

# Error Propagation through the Buffer Operation for Probability Surfaces

Howard Veregin

## Abstract

*This study explores the propagation of error through the buffer operation in GIS. The study focuses on probability based raster databases, or probability surfaces, in which cell values show the probabilities associated with membership in different land-cover classes. Results indicate that there is a strong positive relationship between error levels in source and derived layers. The strength of the relationship is affected by the degree to which source probabilities tend to be under- or over-estimated, and by the interaction between buffer size and spatial covariation in source probabilities.*

## Introduction

This study focuses on the implications of error propagation for GIS-based modeling applications. Error propagation refers to the process of error transference from source to derived data. This process occurs through the application of GIS operations, which transform and merge source data with the objective of deriving specific spatial relationships implicit within these data. Errors in source data are transferred by these operations, and are also modified such that source and derived data may have different error characteristics. Both amplification and suppression of error levels can occur through the application of conventional GIS operations.

Error propagation modeling refers to the attempt to emulate the processes of source error modification and transference, with the goal of estimating the error characteristics of derived data products. Error propagation functions are mathematical and computational techniques designed to propagate error through specific GIS operations. Each GIS operation requires a unique error propagation function (Lanter and Veregin, 1992). Automated error propagation systems must be capable of determining the appropriate error propagation function to employ, and be able to concatenate these functions to mirror the flow of data through a given sequence of GIS operations. Some prototype systems have been developed in recent years (e.g., Heuvelink *et al.*, 1989; Lanter and Veregin, 1990; Carver, 1991). However, these systems are limited to selected subsets of GIS operations and often employ error propagation functions that are applicable only in the controlled conditions found inside the laboratory.

This study concentrates on the propagation of error through the buffer operation. This operation involves the delineation of a geometric zone of specified width around a set of features on the source layer. It is commonly employed in GIS-based modeling applications involving site selection, suitability analysis, and environmental assessment. Despite its widespread use, relatively little is known about error propagation for this operation. Moreover, the mechanisms of error propagation for the this operation are quite complex, due to interaction between spatial and aspatial error components.

Error propagation for the buffer operation is also affected by characteristics of the data model. For the vector model, a buffer is a geometric feature defined in terms of distance from a set of features on the source layer. Thus, error propagation is most appropriately modeled in terms of locational error in the source features. For the raster model, in contrast, the buffer zone comprises all cells within a specified distance of a set of "feature cells" on the source layer. Here, error propagation is most appropriately modeled in terms of thematic error, i.e., errors of omission and commission associated with the feature cells.

Veregin (1994) examines error propagation through the buffer operation for conventional raster databases. The present study extends this work into the realm of probability based raster databases, or probability surfaces, in which cell values indicate the probability of membership in different land-cover classes. Probability surfaces have been proposed as a means of incorporating information on mixed-cell class composition into GIS databases (e.g., Goodchild *et al.*, 1992). Use of these surfaces in GIS-based modeling applications may yield greater precision in model results by accounting for the continuous nature of landscape variation (e.g., Burrough *et al.*, 1992). Despite the potential usefulness of this approach, it has not been widely adopted due to a lack of standard GIS tools for manipulation of probability surfaces.

This study examines the quality of data products derived through the use of probability surfaces in GIS-based analysis procedures. The study focuses specifically on the factors that affect the quality of data derived through the application of a probability based analog of the conventional buffer operation. A brief explanation of probability surfaces and how buffers can be derived for such surfaces is described. The mechanics of error propagation through the buffer operation for conventional raster databases and probability surfaces is then discussed. The hypotheses to be tested are outlined and the methodology employed in the study is then detailed. Finally, results are presented and discussed.

Results suggest that error propagation for probability surfaces has much in common with error propagation for conventional raster databases. In both cases, source and buffer errors are positively correlated, with a decline in the rate of error propagation as buffer size increases. For both probability surfaces and conventional databases, error propagation is affected by spatial covariation. In contrast to conventional databases, however, levels of buffer error for probability surfaces tend to be larger than levels of source error. This suggests that error propagation may have significant effects on the results of GIS-based analysis procedures for probability surfaces.

Photogrammetric Engineering & Remote Sensing,  
Vol. 62, No. 4, April 1996, pp. 419-428.

Department of Geography, Kent State University, Kent, OH 44242-0001. Presently with the Department of Geography, University of Minnesota, Minneapolis, MN 55455.

0099-1112/96/6204-419\$3.00/0  
© 1996 American Society for Photogrammetry  
and Remote Sensing



## Probability Surfaces

### Rationale

Traditional classification methods used to derive raster land-cover data assume that each cell contains only one class. However, this assumption is often not borne out by empirical observation and is known to be unrealistic for geographical phenomena that exhibit gradual transitions between classes or contain mixtures of classes that become apparent at different spatial scales. For such phenomena, pure cells containing only one class are rare, and it is more common to find cells that contain a mixture of classes. The assumption of pure cells is a limitation that can lead to imprecision in the results of GIS-based analysis procedures, especially in areas of high spatial heterogeneity. This suggests that there is a need for alternate data models that account for the continuous nature of landscape variation (Burrough *et al.*, 1992; Goodchild *et al.*, 1992).

Information about mixed-cell class composition can be encoded in GIS databases using a variety of different data models. Perhaps the most flexible of these is the probability surface model, in which a vector of probability values is encoded for each cell. Let the variable  $p_{ic}$  represent the probability that cell  $i$  belongs to land cover class  $c$ . Given  $m$  cover classes, a probability vector of the form  $[p_{i1}, p_{i2}, \dots, p_{im}]$  exists for each of the  $n$  cells in the layer. Individual elements of the vector reflect the probability of observing a particular class and may be referred to as class probabilities (Goodchild *et al.*, 1992).

The probability surface model is a generalization of the conventional raster data model rather than a radical departure from it. In conventional raster databases depicting land-cover classes, each cell  $i$  is coded with a nominal value  $q_i$  (with a value of 1 through  $m$ ) that serves to identify the most probable of  $m$  cover classes. The conventional model can easily be recast in the context of a probability surface model, in which the probability vector for any cell contains  $m-1$  elements with values of zero, and a single element with a value of one corresponding to the most probable class. The nominal value for cell  $i$ ,  $q_i$ , is simply the index value of the most probable class. That is,

$$q_i = c \mid p_{ic} = \max(p_{iv}), \quad v = 1, 2, \dots, m \quad (1)$$

Thus, the conventional raster model can be viewed as a transformation of the more general probability surface model, in which the elements of the probability vector for a cell defined in the closed interval  $[0,1]$  are mapped to the set  $\{0,1\}$ . This transformation process imparts a reduction in the level of taxonomic resolution and, hence, a potential loss of detail in model output.

### Derivation

Land-cover information is a basic requirement of many GIS-based modeling applications, and remote sensing is often used as a source of this information. However, while numerous classification methods have been developed to derive land-cover information from remotely sensed data, few of these methods preserve information about mixed-cell class composition. Rather, these methods typically assume that only the dominant land-cover class is of interest, despite widespread agreement that the notion of pure cells is often unrealistic.

Some classification methods do preserve information about mixed-cell class composition. Indeed, many classification procedures calculate a measure of the strength of membership for all cover classes prior to assignment of class values to cells (Fisher, 1994). For example, discriminant analysis yields probability estimates defining cell membership in a set of  $m$  classes. Maximum-likelihood classifiers

likewise compute the probability that each cell belongs to each of a set of  $m$  classes, in order to then determine which class has the highest probability. Probability-like measures can even be derived for classification procedures based on distances defined in spectral space (Fisher, 1994).

