its investigation is definitely warranted in future research.

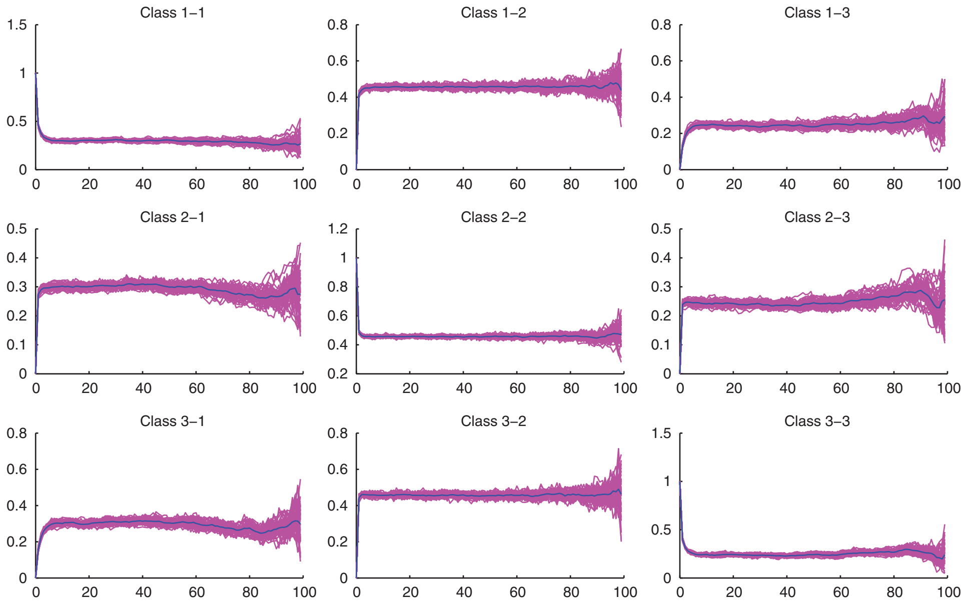

Figure 9. Auto- and cross-transiograms of simulated class labels generated from conditional simulations with the Tau model with OK weights.

Note: OK, ordinary kriging.

Acknowledgment

We gratefully acknowledge the funding provided by the National Geospatial-Intelligence Agency (NGA) to support this research.

References

Arpat, B., 2005. *Sequential simulation with patterns*. Thesis (PhD). Stanford University.

Arpat, B., Caers, J., and Strebelle, S., 2002. Feature-based geostatistics: an application to a submarine channel reservoir. *In*: Proceedings of the SPE Annual Technical Conference and Exhibition: Sam Antonio, Texas, Society of Petroleum Engineers, Paper No. 77426, pp. 1–9.

Benediktsson, J. and Swain, P., 1992. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man and Cybernetics*, 22 (4), 688–704.

Berchtold, A. and Raftery, A., 2002. The mixture transition distribution model for higher-order Markov Chains and non-Gaussian time series. *Statistical Science*, 17, 328–356.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Seris B (Methodological)*, 36 (2), 192–236.

Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment*, 16, 425–448.

Bonham-Carter, G., 1994. *Geographic information systems for geoscientists: modelling with GIS*. New York: Pergamon.

Bordley, R.F., 1982. A multiplicative formula for aggregating probability assessments. *Management Science*, 28, 1137–1148.

Boucher, A. and Kyriakidis, P.C., 2006. Super-resolution land cover mapping with indicator geostatistics. *Remote Sensing of Environment*, 104 (3), 264–282.

Cao, G. and Kyriakidis, P.C., 2008. Combining transition probabilities in the prediction and simulation of categorical fields. *In*: J. Zhang and M. Goodchild, eds. *Proceedings of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences*, 25–27 June 2008, Shanghai, China: World Academic Union.

Carle, S.F. and Fogg, G.E., 1996. Transition probability-based indicator geostatistics. *Mathematical Geology*, 28 (4), 453–476.

Carle, S.F. and Fogg, G.E., 1997. Modeling spatial variability with one and multidimensional continuous-lag Markov chains. *Mathematical Geology*, 29 (7), 891–918.

Chugunova, T. and Hu, L., 2008. An assessment of the Tau model for integrating auxiliary information. *In*: J. Ortiz and X. Emery, eds. *Proceedings of the eighth international geostatistics congress*, Santiago, Chile: Gecamin Ltd.

Deutsch, C., 1996. Correcting negative weights in ordinary kriging. *Computers & Geosciences*, 22, 765–773.

Deutsch, C.V. and Journel, A.G., 1998. *GSLIB: geostatistical software library and user's guide*. 2nd ed. New York: Oxford University Press.

D'Or, D. and Allard, D., 2008. Simulating categorical random fields using the multinomial regression approach. *In*: J. Ortiz and X. Emery, eds. *Proceedings of the eighth international geostatistics congress*, Santiago, Chile: Gecamin Ltd.

Emery, X., 2004. Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, 18 (6), 414–424.

Friedman, N., Geiger, D., and Goldszmidt, M., 1997. Bayesian network classifiers. *Machine Learning*, 29, 131–163.

Geman, S., *et al*., 1993. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, 20 (5), 25–62.

Goodchild, M., 2008. Statistical perspectives on geographic information science. *Geographical Analysis*, 40, 310–325.

Goodchild, M., Guoqing, S., and Shiren, Y., 1992. Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*, 6, 87–104.

Guardiano, F. and Srivastava, R., 1993. Multivariate geostatistics: beyond bivariate mements. *In*: A. Soares, ed. *Geostatistics Troía 1992*. Dordrecht, the Netherlands: Kluwer Academic Press, 133–144.

Johnson, M., 1987. *Multivariate statistical simulation*. New York: John Wiley.

Journel, A., 2002. Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses. *Mathematical Geology*, 34 (5), 573–596.

Journel, A.G. and Posa, D., 1990. Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, 22 (8), 1011–1025.

Keogh, E. and Pazzani, M., 1999. Learning augmented Bayesian classifiers: a comparison of distribution-based and classification-based approaches. *In*: *Proceedings of the 7th international workshop artificial intelligence and statistics*. Ft. Lauderdale, FL. 225–230.

Krishnan, S., 2008. The Tau model for data redundancy and information combination in earth sciences: theory and application. *Mathematical Geosciences*, 40, 705–727.

Li, W., 2006. Transiogram: a spatial relationship measure for categorical data. *International Journal of Geographical Information Science*, 20 (6), 693–699.

Li, W., 2007. Markov chain random fields for estimation of categorical variables. *Mathematical Geology*, 39, 321–335.

Raftery, A., 1985. A model for higher-order Markov chains. *Journal of Royal Statistical Society*, 47, 528–539.

Sahami, M., 1996. Learning limited dependence Bayesian classifiers. *In*: *Proceedings of the second international conference knowledge discovery and data mining (KDD)*. Portland, OR: AAAI Press, 335–338.

Solow, A.R., 1986. Mapping by simple indicator kriging. *Mathematical Geology*, 18 (3), 335–352.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34 (1), 1–21.