Transactions in GIS, 2009, 13(1): 7–23

Research Article

Discriminant Models of Uncertainty in Nominal Fields

Michael Goodchild
National Center for Geographic Information and Analysis
Department of Geography
University of California, Santa Barbara

Jingxiong Zhang
School of Remote Sensing Information Engineering
Wuhan University

Phaedon Kyriakidis
Department of Geography
University of California, Santa Barbara

Abstract
Despite developments in error modeling in discrete objects and continuous fields, there exist substantial and largely unsolved conceptual problems in the domain of nominal fields. This article explores a novel strategy for uncertainty characterization in spatial categorical information. The proposed strategy is based on discriminant space, which is defined with essential properties or driving processes underlying spatial class occurrences, leading to discriminant models of uncertainty in area classes. This strategy reinforces consistency in categorical mapping by imposing class-specific mean structures that can be regressed against discriminant variables, and facilitates scale-dependent error modeling that can effectively emulate the variation found between observers in terms of classes, boundary positions, numbers of polygons, and boundary network topology. Based on simulated data, comparisons with stochastic simulation based on indicator kriging confirmed the replicability of the discriminant models, which work by determining the mean area classes based on discriminant variables and projecting spatially correlated residuals in discriminant space to uncertainty in area classes.

Address for correspondence: Jingxiong Zhang, School of Remote Sensing Information Engineering and Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 LuoYu Road, Wuhan 430079, China. E-mail: jxzhang@whu.edu.cn

doi: 10.1111/j.1467-9671.2009.01141.x

1 Introduction

Categorical information about land cover, soil, and other properties is important for environmental modeling and various applications. Maps of spatial classes are usually represented as discrete polygons associated with class labels or contiguous groups of class-coded raster cells. This kind of map is known as an area-class map (Mark and Csillag 1989), a term which is used interchangeably with categorical map in this article. Developments in remote sensing and statistics have led to the proliferation of spatial categorical information at a range of scales and over time, with multi-source data being integrated using increasingly sophisticated algorithms (Goodchild 1994, Lambin and Strahler 1994, Richards 1996, Tso and Mather 2001, Foody and Mathur 2004).

Despite progress made in remote sensing for resource inventory, ecosystem modeling, and global change research (Skole and Tucker 1993, DeFries and Townshend 1994, Bruzzone and Serpico 1997, Muchoney and Strahler 2002), information derived from remote sensing often suffers from various uncertainty because many biophysical processes underlying landscape dynamics cannot be remotely monitored with adequate accuracy due to the difficulties of separating one class from another when both show similar spectral signatures and, in a changing environment, discriminating real changes from natural variations such as plant phenology (Lambin and Ehrlich 1996, Coppin et al. 2004). Besides, remote sensor data are subject to measurement error and man-machine limitations. Inaccuracy in remote sensing information will have profound implications on the resultant models concerning landscape and ecosystem processes, as uncertainty enters and creates weak links in the chain of information extraction and knowledge construction. As thematic mapping based on remote sensing has been extended from local, regional, national, to global scale, it is required that error in remotely sensed information and its propagation in derivative products be quantified and handled correctly (Stehman et al. 2003).

Uncertainty in categorical maps can be approached from inaccuracy in position, i.e. area-class boundary placement, and class labeling, respectively (Hunsaker et al. 2001, Foody 2002). Positional errors are handled by identifying and removing sliver polygons, i.e. misfits between versions of polylines or polygon boundaries (Chrisman 1989). Sliver removal is often facilitated by carefully setting the tolerance of positional errors. Errors in classification are analyzed by using confusion or error matrices where classification maps and their reference are compared and the hits and misses between them are cross-tabulated on samples of pixels or parcels. Error matrices provide the basis for computing percent correctly classified pixels located randomly over specific domains, and kappa coefficients of agreement that take account of the so-called chance correctness in labeling pixels (Congalton 1991). Methods for estimating variance of classification accuracy are well established (Stehman et al. 2003), and recent research has explored the use of spatially varying probabilities of misclassification evaluated at individual locations, usually grid cells, to paint spatially varied pictures of uncertainty in mapped classes (Steele et al. 2003).

It is common to apply the law of variance and covariance propagation to quantify errors in derivatives given knowledge of the input variables' variance and covariance and the functions involved. For error modeling with spatial data, however, simple applications of the law of variance and covariance propagation assuming spatial and cross-variable independence will lead to biased quantification of standard errors in derivatives. This is because geo-processing often operates over certain neighborhoods and with multi-source

data, as exemplified by simple summation of grid cells labeled with certain land cover types and differencing of percent tree cover maps for change detection, implying that spatial dependence is often induced in spatial analysis and modeling, which must be accounted for in quantification of uncertainty. Therefore, stochastic simulation utilizing an indicator transform has been widely applied as a non-parametric technique for quantifying spatial uncertainty in categorical information and process models. Unfortunately, indicator stochastic simulation does not produce replicable results for error modeling as stochastic simulation does in the domain of continuous fields (Goodchild 2003). This deficiency originates from a flaw in indicator techniques: realizations are drawn from kriged probability vectors of potentially arbitrary class orders, resulting in non-invariant moments in simulated maps.

A more fundamental issue in thematic mapping of land cover and other nominal fields concerns how spatial classes are defined. As different classifications are performed using different criteria and to serve different purposes, it is rare for classifications derived from different systems, as different generalizations, to be comparable, even if the same classification schemes (e.g. Anderson et al. 1976), the same classifiers, and the same analysts are employed with the same raw data and ancillary information. To achieve greater consensus in thematic mapping and to make the process of categorical mapping more uncertainty-informative, a certain logic must be followed to ensure meaningfulness in the resultant products (Whittaker 1973, Robinove 1981, Laut and Paine 1982). For delineation of vegetation, canopy structures are often used: perennial versus annual biomass (to separate forests and woody-stemmed shrubs from annual crops and grass), leaf longevity (to detect evergreen versus deciduous canopy), and leaf type (needle-leaved, broadleaved, and grass). Soils are classified on the basis of observable or measurable properties affecting or resulting from soil genesis, with underlying processes providing a framework for understanding the soil taxa, as it is believed that different soils originate from different external environmental conditions, primarily of climate and vegetation (Jenny 1941).