Fuzzy set theory (first articulated by Zadeh (1965)) provides another model for mixed cells. In contrast to conventional set theory, fuzzy set theory allows for a variable degree of membership in a set. The degree of membership is defined by a membership value, which is normally expressed as a value in the closed interval  $[0,1]$ . Values closer to 1 indicate a higher degree of membership. Classification algorithms based on fuzzy set theory can be used to obtain an  $m$ -dimensional vector of membership values for each cell, where each element of the vector indicates membership in a particular cover class.

Despite their apparent similarities, there are important differences between this model and the probability-based model described earlier. In fuzzy set theory, there is no formal requirement that the membership values for any observation sum to one, as in the case of probability-based models (Robinson and Thongs, 1985). In fuzzy set theory, membership values define possibility rather than probability. It is generally held that possibility is based on a broader interpretation of uncertainty than probability. Certainly, the two theories address a different kind of uncertainty and handle it in quite different ways.

Despite these differences, the operational distinction between membership values and probabilities is itself rather fuzzy. For example, some algorithms for fuzzy classification, including the well-known fuzzy c-means algorithm (Bezdek *et al.*, 1984), have a probabilistic interpretation, as they force the membership values for any observation to sum to one. Moreover, numerous studies have shown that membership values derived from fuzzy classifiers are correlated with coverage areas (or proportions) for different classes within mixed cells (Fisher and Pathirana, 1990; Veregin and Sultana, 1992; Foody, 1994). Thus, membership values yield useful predictions of the probability of observing a particular class within a cell.

Undoubtedly, many proponents of fuzzy set theory would be disinclined to argue for a convergence in interpretations of membership values and probabilities. However, it is not the intention of this study to argue that the two are synonymous. Rather, the starting point for the study is an assumed ability to derive information on class composition at sub-cell spatial scales and to encode this information in a raster database. The subtle differences between the possibilistic and probabilistic interpretations of this information, while an important topic in its own right, is necessarily somewhat tangential to the objectives of the study.

### The Buffer Operation

As in the case of conventional raster databases, the buffer operation is applied to probability surfaces to derive information on proximal relationships, i.e., relationships associated with proximity or distance. For conventional raster databases, a buffer is defined as the set of cells within a specified distance (the "buffer size") of a set of "feature cells." Feature cells are the cells on the source layer that correspond to the cover class (or classes) of interest (e.g., water or wetlands). Cell values on the source layer indicate membership in the set of feature cells. Typically, values of 1 and 0 indicate membership and non-membership, respectively. The value for a given cell  $i$  on the derived buffer layer is obtained by examining the values of all cells on the source layer for which the distance to cell  $i$  is less than or equal to the buffer size. If any of these cells has a value of 1, then the value for cell  $i$  on the buffer layer is also equal to 1. If no



cell with a value of 1 is found within the distance neighborhood, cell  $i$  on the buffer layer is assigned a value of 0.

For probability surfaces, cell values on the derived buffer layer are likewise defined in terms of proximity to the cover class of interest. However, in this case cell values on the source and buffer layers are defined as the closed interval  $[0,1]$  rather than as elements of the set  $\{0,1\}$ . The value of a given buffer cell reflects the probability that the cell is within the buffer distance of the class of interest. For probability surfaces, the buffer operation is in essence a logical union operation. For any cell  $i$ , the probability that the cell is within the buffer distance of the class of interest is equal to the union of the probabilities of all cells for which the distance to cell  $i$  is less than or equal to the buffer size. Each cell with a non-zero probability within the distance neighborhood of cell  $i$  adds to the buffer probability of the cell, because each of these cells increases the probability that cell  $i$  is within the buffer distance of the class of interest.

Implementation of the union operation depends on the degree of independence assumed to exist in cell probabilities. At one extreme, the contribution of any cell can be seen as independent of any other cell. This assumption is valid if the process that generated the probability surface operates independently at each location (e.g., a random value between 0 and 1 is generated independently for each cell). Under this assumption, the buffer layer can be derived as follows:

$$b_{ic} = 1 - \prod_{j | d_{ij} \leq b} (1 - p_{jc}) \quad (2)$$

In this equation,  $b_{ic}$  is the buffer probability for cell  $i$  for class  $c$ ,  $p_{jc}$  is the source probability value for cell  $j$  for class  $c$ ,  $d_{ij}$  is the Euclidean distance between cells  $i$  and  $j$ , and  $b$  is the buffer size. This equation is the same as that used to compute reliability for accumulation of independent evidence (Tikunov, 1986).

When autocorrelation is present, Equation 2 is inappropriate, because the probability value at any location will not be independent of the value at neighboring locations. In order to account for autocorrelation, the contribution of individual cell probabilities must be reduced. It is proposed that, under the assumption of positive autocorrelation, buffer probabilities be derived using the following equation:

$$b_{ic} = \max(p_{jc}), \quad \forall j | d_{ij} \leq b \quad (3)$$

The equation specifies that the buffer probability for cell  $i$  be equal to the maximum probability of the set of cells within the buffer distance of cell  $i$ . The equation discounts the contribution of neighboring cells under the assumption of positive autocorrelation.

Equations 2 and 3 represent extremes of a continuum. Other definitions of buffer probability can be defined between these two extremes. Equation 2 makes the most liberal allowance for autocorrelation by assuming that each cell contributes independently to the combined probability. Equation 3 uses the most conservative interpretation of autocorrelation, as it discounts the contribution of all cells other than the cell within the maximum probability within the buffer zone. As noted above, the buffer operation may be viewed as a logical union operation; Equation 3 parallels the way in which the logical union operation is typically performed for fuzzy membership data (Leung, 1988).

Either of the two equations can be used to produce a buffer for conventional raster data, given a recasting of the conventional raster data model into a probability surface model as described earlier. Note also that, whether one uses Equation 2 or Equation 3, the buffer can be defined for any class  $c$ .

Figure 1 shows examples of buffers produced by the two

equations. The original probability surface is shown in Figure 1a. (The source of this surface is described below.) Figures 1b and 1c show the buffer probability surfaces based on a buffer size of two cells (60 m). Note that Equation 2 produces a smoother probability surface than does Equation 3. Indeed, even for fairly small buffer sizes, Equation 2 often saturates the buffer probability surface with values close to 1.

## Error Propagation for the Buffer Operation

### Conventional Databases

As noted earlier, the issue of error propagation for the buffer operation has not received a great deal of attention in the GIS literature (but see Veregin (1994)). Error propagation research in the context of probability surfaces is rarer still. Most error propagation research has focused on conventional categorical or numerical raster and vector data layers (e.g., Newcomer and Szajgin, 1984; Burrough, 1986; Heuvelink *et al.*, 1989; Veregin, 1989; Wesseling and Heuvelink, 1991).

