Categorical models for spatial data uncertainty

Sarah L. Hession, Ashton M. Shortridge and Nathan M. Torbick

Department of Geography
Michigan State University
East Lansing, MI 48824, USA
Tel.: + 001 517 355 4649; Fax: + 001 517 432 1671
hessions@msu.edu; ashton@msu.edu; torbick@msu.edu

Abstract
Considerable disparity exists between the current state of the art for categorical spatial data error modeling and the current state of the practice for reporting categorical data quality. On one hand, the general Monte Carlo simulation-based error propagation framework is a fixture in spatial data error handling; researchers have identified potentially powerful approaches to characterizing categorical data error so that its effects on application uncertainty may be assessed. On the other hand, standard data quality assessments for categorical data are 'spatially unaware,' fail to provide critical information for error propagation models, and neglect the fitness for use paradigm underlying the longstanding rationale for accuracy metadata. Many error assessments rely on area-averaged indicators of map error that do not reflect spatial variability, such as the confusion matrix. How might this gulf between state of the art and state of the practice be bridged? In the present work we lay the foundation for such an edifice: we contrast several categorical error models proposed in the literature in terms of input parameters and performance for a heterogeneous land cover dataset. Familiar methods such as the confusion matrix are considered for their utility in developing error propagation models, as well as theoretically-based, spatially explicit methods like indicator simulation that are not commonly employed in applied research. We develop a comparative matrix to summarize different model requirements, characteristics, and performance, and utilize available secondary data sources where possible to develop improved inputs for the analysis of uncertainty propagation.

Keywords: categorical data, land cover data, indicator kriging, indicator simulation, uncertainty propagation

1 Introduction
For GIS users to make informed decisions about the fitness of spatial data for their specific applications, information about these decisions must be transmitted or integrated with the data. There is "a strong need...to obtain detailed understanding of how errors propagate through the large number of possible combinations of model types, data types, data sources, and kinds of error, and to make this available to users in an easily accessible form" (Burrough et al., 1996). Standard tools for documenting classification accuracy include the error or confusion matrix and a variety of statistics that summarize different aspects of this matrix. Parallel research has been conducted by others on spatial data uncertainty models. These models are capable of generating land cover realizations that can be used as inputs for error propagation analysis via Monte Carlo simulation (e.g. Heuvelink, 1998). A variety of geostatistical and other methods can utilize information provided by the confusion matrix to inform prior probabilities of error that are conditioned on image-derived land cover classifications. These probabilities can be updated into local posterior probabilities and mapped using indicator kriging (Goovaerts, 1997;
Kyriakidis and Dungan, 2001). Stochastic simulation can then be performed to evaluate uncertainties in the spatial distribution of categorical data error by generating multiple alternative realizations (maps) of land cover classifications conditional on all of the information provided. This paper considers and implements several methods for employing the confusion matrix to develop uncertainty models for categorical data. These fundamental concepts are reviewed in the following subsections.

1.1 The confusion matrix

Characterizing thematic classification accuracy has occupied the attention of researchers in remote sensing and GIS for decades. The general scenario involves “ground truth” or reference data at a set of locations in the region of interest. The degree of correspondence between categories in the reference data and their collocated classified labels is employed to characterize map accuracy in a variety of ways.

The confusion matrix is typically the most complete representation of classification accuracy assessment employed by researchers; a large number of common metrics may be derived from it. Formally, consider \( n \) land cover categories \( K_1, K_2, \ldots, K_n \) for a region and data set of interest. The confusion matrix is an \( n \times n \) matrix in which each cell value \( p_{ij} \) is the proportion of locations classified as \( K_i \) and with reference category \( K_j \). Congalton and Plourde (2002) and Stehman (1997) provide summaries of the calculation and use of the confusion matrix for thematic classification accuracy.

Although it represents accuracy at a high degree of thematic precision through class-wise measures of agreement, as a global metric the confusion matrix provides no spatially varying information about accuracy (Canters, 1997; Steele et al., 1998; McGwire and Fisher, 2001). There are two components to this limitation. First, there is an implicit assumption that the class-wise agreements are stationary: that the interclass relationships observed in the confusion matrix are invariant across the region of interest. Second, no information about the geographic structure of error is preserved: that second order spatial patterns do not exist.