Goodchild and Dubuc (1987) proposed a logic for categorical mapping based on what they termed phase space, by which land cover types may be differentiated into zones of unique combinations of precipitation and temperature. Recently, along similar lines, Busby (2002) reported a test using a tool known as BIOCLIM that estimates distributions of species such as temperate rainforest trees and bats based on climate data.

It is important to develop classifications based on simple, observable, unambiguous predictors that are remotely sensible and field-measurable, with the defined classes directly translatable to desirable parameters. In this article, we introduce the concept of a *discriminant space*, as it is in this space defined by the environmental variables, or covariates, that individual locations are mapped into class codes. The discriminant-space model emphasizes the importance of understanding the process that has shaped the landscape, as spatial classes are mathematically related to covariates in this model, which are themselves interacting spatial processes. This is significant as area-class occurrences are spatially and temporally varied and heterogeneous, and a process perspective makes the process of categorical mapping and error analysis more objective and tractable.

The major contribution of this process-based methodology is its capability of decoding the logic of categorical mapping and making the information process more accountable. In process-based models for land classification, the deterministic components

are parameterized by biophysical variables and other environmental factors (both natural and built) which have driven the evolution of landscape, while stochastic error terms refer to residuals that cannot be described neatly by the models (Carre and Girard 2002). The predictive models of classes help to maintain consistency across observers, as the deterministic components are described statistically, perhaps as generalized linear models (Moisen and Edwards 1999, Miller and Franklin 2002, McBratney et al. 2003).

The error components are responsible for the differences between realizations of categorical maps. The discriminant-space-based strategy will meet all the requirements for a model of error in area-class maps, because of its flexibility to emulate the variation found between observers, who will vary classes, boundary positions, and also numbers of polygons and boundary-network topology (Goodchild 2003). In this model, class definition, biophysical parameterization, scaling, and error modeling are carried out in the space spanned by discriminant variables. Geostatistics can be utilized for mapping spatial classes and modeling their uncertainty as they propagate from measurement to categorical information, providing a unified strategy for both quantitative remote sensing and predictive modeling of landscape dynamics (Journel and Huijbregts 1978).

The next section will discuss how area classes are defined and parameterized in the discriminant space of dimension b via analysis of the deterministic components in the discriminant models. Simple thresholding is described for univariate cases, i.e. $b = 1$, while generalized linear models are introduced to handle multivariate cases where $b > 1$, although there is no straightforward extension from $b = 1$ to $b > 1$. Section 3 will discuss both linear and indicator geostatistics for propagating error in measurement to uncertainty in area classes. It will be seen that indicator stochastic simulation suffers from non-replicability in realized categorical maps due to class ordering in probability vectors. This problem is solved by performing stochastic simulation in the discriminant space before labeling a location by referring to the mean class models. Section 4 presents empirical results with simulated data, where non-invariant behaviors of indicator stochastic simulation are demonstrated in contrast to replicability of the discriminant model. This is followed by some concluding remarks, outlining topics for further research, in section 5.

2 Demarcating Area Classes in the Discriminant Space

Suppose a domain is discretized into locations denoted by x, which takes values in a chosen coordinate system that may be 2-dimensional, 3-dimensional, or 4-dimensional if a spatiotemporal frame is assumed. Let $\mathbf{Z}(x) = (Z_1(x), \ldots, Z_b(x))$ be a vector field of multi-variables defined on a discriminant space of dimension b with b being a positive integer. At location x, the prevailing class is denoted $C(x)$, which takes values in a set of class codes $\{1, \ldots, K\}$ where K stands for the total number of classes under consideration, or may be a vector of length K, with individual components pertaining to some measures of class likelihood or membership.

Let η denote a classification rule and $\eta(\mathbf{Z}(x))$ a prediction of $C(x)$ obtained by applying the classification rule to a measurement vector $\mathbf{Z}(x)$ at a location (e.g. pixel) x. The process of categorical mapping whereby layers of discriminant variables (measurements) are converted into categorical information can be expressed as:

$$\hat{C}(x) = \eta(\mathbf{Z}(x)) = \arg \max_{k=1,\ldots,k} F_k(\mathbf{Z}(x)) \quad (1)$$

where the F_k are measures of class proximity or similarity to an ideal or stereotype of class k, with the prevailing class at location x taking the maximum value.

The dimensionality b of the discriminant space deserves serious study. It makes sense to start the discussion with $b = 1$ where class labeling can be furnished by thresholding. By Tobler's First Law, a univariate surface of Z will be continuous. Thresholding a continuously varying Z surface gives rise to contours. If area classes are defined by a series of Z intervals, they will take zones between contour lines, implying that area-class maps will resemble contour maps with no three-valent nodes, which are typical for area-class maps and can be constructed in higher-dimensioned discriminant space. In addition, imposition of constraints on the discriminant space is possible when $b > 1$, such as by requiring class definitions to be rectangles or ellipsoids. Therefore, there is no simple extension from $b = 1$ to $b > 1$, as is reflected in the following discussion about class modeling in the discriminant space.

Assume that measurement has been carried out with location x taking a value $z(x)$, with the lower case letter z distinguished from the upper case letter Z and denoting a variable. Given a set of $K - 1$ threshold values z_k, $k = 1, \ldots, K - 1$, it is possible to label the class type at x by simply finding the appropriate interval $z(x)$ falling in:

$$\hat{C}(x) = \begin{cases} 1 & \text{if } z(x) \in [z_{\min}, z_1) \\ 2 & \text{if } z(x) \in [z_1, z_2) \\ \vdots & \\ K & \text{if } z(x) \in [z_{K-1}, z_{\max}] \end{cases} \tag{2}$$

Parallelepiped classifiers can be considered as extensions of thresholding of univariate quantities. For the example of using temperature and precipitation to map land cover, one might be able to define classes as sets of rectangles in the discriminant space.

