A Research Paradigm for Propagating Error in Layer-Based GIS

David P. Lanter and *Howard Veregin**
Department of Geography/NCGIA, University of California, Santa Barbara, CA 93106

ABSTRACT: This paper focuses on the nature of error in spatial databases and the implications of this error for spatial data transformations in GIS applications. It describes an error propagation research paradigm as an information flow linking successively more formal components of error propagation in a GIS context. These components include development of conceptual models of error, creation of formal indices to measure error in spatial databases, implementation of mathematical functions to transform error indices and model the propagation of error as it is processed, and evaluation of the indices to gain insight into the utility of conceptual models used in error measurement and propagation. The paradigm enables researchers to formulate, manipulate, and experiment with components of error propagation to determine their implications for decision making. The applicability of the paradigm is illustrated with a simple GIS application in which error is propagated from sources to final product through a sequence of data transformation functions.

INTRODUCTION

GEOGRAPHIC INFORMATION SYSTEMS PROVIDE USERS WITH convenient and consistent mechanisms for applying automated transformation functions to manipulate and analyze spatial data. These capabilities expand the role and increase the value of spatial databases used in a variety of decision-making contexts. Such systems, however, often lack capabilities for establishing the accuracy and validity of products derived to support decisions. That is, a GIS provides a means of deriving new information without simultaneously providing a mechanism for establishing its reliability. The literature detailing GIS applications shows that there is a lack of concern for error in spatial databases and its propagation through sequences of data transformation functions. In such applications input data quality is often not ascertained, functions are applied to these data without regard for the accuracy of derived products, and these products are presented without an associated estimate of their reliability or an indication of the types of error they may contain.

Such omissions do not imply that errors are of such low magnitude that they can simply be ignored. Rather, they reflect the lack of a standard framework for modeling how error is propagated through sequences of data transformation functions. Paradoxically, an enormous volume of research has been carried out on the question of spatial database accuracy and the errors introduced by various types of data transformation (Goodchild and Gopal, 1989; Veregin, 1989a). Numerous indices have been developed to measure spatial and aspatial dimensions of error in databases, and methods have been proposed for modeling the ways in which data transformation functions modify and introduce error. Much of this research, however, has been carried out in isolation from the broader context of error propagation modeling in a GIS environment. There is a lack of a methodology for specifying the interactions among these various error indices and models of error propagation. That is, there is no accepted paradigm for modeling error propagation that explicitly recognizes the interdependence between basic concepts of spatial database accuracy and formal methods of error propagation in an actual system.

Figure 1 illustrates an informational flow linking successively more formal components of error propagation modeling and is presented as a possible error propagation research paradigm. The conceptual model of error reflects notions of what error signifies in a particular context. This ontological issue is of fundamental importance because error in spatial databases is inherently multi-dimensional. The utility of different dimensions of error is a function of context defined by the requirements of the uses and the classes of geographical data under consideration. Once determined, significant dimensions of error must be represented numerically as an index or set of indices for error measurement. This permits error propagation to be implemented by an error propagation function. Such functions model how a particular type of error is modified as spatial data are processed by a given data transformation function. Automated error propagation functions can be used to track errors present in source data through specific sequences of data transformation functions to determine the quality of a GIS derived data product.

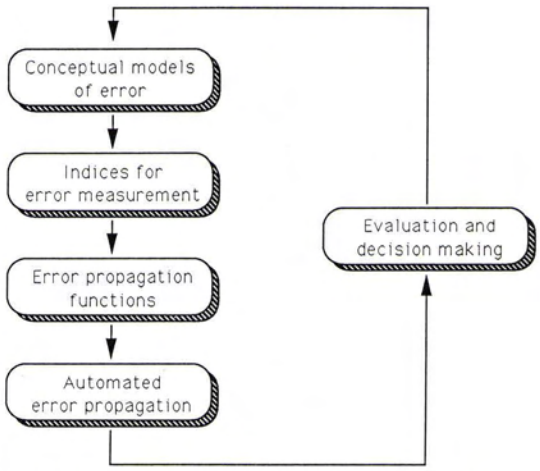

FIG. 1. An error propagation research paradigm.

The sections that follow discuss conceptual models of error for geographic data, indices to measure those errors, and functions to propagate the indices in a GIS application. Error propagation research is facilitated by a computer program for testing error indices and error propagation functions. The program utilizes a meta-data model of a GIS application allowing users to characterize data sources with error indices and implement

*Presently with the Dept. of Geography, Kent State University, Kent, OH 44242.

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING,
Vol. 58, No. 6, June 1992, pp. 825–833.

0099-1112/92/5806-825$03.00/0

functions to propagate these indices through the course of GIS applications processing. This enables researchers to experiment with error measurement indices and functions to propagate them. Such experimentation provides feedback that aids in identifying limitations of existing indices and propagation methods. This serves to elucidate problems associated with error propagation for a particular conceptual model of error.

CONCEPTUAL MODELS OF ERROR

Numerous conceptual models of error exist for spatial data. Error is perhaps most commonly conceived of as a deviation from some accurate, real-world standard, such that any measurement is assumed to represent an approximation of some true but unknown value. This conception of error has much in common with the statistical treatment of error in terms of bias and precision. It also implies that error is largely a function of the reliability of data acquisition methods, such that the distribution of error can be characterized by obtaining repeated measurements of the same phenomenon, and more accurate measurements can be obtained by using more sensitive instruments.

Alternatively, error may be viewed as inherent uncertainty in some abstracted characteristic of the real world. The map or other spatial data product is not intended as an accurate description of the real world, but as an abstract representation of some characteristic of the world. Uncertainty results from indeterminacy in the spatial distribution of this characteristic, because no accurate real-world standard exists against which the mapped characteristic can be compared. Nor is such uncertainty necessarily inadvertent. It may, for example, result from generalization methods used to enhance the graphic representation of the characteristic of interest. According to this view, a map or other spatial data product is a model of the real world, necessarily incomplete and generalized. Uncertainty is therefore propagated and transformed each time a conceptual or physical model is constructed in the course of GIS applications processing (Bedard, 1987).