1.2 Uncertainty propagation

The preceding brief discussion of the confusion matrix is intended not only to illustrate some problems for the metric but also to emphasize its utility for characterizing thematic accuracy. The substantial volume of literature on the confusion matrix and its offspring highlight its importance for reporting data quality. In this paper, however, we are primarily concerned with the potential of the metric for modeling data uncertainty in a spatially explicit manner. A growing body of research on spatial uncertainty modeling using a Monte Carlo simulation approach indicates a diversity of model approaches, methods, and results. Burrough (1999) reviews the case for greater inclusion of geostatistics in spatial analysis, and identifies the statistical approach to characterize error propagation via stochastic simulation as a key advantage. In general, any model that can account for spatial autocorrelation to estimate local (cell or grid point-based) distributions can be used to generate realizations either of error or of the actual phenomenon.

In section two we briefly review a number of approaches for modeling categorical data uncertainty using information from the confusion matrix. Section three compares two of these methods using a study site in Kenya. Finally in section four we discuss results and prospects for a more robust approach to categorical accuracy reporting using an error propagation framework.
2 Modeling approaches

The methods described in this paper utilize both image-derived (soft) data that are generally exhaustive and higher accuracy (hard) data, which are typically sparse. Hard data, or reference data, are often obtained through ground-based surveys, as in our case study, but may be obtained from higher accuracy remotely-sensed land cover products. Table 1 presents a survey of eight methods capable of generating multiple realizations of categorical data.

Table 1. Matrix of methods for characterizing categorical data error distributions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Implementation</th>
<th>Input data requirements/ parameterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Confusion frequency simulation (Fisher, 1994)</td>
<td>Simulating class labels via user's accuracy perspective</td>
<td>Land cover class map and a confusion matrix</td>
</tr>
<tr>
<td>2. Indicator cokriging with sequential indicator simulation (SIS) (Boucher and Kyriakidis, 2005)</td>
<td>Super-resolution mapping (sub-pixel mapping or downscaling) using Landsat thematic mapper (TM) data</td>
<td>Coarse resolution satellite data; set of class labels at the fine (target) resolution, e.g., via ground surveys; indicator variogram models</td>
</tr>
<tr>
<td>3. Simple indicator kriging with varying local means and SIS (Kyriakidis and Dungan, 2001)</td>
<td>Simulating class labels and uncertainties using Landsat TM data for propagation of uncertainty</td>
<td>Land cover class map subject to classification error and/or limited resolution; set of validation points (class labels); variogram models of residuals (hard labels minus soft conditional probabilities)</td>
</tr>
<tr>
<td>4. Indicator kriging and SIS using quadtree segmentation (de Bruin et al., 2004)</td>
<td>Error modeling in categorical raster data</td>
<td>Thematic map (categorical raster data); reference data; indicator variogram models</td>
</tr>
<tr>
<td>5. SIS using collocated indicator cokriging (Magnussen and DeBruin, 2003)</td>
<td>Updating forest cover type maps</td>
<td>Classified image (soft data); reference sample (hard data); indicator variogram models</td>
</tr>
<tr>
<td>6. Spatial data uncertainty model (SDUM) using Monte Carlo simulation (Ehlschlager, 2002)</td>
<td>Simulating errors in spatial data to evaluate uncertainty propagation</td>
<td>Inaccurate, coarse resolution elevation data; validation data points</td>
</tr>
<tr>
<td>7. Hopfield neural network (Tatem et al., 2002)</td>
<td>Super-resolution mapping using Landsat TM data</td>
<td>Prior information on typical spatial arrangement of land cover types incorporated into energy function as a semivariance constraint</td>
</tr>
<tr>
<td>8. Markov chain-based probability vector (PV) approach to conditional simulation (Li et al., 2005)</td>
<td>Modeling spatial uncertainties of soil classes</td>
<td>Survey line data (generated by observing soil class boundary changes along regular survey lines)</td>
</tr>
</tbody>
</table>
Each of these methods can be used to characterize the error distributions of categorical data and generate alternate realizations (maps) for error propagation. Method one does not explicitly account for spatial variability. Methods two through five each require input data in the form of image-derived (soft) data, reference (hard) data, and indicator variogram models. Method eight was developed based on survey line data and cannot yet be applied to raster data. In this paper, methods one and two are applied to a case study; their implementation will be discussed in more detail in the next subsections.