Thematic mapping can also be considered as regression analysis where response variables, e.g. land cover, are statistically associated with certain combinations of explanatory or predicator variables, e.g. radiance measurements and biophysical variables. In multiple linear regression, the expected value of the response variable is statistically modeled as a linear combination of the explanatory variables. With categorical response variables, the problem of nonlinearity is handled through link functions that transform the expected values of the categorical variable into linear functions of the explanatory variables (McCullagh and Nelder 1989, Gotway and Stroup 1997). Generalized linear models provide a unified framework which can be applied to various linear models. Generalized linear models take the form:

$$g(E(\eta_k(x)) = \mathbf{Z}(x)^T \boldsymbol{\beta}_k, \; k = 1, \ldots, K \tag{3}$$

where the regionalized variable $\eta_k(x)$ may come from any exponential family, and $g(\cdot)$ is the link function, which is often a logit function, $\text{logit}(\pi_k) = \log(\pi_k/(1 - \pi_k))$, for categorical variables. In linear regression, the distribution family is Gaussian and the link function is identity, i.e. $g(E(\eta_k)) = E(\eta_k)$. The mapping logic by Equation (3) is to label location x as belonging to class i, if $g(E(\eta_i(x)) > g(E(\eta_j(x))$ for all $j \neq i$. The response $C(x)$ may be predicted by the inverse of Equation (3):

$$\hat{C}(x) = \underset{k=1.....k}{\arg\max}\, E(\eta_k(x)) = \underset{k=1.....k}{\arg\max}\, g^{-1}(\mathbf{Z}(x)^T \boldsymbol{\beta}_k) \tag{4}$$

In principle, any compositions of discriminant variables fit well in such a framework, which may come from remote sensor measurement, biological surveys, climate and

environment monitoring, where ancillary information concerning categorical distributions in the form of prior probabilities or domain knowledge may also be utilized. A key issue, though, is to make inferences about class definition from maps, i.e. estimating class models from combined uses of measurement and ancillary data, such as elevation, soil, biome, and multitemporal normalized difference vegetation index (NDVI) statistics (Brown et al. 1993, Running et al. 1995, Muchoney and Strahler 2002). It is, therefore, important to explore non-parametric techniques for class modeling in the discriminant space so that area classes may be defined through irregular networks of samples. In addition, reduction of dimensionality of discriminant space may become sensible in the light of adequate accuracy and improved interpretability of the resultant class models.

3 Propagating Error from Measurement to Area Classes

After discussion about deterministic components, we now address the stochastic components (i.e. errors) of discriminant space models. As noted earlier, conventional methods for error handling, i.e. epsilon bands and confusion matrices (Chrisman 1989), do not lead to the successful emulation of variation often encountered in area-class maps in terms of class labeling and boundary geometry/topology (Goodchild 2003). Better models of uncertainty can be built upon the discriminant-space model, with the mean of $\mathbf{Z}$ representing the mean class "signature" and the error $\delta\mathbf{Z}$ reflecting stochastic but often spatially dependent deviates from the mean.

In the following discussion, a particular discriminant variable Z is considered, although generalization to a measurement vector $\mathbf{Z}$ and $b > 1$ should be pursued where necessary. The discriminant-space model proposes that each variable has a true value at every location, and that individual realizations are distortions of this true value using an error model of the type $Z = m_z + \delta Z$. This can be generalized to measurement vector $\mathbf{Z}$:

$$\mathbf{Z}(x) = m_z(x) + \delta\mathbf{Z}(x) \tag{5}$$

which states that measurement $\mathbf{Z}(x)$ is the true or mean vector $m_z(x)$ corrupted by an error vector $\delta\mathbf{Z}(x)$. Outcomes vary from realization to realization, but there is strong persistence between realizations as well, which is implied by the mean $m_z(x)$, the distortion $\delta\mathbf{Z}(x)$, and the class semantics and labeling rule embedded in the discriminant space.

Consider the observed version of localized mean class response $E(\eta_k)$ in Equation (3) at location x, which is regressed against discriminant variables $\mathbf{Z}$. The linear model in Equation (5) may be transformed to one concerning class responses:

$$p'_k(x) = p_k(x) + \varepsilon_k(x),\ k = 1, \ldots, K \tag{6}$$

where $p'_k(x)$ is the observed class response, $p_k(x)$ is the mean class response, and $\varepsilon_k(x)$ is the error term, with subscript k indicating a class code.

In geostatistical analysis of a regionalized variable Z, it is usual to assume a constant local mean and a stationary covariance between places separated by a given distance and direction. Spatial covariance, $\mathrm{cov}_Z(h)$, is defined as the expectation of the product of a variable's deviates from local means for locations separated by a lag h:

$$\mathrm{cov}_z(h) = E((Z(x_{s1}) - m_z(x_{s1}))(Z(x_{s2}) - m_z(x_{s2}))) \quad \text{for } x_{s1} - x_{s2} = h \tag{7}$$

where $Z(x_1)$ and $Z(x_2)$ are the values of variable Z at locations x_1 and x_2 respectively, which are separated by lag h.

Given a set of samples, a kriged estimate for a variable Z at location x may be pursued as the linear combination of data values at sampled locations. With knowledge of $m_Z(x)$, a simple-kriging estimate for an unsampled location x is obtained as:

$$\hat{z}(x) = m_z(x) + \sum_{i=1}^{n} \lambda_i (z(x_i) - m_z(x_i)) \tag{8}$$

where λ_i stands for the weight attached to the sample located at x_i (i =1, 2, . . . , n) within the search neighbourhood. The weights, in turn, are determined by:

$$\sum_{j=1}^{n} \lambda_j \mathrm{cov}_z(x_j, x_i) = \mathrm{cov}_z(x, x_i), \text{ for } i = 1, \ldots, n \tag{9}$$

where the elements $\mathrm{cov}_z(x_j, x_i)$ denote the covariance between sampled locations x_j and x_i, and $\mathrm{cov}_z(x, x_i)$ stands for the covariance between the unsampled location x and location x_i.

Consider discretising the underlying variable of known range $[z_{min}, z_{max}]$ with $K - 1$ increasing cut-off values z_k (k = 1, . . . , $K - 1$), resulting in K intervals $[z_{min}, z_1]$, $[z_1, z_2]$, . . . , $[z_{K-1}, z_{max}]$, as in Equation (2). For each class-defining interval, i.e. a class, it is possible to define an indicator transform $i_k(x)$ for any location x. Indicator variables can also be derived from analysis of the discriminant function in Equations (3) and (4). By analogy to Equations (8) and (9), it is possible to determine the local probability of occurrence of each class k, $\hat{p}_k(x)$, conditioned to existing data, using simple indicator kriging:

$$\hat{p}_k(x) = \hat{i}_k(x) = p_k(x) + \sum_{s=1}^{n} \lambda_S (i_k(x_S) - p_k(x_S)), \tag{10}$$

where $i_k(x_s)$ represents the indicator transform of a sample point x_s (s = 1, 2, . . . , n) to class k, λ_s is the weight associated with the sample point x_s, and $p_k(x)$ is the local prior probability of class k inferred from sample data or, better, obtained from Equation (4).