Veregin (1994) examines error propagation for the buffer operation in the context of conventional raster databases, in which cells are coded with values of 0 or 1 to indicate membership in a set of feature cells around which the buffer is to be generated. The accuracy of the derived buffer layer is found to be dependent on the following factors:

- *Source PCC* is the probability that cells in the source layer are correctly classified as either feature or non-feature cells. A positive relationship exists between source accuracy and buffer accuracy, indicating that the higher the accuracy of the source data, the more accurate the buffer layer.
- *Buffer size* is the width of the buffer. A positive relationship exists between buffer size and accuracy. The buffer layer tends to be more accurate when buffer size is large. The effects of buffer size depend on two competing forces, the first being the tendency for thematic error to grow in direct proportion to the width of the buffer, and the second being the tendency for saturation to occur for large buffer sizes.
- *Feature probability* is the proportion of cells in the source layer that are defined as feature cells. In general, a higher probability implies a higher buffer accuracy, due to the enhanced tendency for saturation to occur, even for small buffer sizes.
- *Feature geometry* is the degree to which feature cells tend to cluster in space. An inverse relationship exists between feature geometry and buffer accuracy. The less clustered the feature cells, the more accurate the buffer layer. This is due to the propensity for dispersed feature cells to produce a greater number of buffer cells, which in turn is associated with an enhanced propensity for saturation.
- *Error distribution* is the degree to which misclassified cells tend to cluster in space. A positive relationship exists between the error distribution and buffer accuracy. The less clustered the misclassified cells, the less accurate the buffer layer. Clustering of misclassified cells tends to minimize the number of misclassified buffer cells that will be produced for a given buffer size.

### Probability Surfaces

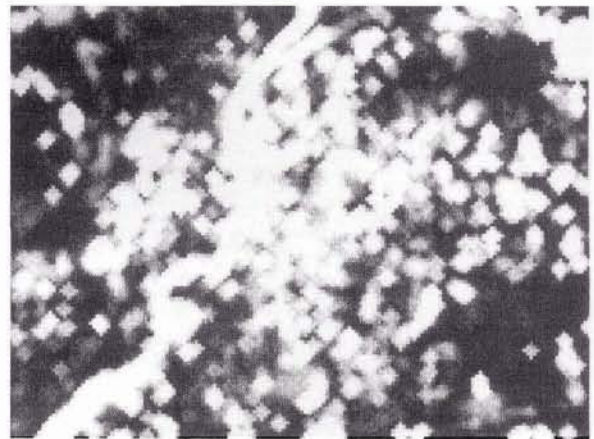
For probability surfaces, the amount of error propagated to the buffer layer depends on the amount of error present in source layer probability values. The relationship is affected by the direction (or sign) of the source errors. Over- and under-estimation of source probabilities have different effects on the degree of buffer error. These effects are mitigated by the interaction between buffer size and spatial covariation in source errors.

The discussion that follows focuses on the computation of buffer probabilities as given in Equation 3. This equation provides the most conservative estimate of buffer probabilities, and therefore serves to define the lower bound of propagated error for the buffer operation.

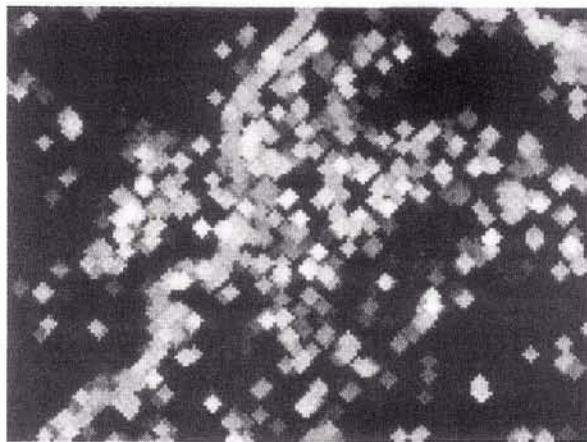




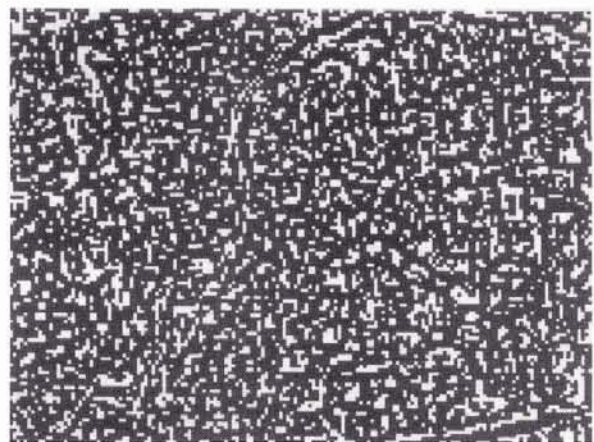
(a)



(b)



(c)



(d)

Figure 1. (a) Probability surface for water and wetlands for an area centered on Kent, Ohio. (b) Buffer for probability surface in (a) using Equation 2 and based on a buffer size of two cells (60 m). (c) Buffer for probability surface in (a) using Equation 3 and based on a buffer size of two cells (60 m). (d) Local maxima for probability surface in (a). See text for explanation of equations.

The effects of source error on the buffer layer depend on the degree to which errors occur at local maxima on the source layer. A local maximum is defined as a cell with a locally high probability, such that this probability is radiated over space when the buffer layer is constructed. A given cell  $i$  is defined as a local maximum ( $x_{ic} = 1$ ) if, for any cell  $j$  on the layer, cell  $i$  has the largest probability value of all of the cells for which the distance to cell  $j$  is less than or equal to the buffer size. That is,

$$x_{ic} = \begin{cases} 1 & \text{if } p_{ic} = \max(p_{kc}), \quad \forall k \mid d_{ik} \leq b, j=1,2,\dots,n \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Local non-maxima refer to those cells for which  $x_{ic} = 0$ . Figure 1d shows the local maxima for the probability surface in Figure 1a.

The identification of local maxima provides an alternate definition of the buffer layer as the layer derived through the superimposition of the buffer zones surrounding each local maximum. The probability values of each local maximum are radiated horizontally over space to a distance defined by the buffer size. For a given cell on the buffer layer, the probability is then defined as the largest of the probability values

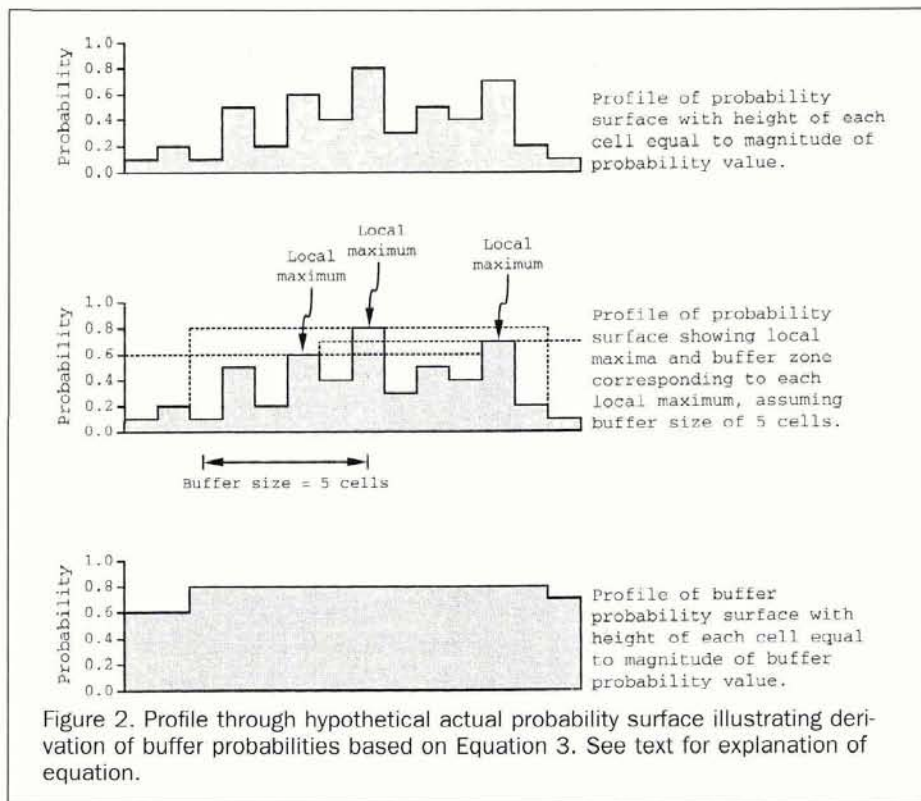
radiated to that cell. Figure 2 provides a schematic representation of the computation of buffer probabilities in this manner, based on a profile drawn through a hypothetical probability surface.