2.1 Confusion frequency simulation

The confusion matrix may be regarded as a set of conditional probabilities: given that category \( K_i \) is observed at a particular location, the probabilities of the location actually being class \( K_1, K_2, \ldots, K_n \) are available. This suggests a simple approach for simulating realizations with appropriate class proportions (Fisher, 1994). For each cell \( i \) in the input land cover raster grid:

1) identify the reported category label \( K_i \); 2.) look up the associated conditional probabilities from row \( K_i \) of the confusion matrix; 3) draw a random value from a uniform \([0,1)\) distribution; and 4) based on the probabilities and the random value, assign class \( K_j \) to the output realization cell \( i \).

This confusion frequency simulation method was implemented by the authors in the C programming language. Its advantages are computational simplicity, speed, and the preservation of the observed inter-category class confusions from the confusion matrix. The resulting proportions of each category in each realization match the expected values from the confusion matrix. Note that this method assumes that cell transitions are spatially independent; cells are processed without regard to surrounding values. In addition, the locations of the reference data do not inform the simulation. In geostatistical parlance, this is an unconditional approach that does not honor any known data.

2.2 Indicator (soft) cokriging

The method selected for use in our case study is indicator cokriging followed by sequential indicator simulation. In this method, soft indicator data are derived by validation of the image-derived classes and used as secondary data to predict multiple realizations of higher accuracy class labels throughout the study area.

A categorical random variable (RV) is defined as \( Z(u) \), that can take on one of \( K \) mutually exclusive and exhaustive categories \( \{z_k, k = 1, \ldots, K\} \), such as land cover classes at any location \( u = (x,y) \) within the study area domain \( D \). The hard data consist of \( n \) class labels representing reference data \( \{z(u_\alpha) = z_k, \alpha = 1, \ldots, n\} \). Soft data consist of \( N \) image-derived class labels \( \{x(u_\alpha') = z_k, \alpha = 1, \ldots, N\} \).

In the indicator framework, each datum is coded into a set of \( K \) local probabilities (Kyriakidis and Dungan, 2001; Journel, 1986), each associated with the \( k \)th outcome \( z_k \):

\[
\text{Prob}\{Z(u) = z_k|\text{info}(u)\}, k = 1, \ldots, K
\]

representing the probability that class \( z_k \) is observed at location \( u \) given information \( \text{info}(u) \) at that location.

As described by Kyriakidis and Dungan (2001), hard probabilities are coded as follows:

\[
i(u; z_k) = 1 \text{ if } Z(u) = z_k \text{, and } i(u; z_k) = 0 \text{ if not, } k = 1, \ldots, K
\]
yielding a vector of K hard indicator probabilities. Soft probabilities are calculated based on the soft categorical data, resulting in probabilities between 0 and 1:

\[ y^*(u'; z_k) = \text{Prob}^*\{Z(u') = z_k | x(u') = z_k\} = p^*(z_k | z_k) \quad (2) \]

where \( p^*(z_k | z_k) \) is the frequency of observing category \( z_k \) given that category \( z_k' \) occurs in the image-derived data at the same location (i.e., the user’s accuracy) and is estimated as:

\[ p^*(z_k | z_k') \approx \frac{\sum_{\alpha=1}^{n} j(u_\alpha; z_k) j(u_\alpha; z_k')}{\sum_{\alpha=1}^{n} j(u_\alpha; z_k')}, \quad k, k' = 1, \ldots, K \quad (3) \]

In soft cokriging, the hard and soft indicator data are combined, where \( i(u; z_k) \) is the primary or dependent variable and \( y^*(u; z_k) \) is the secondary variable. Goovaerts (1997) describes the ordinary indicator cokriging estimator for class \( z_k \) and the method for deriving the cokriging weights. For details, the reader is encouraged to refer to Goovaerts (1997, p. 310). In deriving this system of equations, note that two autocovariance functions \( C_{II}(h; z_k) \) and \( C_{YY}(h; z_k) \) and one cross covariance function \( C_{IIY}(h; z_k) \) must be modeled for each class \( K, k = 1, \ldots, K \).