To model uncertainty in joint computing at multiple locations, stochastic simulation is used (Goodchild et al. 1992). Working with indicators, stochastic simulation proceeds by building a conditional cumulative distribution function (ccdf) as:

$$\mathrm{ccdf}_k(x \mid (\mathrm{data})) = \sum_{k'=1}^{k} p_{k'}(x \mid (\mathrm{data})), k' = 1, \ldots, k \tag{11}$$

where the conditional probabilities $p_{k'}(x \mid (\mathrm{data}))$ are estimated from indicator kriging as defined in Equation (10), and the conditioning data consist of neighbouring original indicator data and previously simulated indicator values, the former being withheld for unconditional simulation. Draw a random number p uniformly distributed between 0 and 1. The category simulated at location x, $c(x)^{(l)}$, is the category that corresponds to the probability interval including p, i.e.

$$c(x)^{(l)} = k \text{ if } p \in (\mathrm{ccdf}_{k-1}(x \mid (\mathrm{data})), \mathrm{ccdf}_k(x \mid (\mathrm{data})) \tag{12}$$

Add that simulated value to the conditioning data set, and continue the previous four steps for the next node along the random path until the problem domain is exhausted.

However, to avoid the non-invariant behaviour of sequential indicator simulation due to class ordering, as can be seen from the algorithm in Equation (12), some special treatment is needed. One can apply conditional simulation to generate equal-probable

realizations of the error vector in Equation (5), which are then added to the mean class response using Equation (6). Lastly, realized area-class maps are derived from applying the rule shown in Equation (1). As the variations in discrimnant space are smooth, only classes adjacent in Z space can be adjacent, overcoming the setback in indicator simulators. Obviously, variogram modeling becomes explicitly and solely associated with the Z space rather than the transformed space of indicators. It should be noted, though, that joint simulation of multiple variables is not trivial (Christakos 1987). This implies that the invariant property of the discriminant space model comes at a price.

Nevertheless, there are cases where class ordering makes sense. For example, in land cover change modeling, it is necessary to have some specific ordering, at least partially, in class labels, e.g. urban invasion of agricultural land, so that sequential class assignment in indicator simulation may be implemented (Goodchild 2003). In such cases, indicator simulation generates realized maps that meet specific requirements. A robust strategy in stochastic simulation concerning area-class maps is, thus, a combination of both indicator and discriminant space models so that the simulators work accurately and flexibly.

4 Results from Simulated Data

To show the drawbacks of indicator stochastic simulation and to demonstrate the effectiveness of the proposed strategy for error modeling in nominal fields, an experiment was implemented with simulated data sets. A univariate discriminant space was considered for its simplicity, as generalization to $b > 1$ is not trivial and needs further research.

First, a mean or true surface of continuous variable Z was simulated using the GSLIB software system, given variogram models (Deutsch and Journel 1998). Gaussian sequential simulation was performed with a variogram model of 1.0*Sph(10), run over a grid system of 100×100, with each cell being a unit square. To make it resemble the typically smooth appearance of a map, the simulated Z field was smoothed using averaging over a 7 by 7 moving window, resulting in a total variance of 0.60. Four threshold values (–0.918, –0.469, –0.063, 0.368) were taken from the smoothed Z's cumulative distribution, giving rise to five classes of approximately equal proportions. The mean map is shown in Figure 1, where the class codes 1 through 5 are depicted in grey scale, lighter to darker. Clearly, the map shown in Figure 1 does not resemble a typical area-class map where three-valent nodes are common, because thresholding of a continuous Z surface ($b = 1$) generates a contour map.

Next, 200 samples were selected from the Z field. These sample points had Z values near the middle of the individual class Z-intervals to ensure their relative class-purity. From these samples, indicator variogram models for five classes were fitted: 0.166 Sph(14.0), 0.141 Sph(8.7), 0.148 Exp(2.0), 0.199 Exp(4.0), and 0.138 Exp(4.5), for classes 1 through 5, respectively. Indicator simulation was applied to generate 100 realizations of area-class maps. Two realizations are shown in Figures 2a and c. These equal-probable maps were post-processed to derive a map of probability vectors and, hence, an area-class map with corresponding probabilities in class labeling, with the former shown in Figure 2e. To see the non-invariant behaviour of indicator-based simulation, class labels 4, 1, 3, 5, and 2 were substituted for class labels "1" through "5" in indicator simulation, and were subsequently reconstituted to the labels as numbered. Again, indicator simulation was run to generate 100 realizations of area-class maps. Two examples are shown in Figures 2b and d. These realizations were post-processed

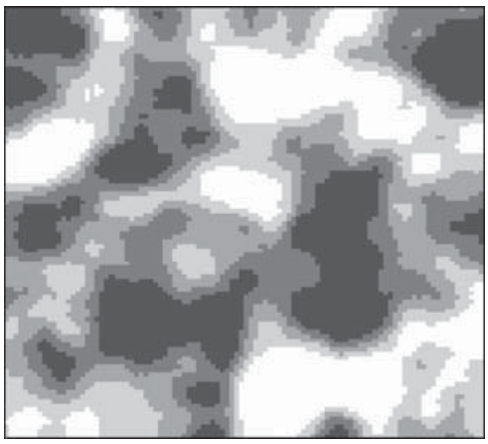

Figure 1 Area-class maps derived from the true *Z* field

to derive a map showing the labels with the maximum per-cell probability of occurrence, as depicted in Figure 2f. There are huge differences between the maps derived from indicator stochastic simulation due to the non-invariant behaviour of indicator-based simulation to the order of class labels in computing. Also notable are the differences between the indicator-simulated maps and the assumed true map, shown in Figure 1.