Geographical entities are defined in terms of spatial, thematic, and temporal dimensions (Figure 2), and each dimension can be described with corresponding dimensions of error. Errors in spatial data are multi-dimensional in character. That is, spatial databases cannot be characterized adequately with a single index of error. For example, spatial accuracy includes both vertical and horizontal components that are not always separable. Thematic accuracy depends on data type (e.g., numerical versus categorical) and is not always independent of spatial accuracy. Temporal accuracy is an important but often overlooked dimension of accuracy in spatial databases. Data reliability is often (but not always) an inverse function of age, because spatial and thematic attributes may change over time. In addition, older data acquisition methods may be of limited or unknown accuracy.

In surveying and related fields, spatial accuracy is more dominant than thematic accuracy, and a variety of methods exist for measuring accuracy with reference to a precise theoretical standard. In fields that focus instead on the derivation and analysis of thematic information (such as land cover and soil and vegetation type) thematic accuracy plays a much larger role. Indeed, for the class of spatial data known as "categorical coverages" (Chrisman, 1989), the spatial attributes of the data are secondary to, and are determined by, the thematic content itself. Therefore, it may not be meaningful to examine spatial and thematic accuracy independently.

The multi-dimensional character of error in spatial databases is also reflected in the data quality proposed by The Digital Cartographic Data Standards Task Force (DCDSTF, 1988). According to the proposed standard, documentation of data quality includes five key components: positional accuracy, attribute accuracy, lineage, logical consistency, and completeness. Positional and attribute accuracy refer to spatial and thematic components of accuracy, respectively. Lineage refers to the data sources, methods of deriving and encoding the data, and the set of all transformations applied to the data. Logical consistency refers to the fidelity of the relationships encoded in the data. This includes consistency of topology, spatial attributes such as perimeter and area values across hierarchical groupings of polygons, and thematic attributes such as Census population estimates for different aggregations of enumeration units. Completeness describes the relationship between objects in the database and the abstract universe of all objects.

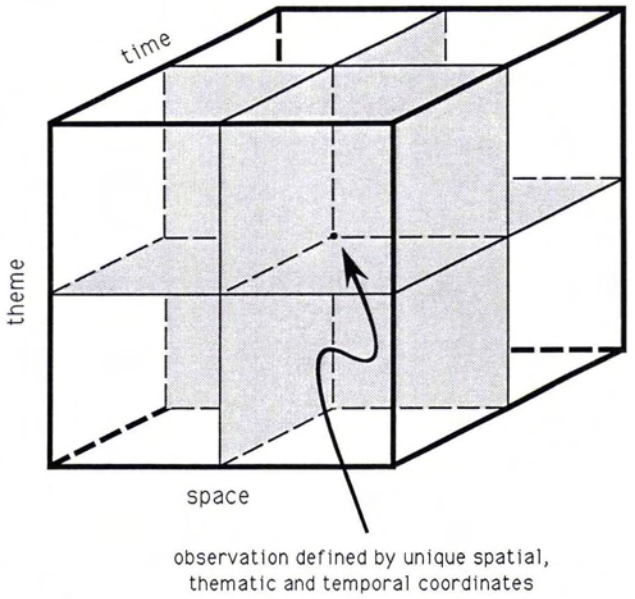

FIG. 2. Geographical data defined in terms of space, theme, and time.

The proposed standard is designed to facilitate "truth in labeling." That is, while it is the responsibility of the data producer to document data quality, it is the user who must interpret this documentation and evaluate the fitness of the data for a particular application. The error propagation paradigm presented here adopts the truth in labeling concept. The paradigm provides a means to propagate error through GIS functions, but does not define what level of error is acceptable in a given situation. Such policy decisions are application-specific and should be based on a consideration of the relevance and significance of different types of error, issues that are largely a function of context.

INDICES FOR ERROR MEASUREMENT

The choice of an index or set of indices to measure and document error in spatial data is dependent on the conceptual model of error that has been determined to be appropriate in a particular context. For example, in surveying and related fields in which spatial accuracy is of paramount importance, error is typically evaluated in terms of the deviations between the actual and estimated locations of a sample of points. These deviations may be measured in the X,Y (i.e., horizontal), and Z (i.e., vertical) dimensions for the points. Often, measurements of error in the X and Y dimensions are collapsed into a single index of horizontal accuracy, as in the case of the National Map Accuracy Standards (NMAS) currently applied to U.S. Geological Survey topographic maps. Using Koppe's formula, which accounts for the effects of terrain slope on mean vertical error, it is also

possible to collapse the horizontal and vertical accuracy components into a single index (Gustafson and Loon, 1981).

Moreover, as the NMAS illustrates, error indices may take the form of a compliance test with an accuracy standard. However, as this form of error index is only a logical test of accuracy, it provides no information about the amount of error at particular sample points. Alternatives to the accuracy standards approach often involve the statistical concepts of bias and precision. Common error indices based on these concepts include mean absolute error, standard error, and root-mean-squared error (RMSE). Again, these may be measured in the X, Y, and Z dimensions, or these dimensions may be collapsed. Error indices may include measurements of both bias and precision (e.g., American Society of Civil Engineers, 1983) or of precision under the assumption that bias is random (Merchant, 1987).

Spatial accuracy may also be measured in terms of statistical functions. For example, the epsilon band concept provides a means of representing the positional error in lines due to digitizing or generalization error (see Burrough, 1986). Typically, a boxcar distribution with a width of twice the value of epsilon and centered on the estimated line is assume to encompass the true line location (e.g., Blakemore, 1983). Others have suggested that the distribution may approximate some other statistical function, such as a Gaussian or bimodal one (e.g., Honeycutt, 1986). These error indices, while appropriate for a conceptual model of error in the statistical sense of inexactness, are not necessarily appropriate for a model of error as inherent uncertainty. In this case, it would be more appropriate to measure error based on notions of spatial variability. Researchers have proposed various error indices that incorporate ideas from fuzzy set, evidential, and probability theory, among others [see Stoms (1987) for a review]. Thus, spatial accuracy may be characterized by one or more components, depending on the purpose for which accuracy evaluation is being carried out.