Sequential indicator simulation then proceeds as follows to generate an alternate realization (Goovaerts, 1997):

1. Define a random path visiting each node of the grid only once.
2. At each location \( u \), determine the K conditional cumulative distribution function (ccdf) values using indicator cokriging with the direct and cross semivariogram models of the hard and soft probabilities. The conditioning information consists of neighboring hard and soft probabilities and previously simulated values.
3. Correct for any order relation deviations and build a complete ccdf.
4. Draw a simulated value from that ccdf and add it to the data set.
5. Proceed to the next point along the random path and repeat steps 2 through 4.

Other realizations can be obtained by repeating the entire process with a different random path for each realization.

3 Case study

To illustrate the methods described above, a case study utilizing both non-spatial confusion frequency simulation and indicator cokriging with sequential indicator simulation is conducted.

3.1 Study Area

The study area is located in central Kenya (Figure 1). The simulation domain extends 56 km east-west and 53 km north-south centered at 37°24'10 east 1°5'43 south. Present within the study area are expansive grasslands, small urban pockets, forest, shrublands, water, barren areas, and agricultures. These land categories comprise seven general land categories that cover the entire study region.
3.2 Data

The image-derived land cover data set (soft data) used in this study is the Global Land Cover (GLC) 2000 product available from the Joint Research Centre’s (JRC) Global Vegetation Monitoring Unit (Mayaux et al. 2004). The overall objective in creating GLC2000 products was to provide a common land cover database for the globe as part of the Millennium Ecosystem Assessment (Bartholome et al. 2002, Mayaux et al. 2003). GLC2000 utilized 14 months of 1 km resolution VEGETATION imagery from the Satellite Pour l’Observation de la Terre (SPOT-4) collected between 11/1/1999 – 12/31/2000.

The classification scheme used the Land Cover Classification System (LCCS) which was developed in part to address problems arising from products made with different resolution data sources and/or different classification objectives. The LCCS provides a standardized, hierarchical, a priori classification system. The central idea of the LCCS method is to use universal, independent diagnostic attributes to create land use land cover (LULC) products that can be compared across scales and sources (Di Gregorio and Jansen, 2005). The landscape is classified by categorical assignments initially from broad characteristics followed by a series of descriptive characteristics. The LCCS establishes major land categories using parameters and key landscape factors including vegetated or non-vegetated surfaces, terrestrial or aquatic systems, cultivated and managed, natural and semi-natural, life-form, cover layer, feature height, spatial distribution, leaf type, and phenology.

An assessment of the LULC product was conducted using airborne videography techniques (Torbick et al., 2005). The videography (ground truth) data consisted of two low-altitude, light aircraft flights in the region totaling approximately 900 km in length. These flight paths were flown over selected socio-ecological gradients. Figure 1 illustrates the flight paths in the region. Using this information, a confusion matrix (Congalton and Green 1999) was constructed to assess GLC2000 classification accuracy (Table 2). A stratified random sampling
scheme extracted video frames to construct the confusion matrix and generate the hard reference data used in this analysis.

Table 9 GLC2000 versus reference data confusion matrix.

<table>
<thead>
<tr>
<th>GLC Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Row Sum</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0.11</td>
</tr>
<tr>
<td>Wood-shrubland</td>
<td>7</td>
<td>17</td>
<td>0</td>
<td>65</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>90</td>
<td>0.18</td>
</tr>
<tr>
<td>Grassland</td>
<td>3</td>
<td>0</td>
<td>29</td>
<td>59</td>
<td>30</td>
<td>2</td>
<td>0</td>
<td>120</td>
<td>0.49</td>
</tr>
<tr>
<td>Agriculture</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0.63</td>
</tr>
<tr>
<td>Barren</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>0.00</td>
</tr>
<tr>
<td>Water</td>
<td>6</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0.00</td>
</tr>
<tr>
<td>Urban</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Null</td>
<td></td>
</tr>
<tr>
<td>Column Sum</td>
<td>18</td>
<td>62</td>
<td>70</td>
<td>193</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Producer</td>
<td>0.55</td>
<td>0.27</td>
<td>0.84</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3 Implementation of confusion frequency simulation

Figure 2 portrays a realization of the error frequency simulation method for the Kenya landcover data. Simulations are also presented for two individual classes (i.e., woodland-shrub and grasslands).