For the proposed method based on the discriminant space, simulation was carried out to generate 100 realizations of error surfaces to be commensurate with the smoothed *Z* field, applying a spherical variogram model (zero nugget, unit sill, range of 10 grid cells). Conditional data were the set of zeros co-located with the pure samples to ensure that they would not be perturbed. Each simulated error field multiplied with a factor of 0.33 was imposed upon the smoothed *Z* field to emulate a noisy *Z* field, which was further processed through thresholding to generate an error-contaminated area-class map. This process resulted in a total of 100 realizations of area-class maps, with two examples shown in Figures 3a and c, which boiled down in a "mean" area-class map, as shown in Figure 3e. It is visually obvious that *Z*-field-based error modeling is able to map the simulated landscape as shown in Figure 1 accurately.

The main reason for the inaccuracy in indicator simulation-derived area-class maps lies in their "void" of mean-class structures. To fix this, indicator simulation with collocated probability vectors was applied. The probability vectors were derived from summarizing the 100 versions of area-class maps simulated via the discriminant space, as shown in Figures 3a and c. In the Markov-Bayes implementation of indicator simulation with soft data, a value of 0.67 was specified for the parameter $B(z)$, reflecting the factor of 0.33 in the linear model $Z = m_z + \delta Z$. The resultant realizations are shown in Figures 3b and d, with the mean map shown in Figure 3f.

The discrepancy of the three versions of maps derived from stochastic simulation from the assumed true map can be analyzed through cross-tabulation. Such statistics are shown in Table 1, where percent correctly classified (*PCC*) grid cells, kappa coefficients of agreement, and standard errors of kappa estimates are reported for mean maps derived from indicator-based and *Z*-field-based stochastic simulation (Fleiss et al. 1969). Very low agreements between indicator-based maps and the true map are in sharp

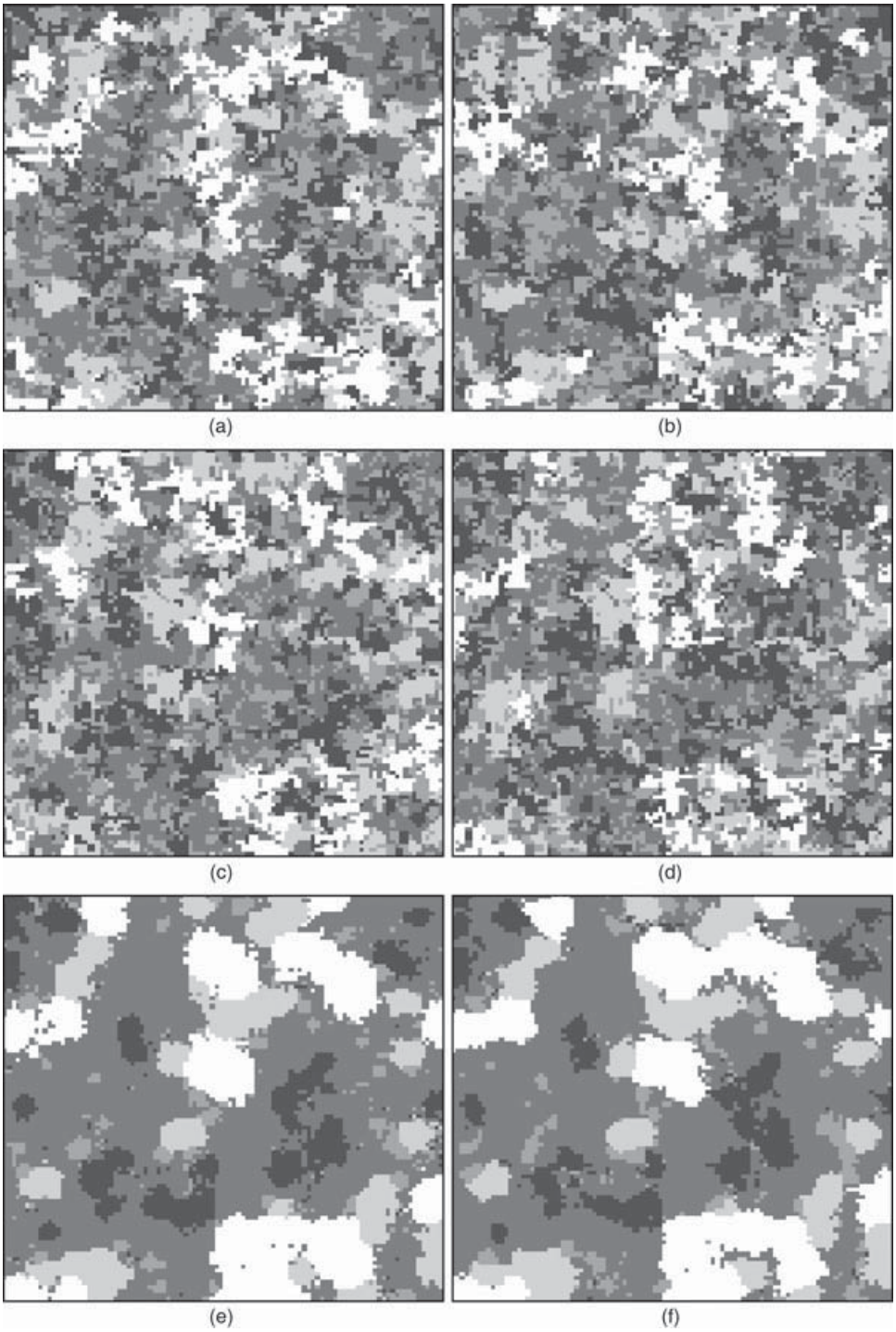

Figure 2 Realized area-class maps and mean maps: (a), (c), and (e) indicator simulation, (b), (d), and (f) indicator simulation in a different class order