Measurement of thematic accuracy depends on the type of data under consideration. For categorical data (e.g., land cover, or soil or vegetation type) it is common to compute an index of classification accuracy from a classification error matrix. This matrix is a cross-tabulation of the estimated and actual thematic values for a sample of points. In the classification error matrix, element c_{ij} represents the number of points assigned to class i that actually belong to class j. Perhaps the most common index of classification accuracy derived from the classification error matrix is the proportion of points correctly classified (PCC). The PCC is defined as the trace of the classification error matrix (i.e., the sum of all c_{ij} where $i = j$) divided by the number of sample points. If the sample has been drawn randomly then the PCC may be viewed as the probability that a point selected at random from the layer is correctly classified. Among the alternatives to PCC are the kappa (or khat) statistic, which accounts for correct classifications that occur by chance alone, and user's and producer's accuracies, which focus on the accuracy of individual thematic classes (Ginevan, 1979; Aronoff, 1982; Story and Congalton, 1986; Hudson and Ramm, 1987). There are also numerous alternatives to the classification error matrix approach. For example, one might compare the area of a sample of polygons on a map to their actual area as determined by ground survey (Fitzpatrick-Lins, 1978), or compute the positional error in polygon boundaries arising from classification error (Hord and Brooner, 1976). For numerical thematic data, indices derived from the classification error matrix are inappropriate and some other index, such as the standard or root-mean-squared error, might be constructed.

Temporal error may also be measured according to different criteria. For example, one might differentiate between time error (i.e., the difference between the recorded time of an observation and the actual time) and synopticity error (i.e., the difference between the recorded time of an observation and the real-world time it is assumed to represent) (Stearns, 1968). The former is a form of measurement error resulting from inexact temporal coordinates, while the latter is more akin to sampling bias due to the inability to measure some phenomenon instantaneously or at the exact reference time. While other types of temporal error could also be defined, this dimension of error has not received much attention in the geographic literature.

The brief synopsis presented above indicates that measuring error in spatial databases entails, not a single index, but a set of indices describing various dimensions of spatial, thematic, and temporal error. Within each of these dimensions, a variety of possible error indices may be constructed, depending on what type of error is deemed to be significant given the nature of the data and the physical system under consideration, the type of data processing required, and the purpose for which the error assessment is being carried out. These indices may reflect simple compliance tests (e.g., NMAS), summary statistical measures (e.g., the PCC index of classification accuracy), a vector of indices (e.g., standard error in the X, Y, and Z dimensions), a matrix of values (e.g., the classification error matrix), or a statistical distribution (e.g., the epsilon band). A hypothetical vector of error indices for spatial, thematic, and temporal error for a layer is shown in Figure 3.

The vector shown in Figure 3 contains a single-valued index for each of several different dimensions of error. However, indices need not be single-valued, and might refer instead to other vectors, matrices, statistical functions, or other data layers. Such indices do not necessarily assume that error is distributed uniformly over space, theme, or time. Rather, they may be said to be spatially, thematically, or temporally differentiated models of error to the degree to which they permit error to vary along these three dimensions. For example, a classification error matrix is a thematically differentiated model of error, because it tabulates error separately for each thematic attribute class. In contrast, the PCC index of classification accuracy is a single-valued index derived from the classification error matrix, and as such does not permit thematic differentiation of error. That is, the PCC value does not describe how error levels may vary from class to class.

Spatially differentiated models consider how indices of spatial, thematic, or temporal error are distributed over space. The result is a non-uniform patterning that has significant consequences for error propagation (see Fisher, 1989; Flowerdew, 1989; Goodchild, 1989). In these models, indices of spatial, thematic, or temporal error vary over space. A common example is the "Reliability Diagram" that accompanies some topographic maps. This diagram differentiates the quality of different parts of the map based on the date and method of data collection (Chrisman, 1983). As a result, it is possible to construct temporally differentiated models of error because the contents of a cartographic database may have been collected at different times using different sources and methods.

In a layer-based GIS, geographic features are organized according to a thematic or temporal scheme (Chrisman and Niemann, 1985; Kjerne and Dueker, 1986; Aronson, 1987; Bracken and Webster, 1989). Individual layers are in some sense indi-

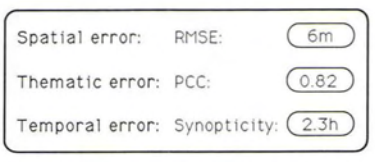

FIG. 3. A hypothetical vector of error indices for a layer.

visible, like algebraic variables. They are 'atoms' from which higher-order constructs (whether algebraic expressions or GIS applications) are build (Tomlin and Berry, 1979). Nevertheless, the ability to define spatially, thematically, or temporally differentiated models of error means that it is not necessary to assume that error is distributed uniformly over space, theme, or time. Error can be divided into different levels along any of these dimensions. As these divisions become fine enough, they become, in effect, attributes of points, lines, areas, or cells.

ERROR PROPAGATION FUNCTIONS

Spatial data transformation functions in a GIS derive new information by making explicit the spatial relationships implicit in source data. The accuracy of this new information depends on the type and level of error present in the sources and the transformations applied to derive the new information. By propagating source errors through data transformation functions, the utility of derived products for decision making can theoretically be established.

An error propagation function may be defined as a mathematical (or otherwise unambiguous) representation of the mechanisms whereby errors present in data sources are modified by a particular data transformation function. In addition, the error propagation function may incorporate the processes whereby the data transformation function itself introduces error where none existed previously. Error propagation functions are therefore process-oriented, as they model changes in error through the course of GIS applications processing. As Figure 4 illustrates, the choice of error propagation function is determined by the GIS data transformation function applied and the index used to measure error in the data input to the transformation function. This assumes some *a priori* knowledge of input data quality and a mathematical model of error propagation mechanisms. For a given GIS data transformation function, there is a vector of error propagation functions (corresponding to a row in Figure 4) that depends on the error measurement index to be propagated. For a given error measurement index there is also a vector of error propagation functions (corresponding to a column in Figure 4) depending on the GIS data transformation through which the error index is to be propagated. Automated error propagation, therefore, involves matching the GIS data transformation function with the error measurement index. The GIS transformation function and error index serve as keys for identifying a cell in Figure 4 and selecting the appropriate error propagation function.