Errors at local maxima have a significant impact on buffer layer accuracy. The concept is similar to the blanket of error defining the difference between unclassified and classed attribute values in choropleth maps (Jenks and Caspall, 1971). The amount of error in buffer probabilities depends on the degree to which local maxima are over- or under-estimated. For a given cell  $i$ , assume that the source probability value,  $p_{ic}$ , is an estimate of the true probability value,  $p_{ic}^*$ . The error,  $e_{ic}$ , can be defined by the difference

$$e_{ic} = p_{ic} - p_{ic}^* \quad (5)$$

Over-estimation of a probability value means that  $e_{ic} > 0$ , while under-estimation implies that  $e_{ic} < 0$ . When a local maximum is over-estimated, the over-estimated component  $e_{ic}$  is radiated over space around the local maximum to a distance defined by the buffer size (Figure 3a). The component is added to the probability of every cell for which the over-estimated local maximum has the largest probability of all probability values radiated to that cell. The total volume of





error is equal to the product of  $e_{ic}$  and the number of cells that have this characteristic. Thus, an increase in buffer size implies that more cells on the buffer layer will contain the over-estimated probability values, resulting in a greater volume of error.

Under-estimation of local maxima has essentially the same effect, except in cases in which the under-estimation causes a cell to lose its status as a local maximum (Figure 3b). In this case, buffer error is likely to be somewhat less than for the same degree of over-estimation.

For a local non-maximum, over-estimation of source probabilities is again more significant than under-estimation. Over-estimation can cause a cell to become a local maximum, which results in propagation of the over-estimated component (Figure 3c). In contrast, under-estimation for local non-maxima has no impact on buffer error (Figure 3d), because the probability value is already lower than another nearby cell.

The effects of over- and under-estimation of source probabilities are mitigated by buffer size. There are two competing forces at work. On the one hand, error is propagated to greater distances as buffer size increases. For larger buffers, error is propagated to a larger number of cells, resulting in a greater volume of error in the buffer layer. On the other hand, errors associated with a local maximum cannot be propagated indefinitely over space. The buffer layer tends to become saturated as buffer size increases, because large buffers have a tendency to overlap. As saturation begins to occur, error no longer accumulates independently from each local maximum. Rather, error propagated from one local maximum tends to be swamped by error propagated from another. Thus, as buffer size increases past some threshold, the rate of error propagation begins to decline. The ultimate cap on propagated error is determined by the size of the grid. When the buffer exceeds the boundaries of the grid, error propagation rates become flat.

These general observations are affected by spatial covar-

iation in probability values. Spatial covariation has an impact on the degree of over- and under-estimation and the propensity for saturation to occur. These effects may be mitigated by covariation in errors, because errors affect the spatial distribution of probabilities on the buffer layer.

## Analysis

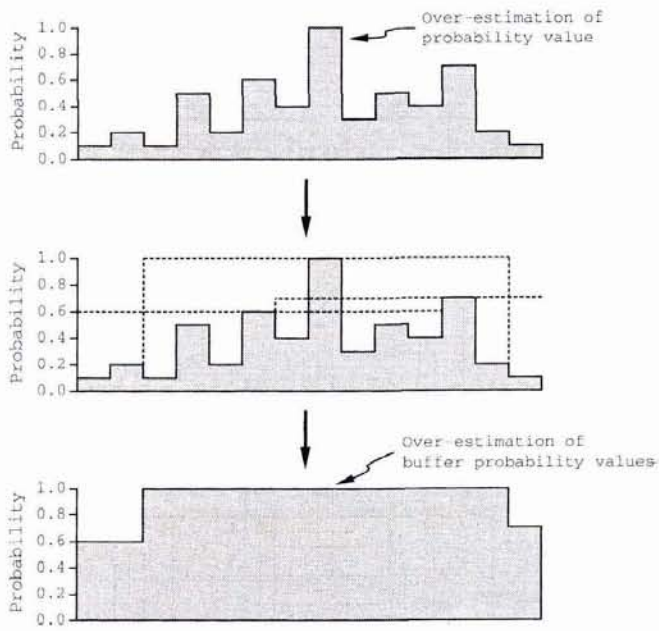
### Hypotheses

Based on the above considerations, the following hypotheses may be advanced with regard to error propagation for the buffer operation in the context of probability surfaces:

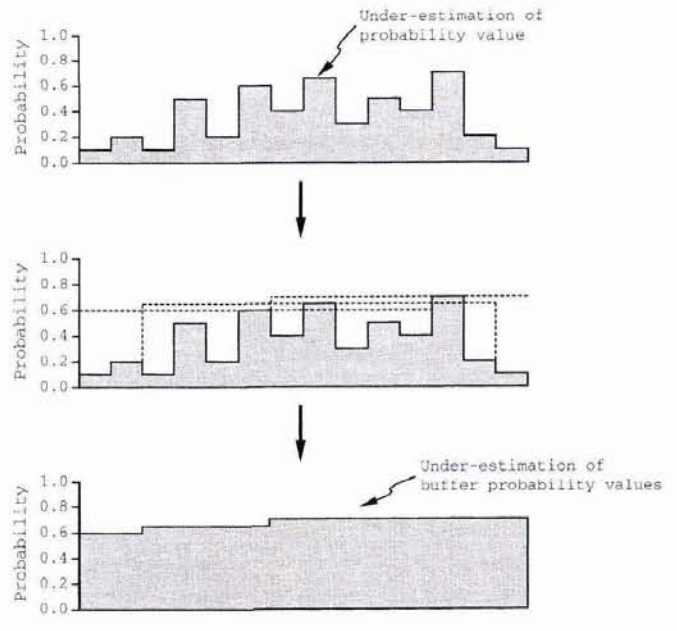
- There should be a positive relationship between the amount of error in source probability values and the amount of error in derived buffer probabilities.
- The amount of error in the buffer layer should depend on buffer size. The rate of error propagation should initially increase with buffer size as error is propagated to an ever larger number of cells. As saturation begins to occur, the rate of error propagation should begin to decline.
- Error indices that differentiate between over- and under-estimation (e.g., bias or mean error) should yield more consistent predictions of buffer error than indices that do not differentiate between these two types of error (e.g., root-mean-squared error or RMSE). This is because over- and under-estimation tend to have different effects on buffer error.
- Over-estimation of probabilities should have a greater effect on buffer errors than under-estimation. As described above, under-estimation has little effect on buffer error in many cases.
- Effects associated with spatial covariation in probabilities should also exert an impact on buffer error.