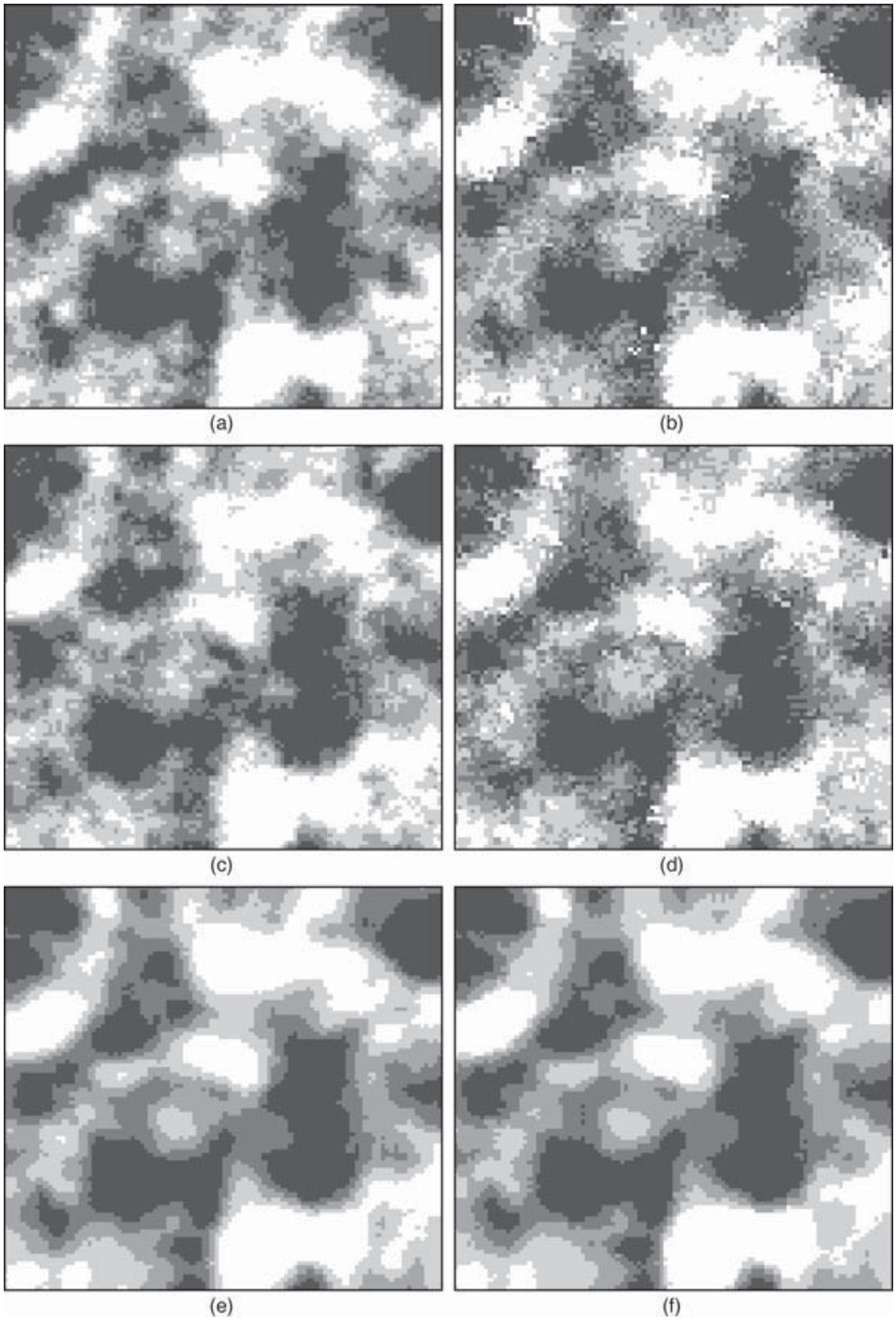

Figure 3 Realized area-class maps and mean maps: (a), (c), and (e) *Z* field-based simulation, and (b), (d), and (f) indicator simulation with colocated probability vectors

Table 1 Confusion matrices for mean maps derived from indicator and *Z*-field simulation

	Indicator simulation I					Indicator simulation II				
Class	1	2	3	4	5	1	2	3	4	5
1	1,487	312	64	10	5	1,497	338	68	13	1
2	192	801	281	26	0	216	799	280	25	2
3	27	103	368	74	29	33	102	392	58	36
4	286	771	1,276	1,779	1,087	250	736	1,240	1,789	1,123
5	8	13	11	111	879	4	25	20	115	838
	PCC = 53.1%, k = 0.414, $s(k)$ = 0.006					*PCC* = 53.2%, k = 0.414, $s(k)$ = 0.006				
	Z field-based simulation					Indicator simulation with soft data				
Class	1	2	3	4	5	1	2	3	4	5
1	1,963	105	0	0	0	1,881	152	0	0	0
2	37	1,834	129	0	0	119	1,722	92	0	0
3	0	61	1,779	84	0	0	126	1,806	95	0
4	0	0	92	1,773	28	0	0	102	1,707	29
5	0	0	0	143	1,972	0	0	0	198	1,971
	PCC = 93.2%, k = 0.915, $s(k)$ = 0.003					*PCC* = 90.9%, k = 0.886, $s(k)$ = 0.004				

contrast with the high accuracy obtained by *Z*-field-based simulation. Indicator simulation with colocated probability vectors registers a significant improvement over the results obtained with simulation with hard indicators alone, although it is a little bit below the *PCC* obtained with *Z*-field-based simulation.

As the non-invariant behaviour due to alternative class orderings was singled out as a key pathology of indicator simulation previously, it is interesting to check the extent of (dis)agreement or (non)replicability observed with the hypothetical data set used in this test. Map comparison was performed between the two mean maps derived from summarizing the two sets of indicator-simulated realizations, generated by applying different class orders. The resultant confusion matrix is reported in Table 2, recording significant differences (almost 20%) between indicator simulators using alternative class orders.

As understood, stochastic simulation can provide statistics about mean and standard deviation for areal extents of different class types, which would otherwise be extremely difficult to compute due to the complications of multi-point spatial correlation. With stochastic simulation, this becomes straightforward so that samples of per-class areal extents can be assembled and summarized from individual map realizations. Results are reported in Table 3, where headers indicator I and indicator II refer to indicator simulation with different class orders, while header indicator III stands for indicator simulation with colocated probability vectors. Interpretation of Table 3 suggests that methods of indicator I and indicator II underestimate classes 1 and 2 while severely overestimating class 4 (by more than 60%). *Z*-field-based simulation provides the best

Table 2 Agreement between mean maps derived from indicator simulation

Indicator I	Indicator II				
Class	1	2	3	4	5
1	1,661	35	11	170	1
2	14	1,176	13	97	0
3	30	12	293	263	3
4	212	99	278	4,369	241
5	0	0	26	239	757
		PCC = 82.6%,	k = 0.738,	$s(k)$ = 0.006	

Table 3 Statistics about per-class areal extents using different methods (units in grid cells, sd = standard deviation)

Methods	true	Indicator I		Indicator II		*Z* field		Indicator III	
Class	mean	mean	sd	mean	sd	mean	sd	mean	sd
1	2,000	1,355	126	1,368	138	2,128	91	2,115	72
2	2,000	1,380	131	1,375	145	1,899	84	1,942	77
3	2,000	2,143	96	2,169	86	1,942	95	1,971	43
4	2,000	3,236	157	3,203	173	1,847	80	1,782	45
5	2,000	1,886	130	1,885	138	2,184	105	2,190	33

overall estimates, while indicator III underestimates class 4 by over 10%, and overestimates class 5 by just under 10%. It is arguable that Markov-Bayes indicator simulation with soft indicator data will produce results closer to those generated by *Z*-field-based simulation, as the parameter $B(z)$ is closer to 1.0. This suggests that Markov-Bayes indicator simulation will provide a viable alternative to *Z*-field-based simulation for replicable uncertainty modeling.