The selection of an error propagation function is also dependent on assumptions about error propagation. These assumptions include the nature of the errors present in the source data, the spatial distribution of the source errors, and the degree to which errors co-occur spatially on different layers. This is represented in Figure 4 in terms of a set of "planes", each containing a matrix of error propagation functions for different combinations of GIS functions and error indices. In short, there is more than one error propagation function possible per combination, each representing a different set of assumptions about error propagation. A fundamental characteristic of the paradigm presented here is that alternate error propagation functions can be selected based on the context within which one is working. In this way, the paradigm serves as a framework for exploring different assumptions about error propagation and permits new functions to be incorporated as they become available.

LAYER-BASED ERROR PROPAGATION

One benefit of layer-based GIS is that provides users with an intuitive conceptual model that facilitates the visualization of geographic features as organized thematic map separates. This allows the application of GIS operators, such as spatial neighborhood, overlay, and attribute manipulation functions, to derive new geographic themes. For example, a GIS application to extract areas where oak regeneration is at risk from cattle grazing is illustrated in Figure 5. Oak woodlands at risk in this example are implicit in the registered source layers: LANDUSE, PERMITS, and VEGETATION. Data transformations applied in this application link each input map to an output map layer. The result is a network linking the application's source maps (LANDUSE, PERMITS, and VEGETATION) to its product (AT_RISK).

The network of input and output relations between spatial data layers is an example of a data flow diagram (Martin and McClure, 1985). Geographers have referred to this data flow diagram as an application's "cartographic model" (Tomlin and Berry, 1979; Berry, 1987). The cartographic model illustrates the propagation of spatial data from source materials to application product. Functions for propagating an index of data error have been incorporated in GEOLINEUS, a lineage information program for GIS (Lanter, 1991). The system is implemented in the LISP

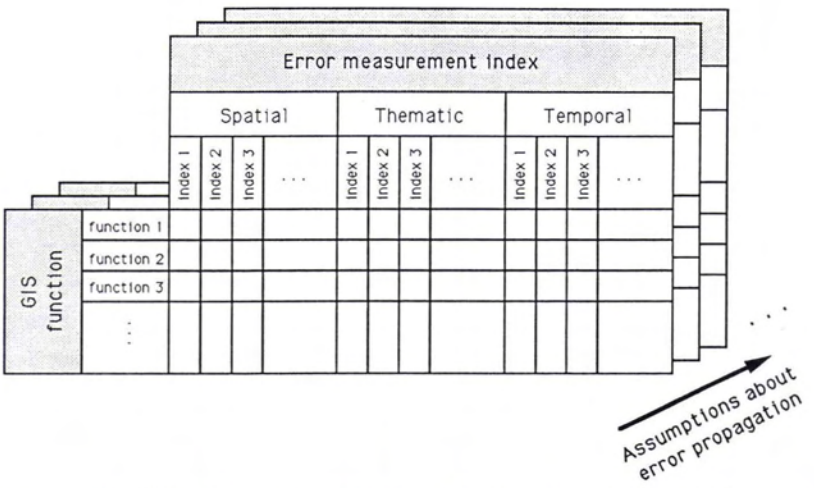

FIG. 4. Each cell in the matrix references a specific error propagation function designed to propagate a specific error index (column) through a particular GIS function (row) based on a set of assumptions about error propagation (plane).

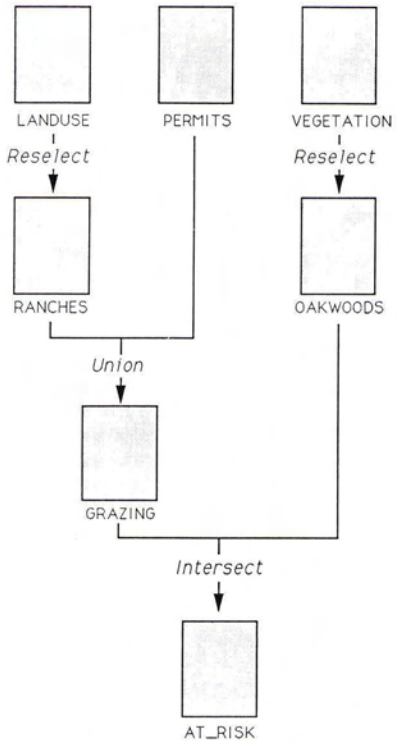

FIG. 5. A GIS application for identifying oak woodlands at risk from grazing activity.

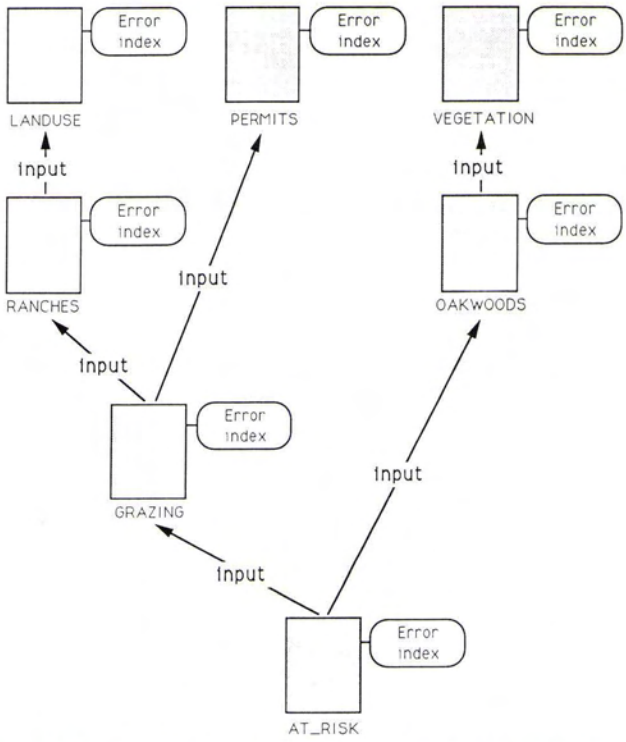

FIG. 6. Input links and error indices for layers in the GIS application.