### Methods

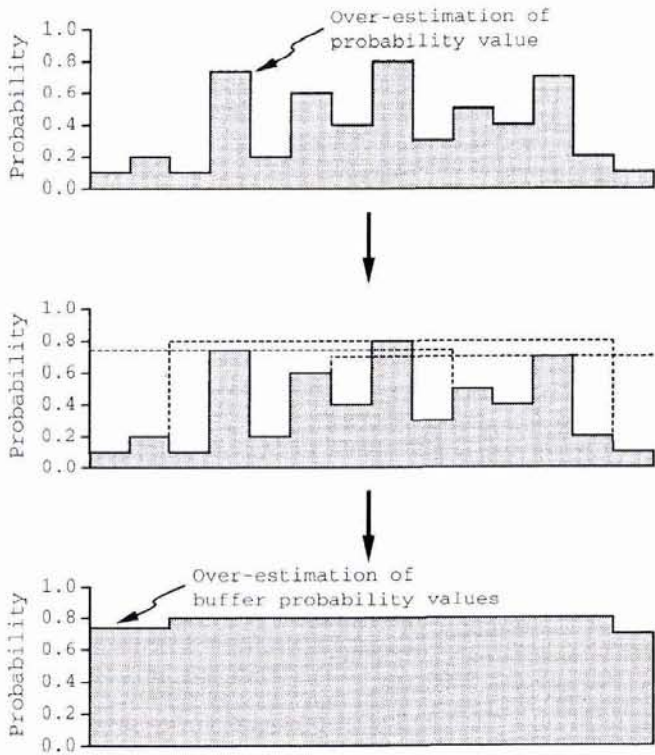
In this study, a simulation-based procedure is used to examine error propagation effects for the buffer operation. The procedure is based on a comparison of the error characteristics of multiple realizations of "actual" and "estimated"



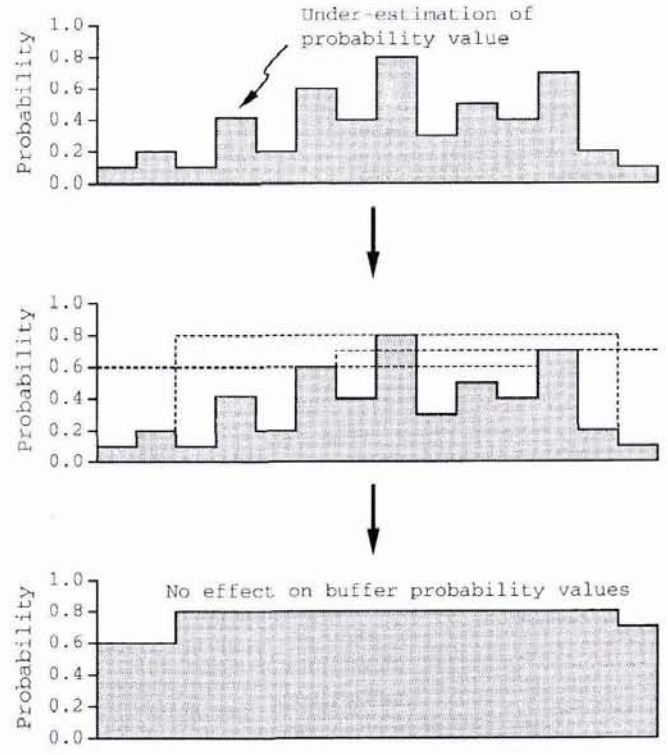
(a)



(b)



(c)



(d)

Figure 3. Profile through hypothetical probability surface, illustrating effects of over- and under-estimation of probabilities for buffer derivation (to be compared with actual distribution in Figure 2). (a) Over-estimation of local maximum causes over-estimated component to be radiated to neighboring cells. (b) Under-estimation of local maximum causes under-estimation of buffer probabilities for some neighboring cells. (c) Over-estimation of local non-maximum can cause error to be propagated if cell becomes local maximum. (d) Under-estimation of local non-maximum has no effect on propagated error.



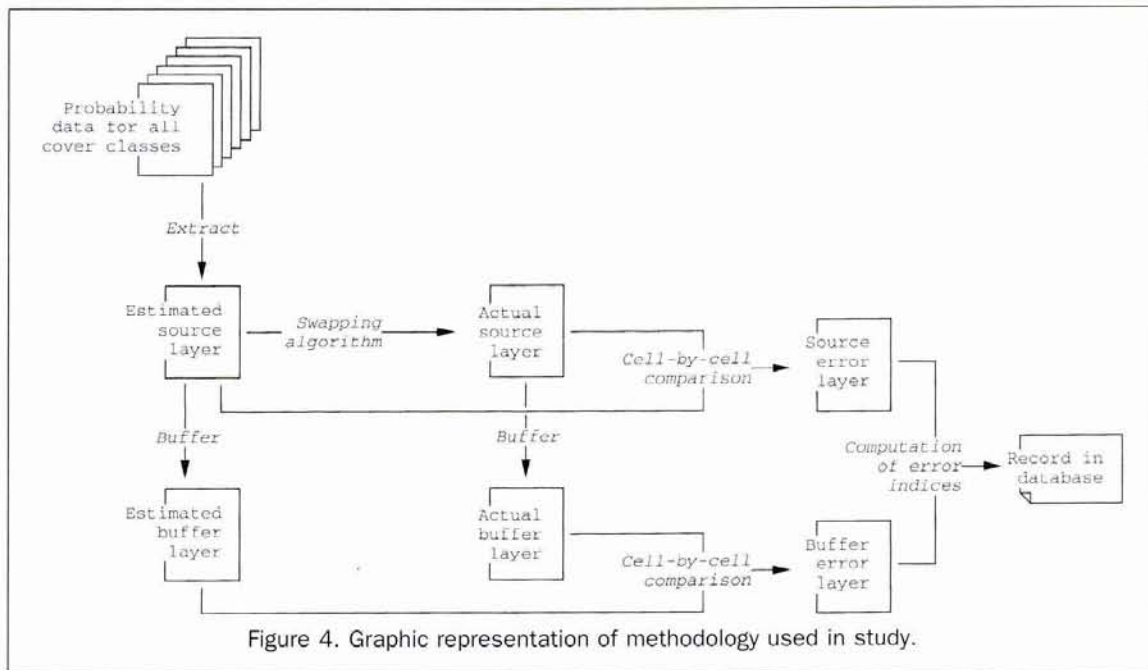


Figure 4. Graphic representation of methodology used in study.

source and buffer layers. For the purposes of this study, the following definitions are used:

- The estimated source layer refers to the set of cell probabilities for a particular cover class, as derived from some classification procedure. This is the layer that would normally be available to a GIS user. The estimated source layer is assumed to be an imperfect estimator of the actual source layer.
- The actual source layer is the theoretical (but operationally unobservable) error-free counterpart of the estimated source layer. In this study, actual source layers are derived through a transformation of the corresponding estimated source layer, subject to certain conditions (as described below).
- The estimated and actual buffer layers are the layers derived by applying Equation 3 to the estimated and actual source layers, respectively.

Data for the simulation procedure were obtained from a 120-row by 160-column subset of a Landsat Thematic Mapper scene for Kent, Ohio. The fuzzy c-means classifier was applied to these data to obtain probabilities for six land-cover classes. The original FORTRAN version of the c-means classifier can be found in Bezdek *et al.* (1984). For this study, the classifier was translated into C.

A schematic representation of the steps involved in the procedure is given in Figure 4. These steps are described in detail below.