It is of further interest to examine the effects of scaling, i.e. either averaging or taking the commonest classes over larger grid cells when post-processing realizations generated by the *Z*-field-based and indicator simulation, respectively. Means and standard deviations of different class types when up-scaling the original 100 by 100 grid cells are reported in Figure 4, where dots denote the original scale, triangles denote twice the original grid cell size, and squares denote four times the original cell size. In Figure 4, solid lines are for results with *Z*-field-based simulation and dashed lines for indicator simulation. As shown in Figure 4, almost-linear and stable patterns are observed with the per-class areal extents derived from *Z*-field-based simulation, while non-linearity is obvious with the results obtained by the indicator approach. The error bars pertaining to the results obtained by *Z*-field-based simulation tend to be narrower than those by indicator simulation.

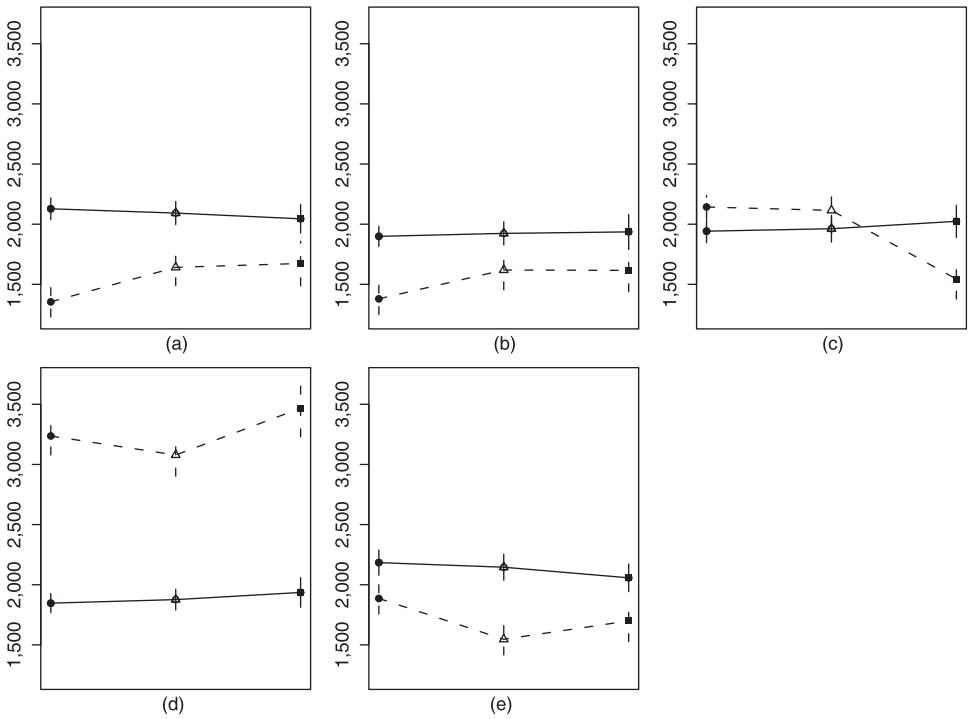

Figure 4 Means and standard deviation in error bars when up-scaling (dots for the original scale, triangles for twice the original grid cell size, and squares for four times the original cell size) by classes: (a) 1, (b) 2, (c) 3, (d) 4, and (e) 5; solid lines indicate results for *Z*-field-based simulation and dashed lines for indicator simulation

5 Discussion and Conclusions

This article has described a new method for categorical mapping and error modeling, which is based on the definition and analysis of a discriminant space and variables therein. Such a method will have great significance for uncertainty-informed categorical mapping, overcoming the drawback of approaches built on error matrices, epsilon error bands, and indicator stochastic simulation. It will enhance consistency, compatibility, and interoperability in the collection and manipulation of area-class information. It will also facilitate scale change as process variables are analyzed using geostatistical downscaling and upscaling to transform originally discrete maps of land cover with minimized loss of inherent variations (Woodcock and Strahler 1987, Thomlinson et al. 1999, Kyriakidis 2004, Atkinson 2005). Further investigations should be directed towards measurement, analysis, and error modeling for area-class maps at multiple scales.

While only a univariate case study ($b = 1$) is considered in this article, real-world applications may well involve multiple discriminant variables, i.e. $b > 1$. It is important to develop stochastic simulation in multi-dimensional discriminant space, with error modeling in area-class maps posing as a familiar task of error propagation (Arbia et al. 2003, Van Niel and Austin 2007). As there is no simple extension from $b = 1$ to $b > 1$,

discriminant class model-constructing methods, especially non-parametric ones, such as decision trees and kernel density estimators, are important and should be further examined, while stochastic simulation integrating linear and indicator geostatistics will lead to avenues of fruitful research. As multi-scale data sets and multi-support application objectives are often encountered in the real world, the interactions of multiple variables and scales pose challenging topics for research. The proposed discriminant models provide a powerful logic and feasible techniques for integrative categorical information and uncertainty analysis.

From a practical perspective, it may be useful to reduce the dimensionality of a discriminant space through principal component analysis if the resultant class models meet accuracy requirements, as easier interpretation of area classes in the discriminant space provides better insights into their signatures and clues for incorporating extra discriminant variables. Clearly, simplified discriminant models will encourage implementation with real data sets, and widespread applications of discriminant models for improved thematic mapping are foreseeable, although some challenges remain in implementing them. Successful applications that produce better information and decision support than conventional practice will say more than mere arguments.