(LISt Processing) programming language and integrated with ESRI's ARC/INFO running under the AIX operating system on an IBM RS/6000 and under UNIX on a SUN SPARCstation. The LISP language is a programming language modeled after McCarthy's (1960, 1963) calculus of symbolic expressions. GEOLINEUS maintains a lineage knowledge representation reflecting the order of data propagation found in a GIS application's cartographic model in its meta-database (Lanter, 1990). To automate error propagation, thematic data layers are represented in the knowledge representation as nodes connected by "input" links pointing from derived layers back toward their input layers. Each layer is associated with an error index. The resulting structure represents both data and accuracy dependencies between derived layers and their sources (Figure 6).

Error propagation functions manipulate this knowledge representation to access error indices associated with input layers, calculate, and store error values for derived layers. As GIS transformations are applied, error propagation functions traverse links emanating from meta-database representations of derived layers to access the input data's error indices. As an illustration of automated error propagation, consider the propagation of thematic classification error through the GIS application shown in Figure 5. The error propagation functions are based on ARC/INFO terminology but are intended in a broader, more conceptual sense, as described below. The application may be explained as follows:

- Source layer LANDUSE (land use classes) is transformed using the RESELECT function to create a binary layer called RANCHES (private ranches).
- Source layer VEGETATION (vegetation classes) is transformed using the RESELECT function to create a binary layer called OAKWOODS (oak woodlands).
- Derived layer RANCHES and source layer PERMITS (grazing permits on private lands) are overlayed using the UNION function to create a binary layer called GRAZING (private and public grazing lands).
- Derived layers GRAZING and OAKWOODS are overlayed using the INTERSECT function to create a binary layer called AT_RISK, which identifies lands in which oak regeneration is at risk from grazing activities.

The UNION and INTERSECT transformations represent Boolean OR and AND operations, respectively. The RESELECT transformation is conceptually equivalent to a reclassification in which certain cover classes on the input data are assigned new classes on the output data. Each of the three GIS data transformation functions used in this application (i.e., UNION, INTERSECT, and RESELECT) induces changes in thematic classification accuracy. All three source layers (i.e., LANDUSE, PERMITS, and VEGETATION) are assumed to be characterized by the proportion correctly classified (PCC) index of classification accuracy. Propagation of error indices parallels the propagation of data. As each GIS function is applied to the input data to derive the output data, the error propagation function is passed the PCC value associated with the input layer in order to calculate the PCC value for the output layer. For source layers LANDUSE and VEGETATION, the PCC index reflects the accuracy with which cover classes are depicted on the layer. In the case of PERMITS, a binary source layer, the PCC index is assumed to reflect uncertainty associated with incompleteness in grazing permit records and changes in the status of permits following publication of the data. The transformation of the PCC index through the application is based on the assumptions that errors are uncorrelated across data layers and are distributed uniformly across classes of the thematic attributes. Figure 7 is an illustration of the propagation of accuracies through the example application based on this assumption.

The error propagation functions discussed below illustrate how error might be propagated through a GIS application, but they do not represent the only way in which error propagation mechanisms might be modeled. A fundamental characteristic

of the paradigm presented here is that different error propagation functions can be employed for any particular combination of GIS function and error index. Each different error propagation function represents a different set of assumptions about the nature of the errors present in source data or the mechanisms whereby these errors are propagated (i.e., each function is from a different "plane" as illustrated in Figure 4). In this way, the paradigm serves as a framework for exploring the effects of using different error propagation functions. New functions can also be incorporated into this framework as knowledge concerning error measurement and propagation mechanisms improves. The discussion that follows also describes how error might be propagated for different assumptions about error propagation mechanisms (e.g., random versus non-random errors and different spatial distributions of error). This is intended to illustrate the flexibility of the paradigm, in that alternate error propagation functions are accommodated, allowing the paradigm to serve as a framework for error propagation modeling.

AN EXAMPLE OF ERROR PROPAGATION

The RESELECT function involves the selective retrieval of a subset of features on an input layer based on their thematic attribute values. In the case of categorical data, the number of thematic attribute classes on the output layer is less than the number on the input layer. That is, we assume that the RESELECT function is applied to a layer to selectively retrieve areas with a particular thematic class. This means that the remaining classes in the layer are effectively collapsed into a single class on the output layer. Therefore, the only misclassifications that occur on the output layer are those cases in the input layer where the selected class is confused with one of the unselected classes. Misclassifications in the input layer occurring across the constituent collapsed classes do not introduce error into the output layer. This collapsing of classes implies that the function will generally tend to increase thematic classification accuracy, because misclassifications that occur across constituent classes of each collapsed class no longer have any impact on the degree of error. In effect, one is trading information (in this case, both spatial and thematic) in exchange for accuracy.

As an illustration, assume that a layer being input to the RESELECT function has five classes labeled A through E, and the RESELECT function is applied to select areas with a class equal to A. Thus, classes B through E will be collapsed into a single "not A" class on the output layer. Error is propagated through the RESELECT function for those locations on the layer where the selected class (i.e., class A) is confused with one of the unselected classes (i.e., classes B through E). However, misclassifications between the constituent classes of the unselected classes (i.e., classes B through E) do not contribute to the error in the output layer, as we are only concerned that these classes are "not A." In order to propagate the PCC index through the RESELECT function in this way, the classification error matrix must be available (i.e., a thematically differentiated model of error is assumed). In this case, it is a straightforward process to compute the matrix for the output layer by collapsing the matrix for the input layer. Note that it may be necessary to weight the contribution of each class to the overall PCC by the estimated area of each class if the matrix was not constructed from a random sample.

If only the PCC, and not the entire classification error matrix, is available for the input layer, then error propagation can be performed assuming that errors are distributed uniformly across classes of the thematic attribute. That is, the probability of misclassification is identical for any two classes. All non-diagonal elements of the classification error matrix are assumed to be identical and each class is assumed to contribute equally to the overall PCC. Based on this assumption, it is possible to reconstruct the classification error matrix based only on the PCC of the input layer and k, the number of classes of the thematic attribute for the input layer. Propagation of the PCC index can then be modeled in terms of the number of classes in the input layer that are collapsed to create each of the new classes on the output layer. The error propagation function in this case may be expressed as

$$\text{PCC\_out} = \text{PCC\_in} + K\,(1 - \text{PCC\_in})$$

where

$$K = \frac{k - 2}{k}.$$

In these equations, PCC_out is the PCC of the output layer, PCC_in is the PCC of the input layer, and k is the number of classes in the input layer.