- Step 1. Obtain probability values,  $p_{ic}$ , for some class,  $c$ , for all cells in the estimated layer. As noted above, the estimated layer is derived using a classification procedure, and is assumed to be an imperfect representation of the actual layer.
- Step 2. Compute  $\gamma$  (the semi-variance) for the estimated layer at a spatial lag of one cell (approximately 30m). Large values of  $\gamma$  indicate that adjacent cells tend to have different probability values (high spatial covariation), while small values of  $\gamma$  indicate that adjacent cells tend to have similar values (low spatial covariation).
- Step 3. Compute a "target" value of  $\gamma$  for the actual layer. The target level is computed by selecting a random number between 0.0 and 0.03 from a uniform distribution. (These extreme values are equidistant from the mean  $\gamma$  value of 0.015 observed for the six land-cover classes.)
- Step 4. Derive the actual layer for comparison with the esti-

ated layer. The probability values for the actual layer,  $p_{ic}^*$ , are initially set to equal those on the estimated layer. Values are then modified to achieve the target value of  $\gamma$ . The procedure is an iterative one in which a cell is selected randomly and its probability then redefined by adding or subtracting a random scalar. Only if this change brings the layer closer to the target level of  $\gamma$  is the change preserved. The procedure continues until the target level of  $\gamma$  is achieved. The procedure is similar to the "swapping" algorithm described by Veregin (1994).

- Step 5. Apply the buffer operator (Equation 3) to the actual layer to derive the actual buffer layer, and to the estimated layer to obtain the estimated buffer layer. Buffer sizes range from 300 m to 1500 m. Larger buffer sizes tend to cause saturation of buffer sizes, due to the relatively small size of the study area.
- Step 6. Compute source and buffer error. Source error is defined in terms of the differences between estimated and actual probability values, as defined in Equation 5. Buffer error is defined in an analogous manner.
- Step 7. Compute the value of  $\gamma$  for source error.
- Step 8. Compute the error indices. In this study, two standard indices are used. The bias index is sensitive to the direction of error, while the RMSE (root-mean-squared error) index is not.

$$\text{RMSE} = \left[ \frac{1}{n} \sum_{i=1}^n e_{ic}^2 \right]^{1/2} \quad (6)$$

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n e_{ic} \quad (7)$$

This sequence of steps is performed repeatedly for different combinations of buffer sizes, target  $\gamma$  levels, and cover classes. Each combination produces a record in a dataset that is later analyzed statistically. The present study employs five different buffer sizes (ranging from 300 m to 1500 m in increments of 300 m), ten different target  $\gamma$  levels (selected randomly from a uniform distribution with a range of 0.0 to 0.03), and six different classes (i.e., the classes derived from the fuzzy c-means classifier). The result is a database with 300 observations. Each observation contains values for each of the following variables:



- *Cover class*: class derived from fuzzy c-means classifier,
- *Buffer size*: width of buffer (in metres).
- *Source RMSE*: RMSE value for actual and estimated source layers,
- *Source bias*: bias value for actual and estimated source layers,
- *Buffer RMSE*: RMSE value for actual and estimated buffer layers,
- *Buffer bias*: bias value for actual and estimated buffer layers,
- *Estimated  $\gamma$* : semi-variance for estimated source layer at spatial lag of one cell,
- *Actual  $\gamma$* : semi-variance for actual source layer at spatial lag of one cell, and
- *Error  $\gamma$* : semi-variance for source error layer at spatial lag of one cell.

## Results

Results of correlation and regression analysis applied to the derived dataset are summarized below.

- In general, there are strong and significant correlations between source and buffer errors across classes and buffer sizes (Table 1). As expected, these correlations are positive, indicating that buffer error tends to increase with an increase in source error.
- As expected, stronger correlations are evident for the bias index than for the RMSE index (Table 1). Correlation results for RMSE are less consistent than those for bias. Occasional negative relationships for source and buffer RMSEs (e.g., for class A for buffer sizes over 600 m) occur because RMSE (in contrast to bias) fails to differentiate between over- and under-estimation of class probabilities.
- Regression lines for buffer error on source error have slopes greater than one, indicating that the degree of buffer error is larger than the degree of source error, and that this difference tends to be magnified as source error increases (Figures 5a and 5b).
- Positive bias in the source layer (indicating over-estimation of

probabilities) produces positive bias in the buffer layer. Negative bias in the source layer (indicating under-estimation of probabilities) produces negative bias in the buffer layer. As expected, over-estimation of source probabilities has a greater impact on buffer error. Given the same level of positive and negative bias in the source, the level of positive bias in the buffer (associated with positive bias in the source) tends to be greater than the level of negative bias in the buffer (associated with negative bias in the source).

- Relationships between source and buffer error are affected by covariation in source and error probabilities. Covariation (measured in terms of  $\gamma$ ) contributes significantly to the explanatory power of regression models of buffer error on source error assessed over all cover classes simultaneously. Regression models incorporating  $\gamma$  computed for the actual layer are presented in Tables 2 and 3. Table 2 presents the linear regression results for the RMSE index for all cover classes combined. One equation is given for each buffer size from 300 to 1500 m. The regression equations define the RMSE for the buffer layer as a function of the RMSE of the source layer and the level of covariation in the actual source layer probabilities. Table 3 presents the linear regression results for the bias index for all cover classes combined. The regression equations define the bias for the buffer layer as a function of bias in the source layer and the level of covariation in the actual source layer probabilities. For both the RMSE and bias indices, there is a negative relationship between buffer error and  $\gamma$ . This appears to be due to the fact that low values of  $\gamma$  are associated with a tendency to over-estimate probability values. As described above, over-estimation has a more significant impact on buffer error than under-estimation, which is associated with high values of  $\gamma$ .
- Relationships between source and buffer error are also affected by buffer size. For the RMSE index, the amount of buffer error tends to decline uniformly as buffer size increases, for any source RMSE level (Figure 6a). For the bias index, the effect of buffer size depends on the level of source bias. If

TABLE 1. CORRELATION COEFFICIENTS (PEARSON PRODUCT-MOMENT) BETWEEN SOURCE AND BUFFER ERROR. COEFFICIENT VALUES ARE GIVEN FOR BOTH ERROR INDICES (RMSE AND BIAS) FOR ALL SIX COVER CLASSES (LABELED A THROUGH F) FOR VARIOUS BUFFER SIZES (300 THROUGH 1500 M). VALUES IN PARENTHESES FOLLOWING CORRELATION COEFFICIENTS ARE P-VALUES INDICATING COEFFICIENT SIGNIFICANCE (LOWER P-VALUES INDICATE HIGHER LEVELS OF SIGNIFICANCE).