Acknowledgement

This research is partially supported by a grant from the Ministry of Science and Technology of China's "973" Program (2007CB714402). Comments from an anonymous reviewer are received with thanks.

References

Anderson J E, Hardy E E, Roach J T, and Witmer R E 1976 *A Land Use and Land Cover Classification for Use with Remote-Sensor Data*. Washington, D.C., U.S. Geological Survey Professional Paper No 964

Arbia G, Griffith D A, and Haining R P 2003 Spatial error propagation when computing linear combination of spectral bands: The case of vegetation indices. *Environmental and Ecological Statistics* 10: 375–96

Atkinson P M 2005 Spatial prediction and surface modeling. *Geographical Analysis* 37: 113–23

Brown J F, Loveland T R, Merchant J W, Reed B C, and Ohlen D O 1993 Using multisource data in global land cover characterization: Concepts, requirements and methods. *Photogrammetric Engineering and Remote Sensing* 59: 977–87

Bruzzone L and Serpico B 1997 An interative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE Transactions on Geosciences and Remote Sensing* 35: 858–67

Busby J R 2002 Biodiversity mapping and monitoring. In Skidmore A (ed) *Environmental Modelling with GIS and Remote Sensing*. London, Taylor and Francis: 145–65

Carre F and Girard M C 2002 Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. *Geoderma* 110: 241–63

Chrisman N R 1989 Modeling error in overlaid categorical maps. In Goodchild M F and Gopal S (eds) *Accuracy of Spatial Databases*. London, Taylor and Francis: 21–34

Christakos G 1987 The space transformations in the simulation of multidimensional random fields. *Journal of Mathematics and Computers in Simulation* 29: 313–9

Congalton R G 1991 A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37: 35–46

Coppin P, Jonckheere I, Nackaerts K, Muys B, and Lambin E 2004 Digital change detection methods in ecosystem monitoring: A review. *International Journal of Remote Sensing* 25: 1565–96

DeFries R S and Townshend J R G 1994 NDVI-derived land cover classification at a global scale. *International Journal of Remote Sensing* 15: 3567–86

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

Fleiss J L, Cohen J, and Everitt B S 1969 Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72: 323–7

Foody G M 2002 Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80: 185–201

Foody G M and Mathur A 2004 A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geosciences and Remote Sensing* 42: 1335–43

Goodchild M F 1994 Integrating GIS and remote sensing for vegetation analysis and modeling: Methodological issues. *Journal of Vegetation Science* 5: 615–26

Goodchild M F 2003 Models for uncertainty in area-class maps. In Shi W, Goodchild M F, and Fisher P F (eds) *Proceedings of the Second International Symposium on Spatial Data Quality*. Hong Kong, Hong Kong Polytechnic University: 1–9

Goodchild M F and Dubuc O 1987 A model of error for choropleth maps with applications to geographic information systems. In Chrisman N R (ed) *Proceedings Auto Carto 8, Baltimore, Maryland*. Falls Church, VA, American Society for Photogrammetry and Remote Sensing and the American Congress on Surveying and Mapping: 165–74

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

Gotway C A and Stroup W W 1997 A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics* 2: 157–78

Hunsaker C T, Goodchild M F, Friedl M A, and Case T J (eds) *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*. New York, Springer-Verlag

Jenny H 1941 *Factors of Soil Formation*. New York, McGraw-Hill

Journel A G and Huijbregts C H J 1978 *Mining Geostatistics*. London, Academic Press

Kyriakidis P C 2004 A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis* 36: 259–89

Lambin E F and Ehrlich D 1996 The surface temperature-vegetation index space for land cover and land cover change analysis. *International Journal of Remote Sensing* 17: 463–87

Lambin E and Strahler A 1994 Change-vector analysis in multitemporal space: A tool to detect and categorize land cover change processes using high temporal resolution satellite data. *Remote Sensing of Environment* 48: 231–44

Laut P and Paine T A 1982 A step towards an objective procedure for land classification and mapping. *Applied Geography* 2: 109–26

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

McBratney A B, Mendonca Santos M J, and Minasny B 2003 On digital soil mapping. *Geoderma* 117: 3–52

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

McCullagh P and Nelder J A 1989 *Generalized Linear Models*. New York, Chapman and Hall

Moisen G G and Edwards T C 1999 Use of generalized linear models and digital data in a forest inventory of northern Utah. *Journal of Agricultural, Biological, and Environmental Statistics* 4: 372–90

Muchoney D and Strahler A 2002 Regional vegetation mapping and direct land surface parameterization from remotely sensed and site data. *International Journal of Remote Sensing* 23: 1125–42

Richards J A 1996 Classifier performance and map accuracy. *Remote Sensing of Environment* 57: 161–6

Robinove C J 1981 The logic of multispectral classification and mapping of land. *Remote Sensing of Environment* 11: 231–44

Running S W, Loveland T R, Pierce L L, Nemani R R, and Hunt E R 1995 A remote sensing based vegetation classification logical for global land cover analysis. *Remote Sensing of Environment* 51: 39–48

Steele B M, Patterson D A, and Redmond R L 2003 Towards estimation of map accuracy without a probability test sample. *Ecological and Environmental Statistics* 10: 333–56

Stehman S V, Sohl T L, and Loveland T R 2003 Statistical sampling to characterize recent United States land-cover change. *Remote Sensing of Environment* 86: 517–29

Skole D and Tucker C 1993 Tropical deforestation and habitat fragmentation in the Amazon: Satellite data from 1978 to 1988. *Science* 260: 1905–9

Thomlinson J R, Bolstad P V, and Cohen W B 1999 Coordinating methodologies for scaling landcover classifications from site-specific to global: Steps toward validating global map products. *Remote Sensing of Environment* 70: 16–28

Tso B and Mather P M 2001 *Classification Methods for Remotely Sensed Data*. London, Taylor and Francis

Van Niel K P and Austin M P 2007 Predictive vegetation modeling for conservation: Impact of error propagation from digital elevation data. *Ecological Applications* 17: 266–80

Whittaker R H (ed) 1973 *Ordination and Classification of Communities*. The Hague, W Junk

Woodcock C E and Strahler A H 1987 The factor of scale in remote sensing. *Remote Sensing of Environment* 21: 311–32