This simple model assumes that, of the k classes in the input layer, $k-1$ classes are "collapsed" into an "other" category. That is, it assumes that the user is interested in extracting only one class from the input layer. The model is easily generalized to account for situations in which more than one class is extracted. In this case, the variable K in the above equation is redefined as

$$K = \frac{r\,(r - 1) + (k - r)\,(k - r - 1)}{k\,(k - 1)}$$

where k is the number of classes in the input layer and r is the number of classes collapsed to form the extracted class.

It is also possible to weight the contribution of each class based on the estimated area of each class. This could provide a more representative output PCC value if the classes in the input

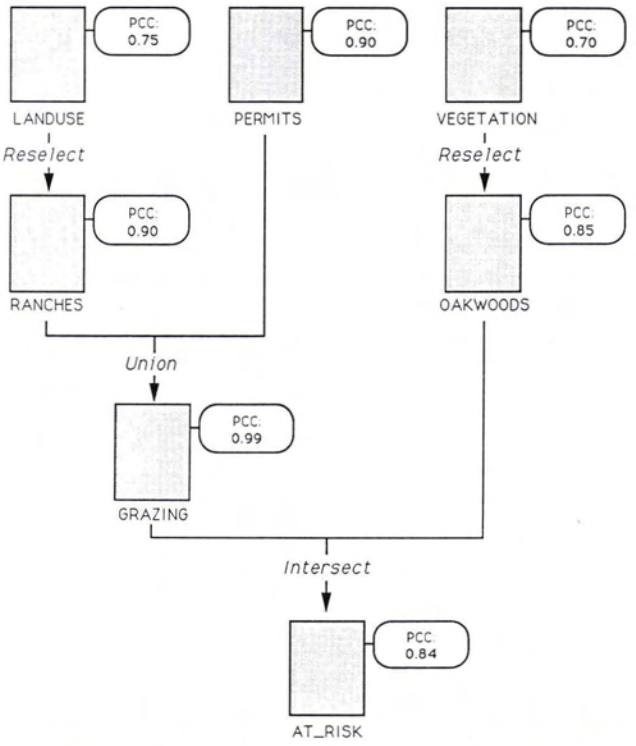

FIG. 7. Meta-database representation of an error index propagated through the GIS application.

layer have significantly different areas. In this case the error propagation function may be written as,

$$\text{PCC\_out} = \text{PCC\_in} + \sum_{i \in R} \frac{K_R\, w_i\, (1 - \text{PCC\_in})}{k - 1} + \sum_{j \in U} \frac{K_U\, w_j\, (1 - \text{PCC\_in})}{k - 1}$$

where

R is the set of reselected classes,
U is the set of unselected classes,
$K_R = r - 1$,
$K_U = k - r - 1$, and
w_i is the weight (relative area) of class i.

Many other error propagation functions could also be proposed here. Our point, however, is simply to illustrate the role assumptions play in matching an error propagation function to a particular combination of GIS function and error index.

The INTERSECT function creates an output layer containing the intersection of the features on two input layers. It is assumed here that the function is applied to binary input layers. That is, layers are assumed to contain only two thematic attribute classes. This implies that the INTERSECT function is equivalent to the Boolean AND operation. Thematic classification accuracy for the output layer is defined as the intersection of the correctly classified portions of the input layers. That is, a point has to be correctly classified on both input layers in order to be considered accurate on the output layer. In terms of the PCC index, this implies that the PCC of the output layer can never be higher than the PCC of the least accurate input layer.

Bayes's Theorem has been used to construct an error propagation model for this function (Newcomer and Szajgin, 1984). However, this model depends on the degree to which the correctly classified portions of the two input layers tend to overlap. More precisely, the model requires the conditional probability of observing a correctly classified point on one input layer given that the point is correctly classified on the other input layer. This means that the locations of the correctly and incorrectly classified portions of the input layers must be known (i.e., a spatially differentiated model of error is assumed). When only the PCC index is available, error propagation can be performed assuming errors are uncorrelated across layers (MacDougall, 1975). In other words, the probability of observing a correct classification at a point on one of the input layers is the same regardless of whether or not that point is correctly classified on the other layer. In this case, the PCC of the output layer (PCC_out) is simply the product of the PCC of the two input layers (PCC_in1 and PCC_in2); that is,

$$\text{PCC\_out} = \text{PCC\_in1} \times \text{PCC\_in2}.$$

The more general form of the error propagation function is

$$\text{PCC\_out} = \text{PCC\_in1} \times \text{PCC}[\text{in2} \mid \text{in1}],$$

where PCC[in2 | in1] is the conditional probability of observing a correctly classified point on layer 2 given that the point is correctly classified on layer 1. The maximum value of this conditional probability is 1 (in which case PCC_out = PCC_in1) and the minimum value is 0 (in which case PCC_out = 0). When PCC[in2 | in1] = PCC_in2, the error propagation function is equivalent to the uncorrelated case presented above. When $1 \leq \text{PCC}[\text{in2} \mid \text{in1}] < \text{PCC\_in2}$, then correctly classified points tend to co-occur spatially (i.e., they tend to occur at the same locations on layer 1 and 2). In this case, the accuracy of the output layer will be higher than for the uncorrelated case. When $\text{PCC\_in2} < \text{PCC}[\text{in2} \mid \text{in1}] \leq 0$, then the correctly classified points tend not to co-occur spatially. In this case the accuracy of the output layer will be lower than for the uncorrelated case.

Using the more general form of the error propagation model, the user can enter an appropriate value for the conditional probability to reflect the degree to which errors are thought to co-occur spatially. This value might be derived through *a priori* knowledge or empirical data. As in the case of the RESELECT function, this discussion shows how the reliability of a propagated error index can be enhanced through the use of ancillary data, and how alternate error propagation models can be employed given different assumptions about error propagation mechanisms.