Class	Index	Buffer Size									
		300m		600m		900m		1200m		1500m	
A	RMSE	0.621	(0.055)	-0.461	(0.180)	-0.468	(0.172)	-0.455	(0.186)	-0.446	(0.197)
	Bias	0.982	(<0.001)	0.960	(<0.001)	0.935	(<0.001)	0.909	(<0.001)	0.898	(<0.001)
B	RMSE	0.964	(<0.001)	0.946	(<0.001)	0.927	(<0.001)	0.912	(<0.001)	0.894	(<0.001)
	Bias	0.972	(<0.001)	0.968	(<0.001)	0.962	(<0.001)	0.956	(<0.001)	0.950	(<0.001)
C	RMSE	0.950	(<0.001)	0.920	(<0.001)	0.907	(<0.001)	0.898	(<0.001)	0.884	(<0.001)
	Bias	0.982	(<0.001)	0.971	(<0.001)	0.964	(<0.001)	0.956	(<0.001)	0.945	(<0.001)
D	RMSE	0.617	(0.057)	0.576	(0.081)	0.572	(0.084)	0.570	(0.085)	0.571	(0.085)
	Bias	0.982	(<0.001)	0.971	(<0.001)	0.962	(<0.001)	0.954	(<0.001)	0.947	(<0.001)
E	RMSE	0.848	(0.002)	0.836	(0.003)	0.825	(0.003)	0.812	(0.004)	0.802	(0.005)
	Bias	0.982	(<0.001)	0.970	(<0.001)	0.956	(<0.001)	0.940	(<0.001)	0.926	(<0.001)
F	RMSE	0.856	(0.002)	0.839	(0.002)	0.834	(0.003)	0.830	(0.003)	0.830	(0.003)
	Bias	0.972	(<0.001)	0.959	(<0.001)	0.947	(<0.001)	0.935	(<0.001)	0.921	(<0.001)

TABLE 2. LINEAR REGRESSION RESULTS FOR RMSE INDEX FOR ALL COVER CLASSES COMBINED. ONE EQUATION IS GIVEN FOR EACH BUFFER SIZE FROM 300 TO 1500 M. VALUES IN PARENTHESES BELOW REGRESSION COEFFICIENTS ARE P-VALUES INDICATING COEFFICIENT SIGNIFICANCE. VALUES IN PARENTHESES BELOW R<sup>2</sup> COEFFICIENTS ARE P-VALUES INDICATING OVERALL SIGNIFICANCE OF REGRESSION EQUATION.

Buffer size	Regression equation	R <sup>2</sup>
300	Buffer RMSE = 0.072 + 1.996 Source RMSE - 4.619 Actual $\gamma$ (<0.001) (<0.001)	0.719 (<0.001)
600	Buffer RMSE = 0.224 + 0.803 Source RMSE - 12.110 Actual $\gamma$ (<0.001) (<0.001)	0.879 (<0.001)
900	Buffer RMSE = 0.217 + 0.608 Source RMSE - 12.41 Actual $\gamma$ (0.010) (<0.001)	0.874 (<0.001)
1200	Buffer RMSE = 0.191 + 0.593 Source RMSE - 11.86 Actual $\gamma$ (0.015) (<0.001)	0.855 (<0.001)
1500	Buffer RMSE = 0.169 + 0.592 Source RMSE - 11.12 Actual $\gamma$ (0.017) (<0.001)	0.836 (<0.001)



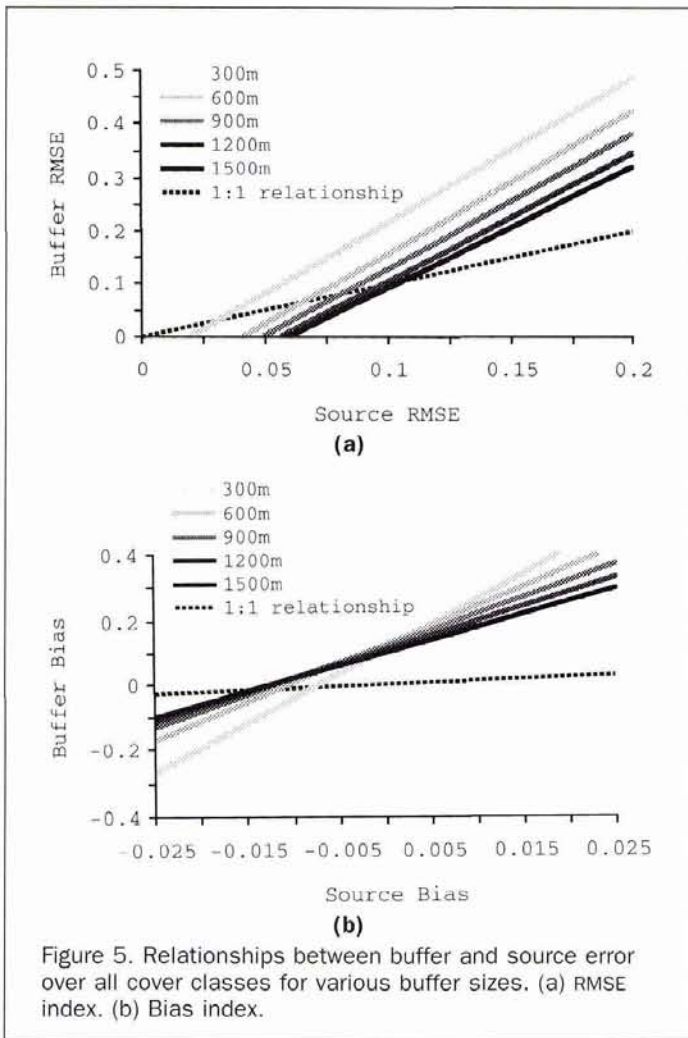


Figure 5. Relationships between buffer and source error over all cover classes for various buffer sizes. (a) RMSE index. (b) Bias index.

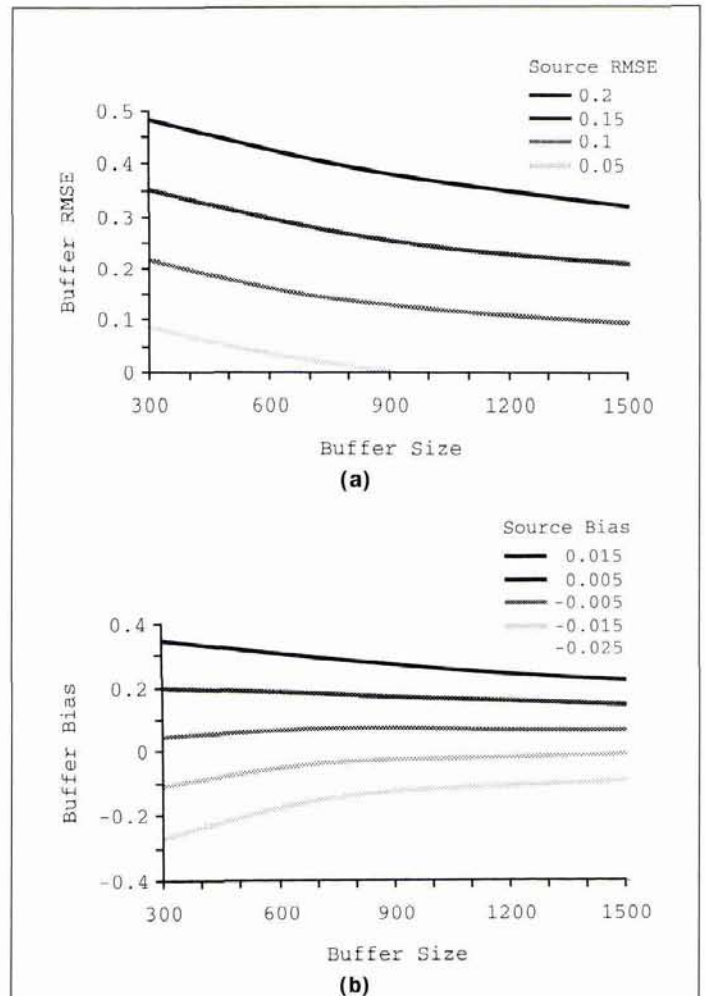


Figure 6. Relationships between buffer error and buffer size over all cover classes for various levels of source error. (a) Relationship between buffer RMSE and buffer size. (b) Relationship between buffer bias and buffer size.

source bias is relatively high, buffer error tends to decline uniformly as buffer size increases. If source bias is low, buffer error initially rises and then declines as the saturation effect described above begins to exert an effect (Figure 6b).