Figure 2. (a) GLC2000 data for the Kenya study area. (b) Error frequency simulation realization. (c) Error frequency simulation realization for woodland-shrub class (urban shown in black). (d) Error frequency simulation realization for grassland class (urban shown in black).
3.4 Implementation of spatially explicit method

Indicator (soft) cokriging with sequential indicator simulation was implemented on a class by class basis. Formulas (1) and (2), respectively, were used to derive hard probabilities for the reference data and soft probabilities for the image-derived data corresponding to the reference locations. Formula (2) was also used to derive soft probabilities corresponding to the GLC2000 classifications for each pixel in the study area. The resulting data consisted of 348 hard and soft class probabilities at the reference locations and 2,968 soft class probabilities.

For each of the $K = 7$ land cover classes, two semivariograms were modeled based on the reference and soft probabilities as well as the cross variogram. Five realizations were simulated for each class where 1 = class is present, 0 = class is not present. Figure 3 depicts one realization for each of three classes below their respective GLC2000 class data sets.

![Realizations for Agriculture, Forest, and Grassland](image)

Figure 3. GLC2000 data for agriculture, forest, and grassland classes compared to realizations generated through indicator cokriging and sequential indicator simulation.

4 Results and Discussion

In contrast to the original data, the realizations generated by confusion frequency simulation (Figure 2) are spatially heterogeneous. Any spatial structure maintained in the realization is solely a function of the original configuration of category labels, as the model does not account for spatial autocorrelation.

Less obvious is the substantial change in class proportions from the original data to the confusion frequency simulation realization. For example, forest comprised 22 percent of the original data. User accuracy for this class was only 11 percent, however; 87 percent of known cells classed as forest were identified as agricultural land cover in the reference data. Forested locations were only infrequently confused as other classes. As a result, forest comprises only 4.6 percent of the error frequency realization. Wood-shrubland was similarly confused with agriculture; however, grassland was confused with wood-shrubland, and the resulting
simulation shows little change in overall wood-shrubland coverage (both are approximately 18.6 percent). Grassland dropped from 41 percent of total area to 20 percent, while agriculture increased from 18 to 55 percent of the total area.

The misclassifications summarized by the confusion matrix (Table 2) and described above also account for much of the difference between GLC2000 class data and the realizations generated through indicator cokriging and sequential indicator simulation (Figure 3). For example, agriculture is predicted throughout the study region as a result of confusion between agriculture and forest or woodland-shrub. The distribution of grassland more closely approximates the GLC2000 data, and portrays more spatial contiguity than the confusion frequency simulation shown in Figure 2, as expected.

The error frequency simulation approach suffers from the major drawback of failing to incorporate any information about spatial structure of land cover categories. Consequently it is an inappropriate method for applications requiring joint probabilities of category membership, including representative spatial maps. If probabilities for a single location (cell) are desirable, the method is adequate, although any nearby information about reference land cover is ignored. For such applications however, full map simulations appear unnecessary.

When simulated data are to be mapped, however, spatially explicit models are preferred. Many of the methods considered in this paper can produce more realistic maps by utilizing information provided by the confusion matrix and extending this information to generate maps conditioned on image-derived land cover classifications. Further, land cover realizations from these models can be used as inputs for error propagation analysis via Monte Carlo simulation. However, data requirements remain an important consideration. Reference or ground-based data are rarely representative of the actual spatial distribution of land cover classes. This issue can be addressed through the development of sampling designs that include adequate numbers of representative samples for each land cover class. Other sources of data can also be considered for reference data, included high resolution remotely sensed data that can provide vast amounts of higher quality data.

Acknowledgements
This work was funded through U.S. Department of Defense grant HM1582-05-1-2011.

References