These functions can also be applied in the more general case in which input layers contain more than two classes (see Veregin, 1989b). The restriction to binary input layers in this discussion reflects a desire to avoid specific implementation issues and, in particular, the raster-vector dichotomy. In the example "oak woodlands at risk from grazing" application previously described, the distinction between raster and vector is not important for error propagation modeling. That is, the same sequence of GIS functions would be used in either case. In many vector-based systems, thematic data are isolated from the spatial data to which they refer. As a result, Boolean overlay is typically not implementable as a single GIS command. Rather, one must first perform the appropriate topological overlay (e.g., INTERSECT or UNION) on the spatial data, and then manipulate the thematic data by selecting those features with a particular combination of thematic attribute values. It is conceptually much simpler to perform the selection process first to derive a set of binary data layers, and then apply the appropriate topological overlay function. This avoids the implications alternative data structures have on error propagation.

The UNION function creates an output layer containing the union of the features on two input layers. As in the case of the INTERSECT function, it is assumed here that the function is applied to binary input layers, such that the UNION function is equivalent to Boolean OR operation. Thematic classification accuracy for the output layer is defined as the union of the correctly classified portions of the input layers. That is, a point needs to be correctly classified on only one input layer in order to be considered accurate on the output layer. In terms of the PCC index, this implies that the PCC of the output layer can never be lower than the PCC of the most accurate input layer. Thus, in contrast to the INTERSECT function, the UNION function tends to increase thematic accuracy (Veregin, 1989b).

As in the case of the INTERSECT function, the error model for the UNION function depends on the degree to which errors on the two input layers tend to overlap. When only the PCC index is available, error propagation can be performed under the assumption of uncorrelated errors. In this case, the PCC of the output layer is defined in terms of the product of the arithmetic inverse of the PCC of each input layer (i.e., the probability of misclassification); that is,

$$\text{PCC\_out} = 1 - (1 - \text{PCC\_in1})\,(1 - \text{PCC\_in2}).$$

In this case, one assumes that the probability of observing a misclassification at a point on one of the input layers is the same regardless of whether or not that point is misclassified on the other layer.

The more general form of the error propagation function is

$$\text{PCC\_out} = 1 - (1 - \text{PCC\_in1})\,((1 - \text{PCC})\,[\text{in2} \mid \text{in1}])$$

where $((1 - \text{PCC})\,[\text{in2} \mid \text{in1}])$ is the conditional probability of observing a misclassified point on layer 2 given that the point is misclassified on layer 1. When $((1 - \text{PCC})\,[\text{in2} \mid \text{in1}]) = \text{PCC\_in2}$, the error propagation function is equivalent to the

uncorrelated case, as defined above. When $1 \leq ((1 - \text{PCC})\ [\text{in2} \mid \text{in1}]) < 1 - \text{PCC\_in2}$, then misclassified points tend to co-occur spatially and the accuracy of the output layer will be lower than for the uncorrelated case. When $1 - \text{PCC\_in2} < ((1 - \text{PCC})\ [\text{in2} \mid \text{in1}]) \leq 0$, then the misclassified points tend not to co-occur spatially and the accuracy of the output layer will be higher than for the uncorrelated case.

If the assumption behind an error propagation function is unrealistic, it might produce an error index that misleads decision makers into placing too much credence in – or discounting the implications of – explicit spatial relations encoded in a GIS derived map. Moreover, the types of harm that might be caused will depend on whether one is a data producer (seeking to minimize the probability of erroneously rejecting data that actually meet a required accuracy standard) or a data consumer (seeking to minimize the probability of erroneously accepting data that actually do not meet the standard). Of course, armed only with the PCC value for an input layer, simplifying assumptions must be made in order to propagate error. The reliability of the propagated error index will likely improve with the availability of relevant ancillary data (e.g., the area of each class, the classification accuracy of each class, etc.). A strong case is thus made for preserving as much information acquired during data quality assessment as possible (i.e., preserving the locations and accuracies of all control points rather than just computing a single-valued error index such as the PCC or a root-mean-squared error).

In absence of such ancillary data, several other strategies can be employed. One option is to define the minimum and maximum possible error, so as to bound the range of error possible in a given application (Veregin, 1989b). One problem with this approach, however, is that the maximum and minimum errors tend to saturate quickly at 1 and 0, respectively, and thus lose all utility. A second problem is that in the context of the example application, the conditions causing error to be inflated for the INTERSECT function are exactly those that cause error to be deflated for the UNION function. In other words, the propagation of error under worst-case conditions for one of these functions is inconsistent with the propagation of error under worst-case conditions for the other function. For this reason, rather than assuming worst- or best-case scenarios, it seems more appropriate to adopt a set of assumptions about error propagation for a data set and employ error propagation functions that reflect these assumptions (as illustrated by the "planes" in Figure 4). Unfortunately, given current understanding of error propagation in a GIS, this ability may not be realizable, and more basic research on the mechanisms of error propagation is clearly needed.

CONCLUSION

Applied error propagation research involves identifying error indices, developing error propagation functions, and testing their utility in assessing spatial, thematic, and temporal accuracies of derived geographic data. Each error propagation function is determined by the specific error index to be propagated, the GIS transformation function to be employed, and a set of assumptions about the nature of errors in spatial data and their propagation mechanisms. This paper demonstrates the use of the GEOLINEUS lineage meta-database system to automatically propagate error indices through GIS spatial data transformation functions. The propagated indices can be evaluated in terms of their utility for judging the quality of GIS derived data products and their appropriateness in decision-making contexts. When necessary, new error indices and error propagation functions can be developed and tested.

The paradigm suggested here provides a framework for research on error propagation. It does not, however, address policy decisions concerning the meaning and use of propagated error indices. Error indices focus attention on the quality of derived data products, but do not define what level of error is acceptable. Such policies should be based on the significance or relevance of different types of error in particular decision-making contexts and as a function of institutional data accuracy requirements.

ACKNOWLEDGMENTS

We would like to thank Prof. Michael Goodchild, the NCGIA (Santa Barbara), and Geographic Designs Inc. for support of this research. We would also like to thank the reviewers for comments that stimulated us to clarify our thinking and express our objectives more precisely.

REFERENCES

American Society of Civil Engineers (Committee on Cartographic Surveying and Mapping Division), 1983. *Map Uses, Scales and Accuracies for Engineering and Associated Purposes*, American Society of Civil Engineers, New York.

Aronoff, S., 1982. Classification Accuracy: A User Approach, *Photogrammetric Engineering & Remote Sensing*, Vol. 48, No. 8, pp. 1299-1307.