- Levels of buffer error measured over all cover classes and buffer sizes can be predicted reliably based on various characteristics of the source data. Table 4 gives the linear regression results for the RMSE and bias indices. One equation is given for each index, computed over all buffer sizes and cover classes. Buffer RMSE is predicted in terms of the RMSE of the source layer, buffer size, and the level of covariation in actual source layer probabilities. (The level of covariation in the estimated or error layer can also be used, but at the expense of some explanatory power.) For the bias index, buffer

size is insignificant, and bias is predicted from source bias and the level of covariation alone. (Buffer size is likely to be significant in situations in which a wider range of buffer sizes is employed.) Polynomial regression models fail to provide significantly more explanatory power than do linear models.

## Conclusions

Error propagation for the buffer operation in the context of probability surfaces has much in common with error propa-

TABLE 3. LINEAR REGRESSION RESULTS FOR BIAS INDEX FOR ALL COVER CLASSES COMBINED. ONE EQUATION IS GIVEN FOR EACH BUFFER SIZE FROM 300 TO 1500 M.

Buffer size	Regression equation	R <sup>2</sup>
300	Buffer bias = 0.264 + 7.684 Source bias - 15.66 Actual $\gamma$ ( $<0.001$ ) ( $<0.001$ )	0.976 ( $<0.001$ )
600	Buffer bias = 0.293 + 3.499 Source bias - 17.450 Actual $\gamma$ ( $<0.001$ ) ( $<0.001$ )	0.954 ( $<0.001$ )
900	Buffer bias = 0.264 + 2.586 Source bias - 15.45 Actual $\gamma$ (0.003) ( $<0.001$ )	0.962 ( $<0.001$ )
1200	Buffer bias = 0.237 + 2.208 Source bias - 13.65 Actual $\gamma$ (0.013) ( $<0.001$ )	0.896 ( $<0.001$ )
1500	Buffer bias = 0.216 + 1.980 Source bias - 12.37 Actual $\gamma$ (0.033) ( $<0.001$ )	0.865 ( $<0.001$ )



TABLE 4. LINEAR REGRESSION RESULTS FOR RMSE AND BIAS INDICES. ONE EQUATION IS GIVEN FOR EACH INDEX, AND IS COMPUTED FOR ALL BUFFER SIZES AND COVER CLASSES COMBINED.

Regression equation	R <sup>2</sup>
Buffer RMSE = 0.267 + 0.918 Source RMSE - 1.03×10 <sup>-4</sup> Buffer size - 10.420 Actual $\gamma$ (<0.001) (0.001) (0.001)	0.828 (0.001)
Buffer bias = 0.255 + 3.591 Source bias - 14.920 Actual $\gamma$ (0.001) (0.001)	0.890

gation for conventional raster data. In both cases, source and buffer errors are positively correlated, with a decline in the rate of error propagation as buffer size increases. For both probability based and conventional data, error propagation is affected by spatial covariation.

In contrast to conventional raster data, however, levels of buffer error for probability surfaces tend to be larger than levels of source error. The simulation results reported in this study suggest that the level of error in a derived buffer layer can easily be several times larger than the level of error in the source layer from which the buffer was derived. The buffer operation can cause source errors to become magnified under certain conditions, including

- a tendency to over-estimate source probabilities (resulting in the propagation of the over-estimated component, especially at local maxima);
- the presence of high levels of spatial covariation in source probability values (resulting in the tendency to over-estimate source probability values); and
- use of small buffer sizes (resulting in a high error accumulation rate, due to a minimal degree of buffer zone overlapping and saturation).

These results suggest that decisions based on probabilistic interpretations of class membership can be significantly impacted by errors in source data. The reliability of buffers derived from probability based data should be interpreted in the light of these observations.

## References

- Bezdek, J.C., R. Ehrlich, and W. Full, 1984. FCM: The fuzzy c-means clustering algorithm, *Computers and Geosciences*, 10:191-203.
- Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment*, Clarendon, Oxford.
- Burrough, P.A., R.A. MacMillan, and W. van Deursen, 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography, *Journal of Soil Science*, 43: 193-210.
- Carver, S., 1991. Adding error handling functionality to the GIS toolkit, *Proceedings EGIS '91*, pp. 187-196.
- Fisher, P.F., 1994. Hearing the reliability in classified remotely sensed images, *Cartography and Geographic Information Systems*, 21:31-36.
- Fisher, P.F., and S. Pathirana, 1990. The evaluation of fuzzy membership of land cover classes in the suburban zone, *Remote Sensing of Environment*, 34:121-132.
- Footy, G.M., 1994. Ordinal-level classification of sub-pixel tropical forest cover, *Photogrammetric Engineering & Remote Sensing*, 60:61-65.
- Heuvelink, G.B.M., P.A. Burrough, and A. Stein, 1989. Propagation of errors in spatial modelling with GIS, *International Journal of Geographical Information Systems*, 3:303-322.
- Goodchild, M.F., S. Guoqing, and Y. Shiren, 1992. Development and test of an error model for categorical data, *International Journal of Geographic Information Systems*, 6:87-104.
- Jenks, G.F., and F.C. Caspall, 1971. Error in choroplethic maps: Definition, measurement, reduction, *Annals of the Association of American Geographers*, 61:217-244.
- Lanter, D., and H. Veregin, 1990. A lineage meta-database program for propagating error in geographic information systems, *GIS/LIS '90 Proceedings*, pp. 144-153.
- , 1992. A research paradigm for propagating error in layer-based GIS, *Photogrammetric Engineering & Remote Sensing*, 58: 526-533.
- Leung, Y., 1988. *Spatial Analysis and Planning under Imprecision*, North Holland, Amsterdam.
- Newcomer, J.A., and J. Szajgin, 1984. Accumulation of thematic map errors in digital overlay analysis, *The American Cartographer*, 11:58-62.
- Robinson, V.B., and D. Thongs, 1985. Fuzzy set theory applied to the mixed pixel problem of multispectral land cover databases, *Geographic Information Systems in Government* (B. Opitz, editor), pp. 871-886.
- Tikunov, V.S., 1986. *Some Issues on the Modeling Approach to Cartography*, unpublished manuscript (available from author).
- Veregin, H., 1989. Error modeling for map overlay, *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors), Taylor and Francis, Basingstoke, pp. 3-18.
- , 1994. Integration of simulation modeling and error propagation for the buffer operation in GIS, *Photogrammetric Engineering & Remote Sensing*, 60:427-435.
- Veregin, H., and F. Sultana, 1993. Resolution-dependent effects on classification accuracy in remote sensing, *Proceedings, 16th Annual Applied Geography Conference*, pp. 24-31.
- Wesseling, C.G., and G.B.M. Heuvelink, 1991. Semi-automatic evaluation of error propagation in GIS operations, *Proceedings EGIS '91*, pp. 1228-1237.
- Zadeh, L.A., 1965. Fuzzy sets, *Information and Control*, 8:338-353.

(Received 13 August 1993; accepted 10 August 1994; revised 29 November 1994)

**To receive your free copy of the ASPRS Publications Catalog, write to:**

ASPRS, Attn: Julie Hill, 5410 Grosvenor Lane, Suite 210,  
Bethesda, MD 20814-2160; jhill@asprs.org.

Please indicate your area of interest:  
Photogrammetry, Remote Sensing, GIS, Cartography, all of the above.