Aronson, P., 1987. Attribute Handling for Geographic Information Systems, *Proceedings of the Eighth International Symposium on Computer-Assisted Cartography*, pp. 346-355.

Bedard, Y., 1987. Uncertainties in Land Information Systems Databases. *Proceedings, Auto Carto 8*, pp. 175-184.

Berry, J.K., 1987. Fundamental Operations in Computer-Assisted Map Analysis, *International Journal of Geographical Information Systems*, Vol. 1, No. 2, pp. 119-136.

Blakemore, M., 1983. Generalization and error in spatial data bases, *Cartographica*, Vol. 21, No. 2/3, pp. 131-139.

Bracken, I., and C. Webster, 1989. Toward a Typology of Geographical Information Systems, *International Journal of Geographical Information Systems*, Vol. 3, No. 2, pp. 137-152.

Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resource Assessment*, Clarendon, London.

Chrisman, N.R., 1983. The Role of Quality Information in the Long-Term Functioning of a Geographic Information System, *Cartographica*, Vol. 21, No. 2/3, pp. 79-87.

———, 1989. Modeling Error in Overlaid Categorical Maps. *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors), Taylor and Francis, London, pp. 21-34.

Chrisman, N.R., and B. Niemann, 1985. Alternative Routes to a Multipurpose Cadastre: Merging Institutional and Technical Reasoning, *Proceedings of the Seventh International Symposium on Automated Cartography*, pp. 84-93.

Digital Cartographic Data Standards Task Force (DCDSTF), 1988. Draft Proposed Standard for Digital Cartographic Data, *The American Cartographer*, Vol. 15, No. 1.

Fisher, P., 1989. Knowledge-Based Approaches to Determining and Correcting Areas of Unreliability in Geographic Databases, *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors), Taylor and Francis, London, pp. 45-54.

Flowerdew, R., 1988. *Statistical Methods for Areal Interpolation*, Research Report No. 16, Northern Regional Research Laboratory.

Fitzpatrick-Lins, K., 1978. Accuracy and Consistency Comparisons of Land Use and Land Cover Maps Made From High-Altitude Photographs and Landsat Multispectral Imagery. *Journal of Research, US Geological Survey*, Vol. 6, No. 1, pp 23-40.

Ginevan, M.E., 1979. Testing Land-Use Map Accuracy: Another Look, *Photogrametric Engineering & Remote Sensing*, Vol. 45, No. 10, pp. 1371-1377.

Goodchild, M.F., 1989. Modeling Error in Objects and Fields, *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors), Taylor and Francis, London, pp. 107-113.

Goodchild, M.F., and S. Gopal, 1989. *Accuracy of Spatial Databases*, Taylor and Francis, London.

Gustafson, G.C., and J.C. Loon, 1981. Updating the National Map Accuracy Standards, *Technical Papers of the American Congress on Surveying and Mapping*, pp. 466-482.

Honeycutt, D.M., 1986. *Epsilon, Generalization, and Probability in Spatial Data Bases*, Unpublished manuscript.

Hord, R.M., and W. Brooner, 1976. Land-Use Map Accuracy Criteria, *Photogrammetric Engineering & Remote Sensing*, Vol. 42, No. 5, pp. 671-677.

Hudson, W.D., and C.W. Ramm, 1987. Correct Formulation of the Kappa Coefficient of Agreement. *Photogrammetric Engineering & Remote Sensing*, Vol. 53, No. 4, pp. 421-422.

Kjerne, D., and K.J. Dueker, 1986. Modeling Cadastral Spatial Relationships Using an Object-Oriented Language, *Proceedings of the Se cond International Symposium on Spatial Data Handling*, pp. 142-157.

Lanter, D.P., 1990. *Lineage in GIS: The Problem and a Solution*, Technical Paper 90-6, National Center for Geographic Information and Analysis, Santa Barbara, California.

———, 1991. A Lineage-Based Meta-Database for GIS, *Cartography and GIS*, (In Press).

MacDougall, E.B., 1975. The Accuracy of Map Overlays. *Landscape Planning*, Vol. 2, No. 1., pp. 23-30.

Martin, J., and C. McClure, 1985. *Diagramming Techniques for Analysts and Programmers*, Prentice Hall, Englewood Cliffs, Inc.

McCarthy, J., 1960. Recursive Functions of Symbolic Expression and Their Computation by Machine: Part 1, *Communications of the ACM*, Vol. 3, No. 4, pp. 184-195.

———, 1963. A Basis for a Mathematical Theory of Computation, *Computer Programming and Formal Systems* (P. Bradford and D. Hirshber, editors), Amsterdam: North Holland.

Merchant, D.C., 1987. Spatial Accuracy Specification for Large Scale Topographic Maps. *Photogrammetric Engineering & Remote Sensing*, Vol. 53, No. 7, pp. 958-961.

Newcomer, J.A., and J. Szajgin, 1984. Accumulation of Thematic Map Errors in Digital Overlay Analysis, *The American Cartographer*, Vol. 11, No. 1, pp. 58-62.

Stearns, F., 1968. A Method for Estimating the Quantitative Reliability of Isoline Maps, *Annals of the Association of American Geographers*, Vol. 58, No. 3, pp. 590-600.

Stoms, D., 1987. Reasoning with Uncertainty in Intelligent Geographic Information Systems, *GIS '87*, pp. 693-700.

Story, M., and R.G. Congalton, 1986. Accuracy Assessment: A User's Perspective, *Photogrammetric Engineering & Remote Sensing*, Vol. 52, No. 3, pp. 397-399.

Tomlin, C.D., and J.K. Berry, 1979. A Mathematic Structure for Cartographic Modeling in Environmental Analysis, *Proceedings of the American Congress on Surveying and Mapping*, pp. 269-283.

Veregin, H, 1989a. *A Taxonomy of Error in Spatial Databases*, Technical Paper 89-12, National Center for Geographic Information and Analysis, Santa Barbara, California.

———, 1989b. Error Modeling for the Map Overlay Operation. *Accuracy of Spatial Databases* (M. Goodchild and S. Gopal, editors), Taylor and Francis, London, pp. 3-18.

(Received 14 February 1991; revised and accepted 2 October 1991)