A User-friendly Tool for Error Modelling and Error Propagation in a GIS Environment

Frank Forier and Frank Canters

Abstract. - In spite of a growing concern for spatial database accuracy today most commercial GIS software products do not incorporate tools for the modelling of error in individual data layers and for the tracking of error when data layers are manipulated and combined in GIS-based spatial analysis. This paper reports on design considerations that have been taken into account in the development of a system-independent tool for error modelling and error propagation in raster-formatted categorical data. The tool aims at an optimal use of available methods for formal error propagation and simulation modelling. Special attention is given to the parameterization of different error models, i.e. how to make optimal use of existing information on error in the source data. When simulating error in the data layers spatial autocorrelation in the error field can be taken into account. This allows a more realistic assessment of uncertainty for spatial analysis which involves the application of neighbourhood or global operations.

INTRODUCTION

The widespread availability and use of geographical information systems in recent years has greatly enhanced the opportunities for a more rationalized approach to spatial decision making. There is, however, a growing awareness that a successful use of GIS can only be achieved if it becomes possible to attach a quality label to the output of each GIS analysis. Therefore issues of spatial data quality, error modelling and error propagation have gained considerable interest from the GIS research community. The number of publications on the subject is rapidly increasing and several techniques for the modelling of uncertainty in spatial data sets as well as formal methods describing the mechanisms of error propagation have been developed.

Despite the impressive volume of research on the topic of error in spatial databases there is still a lack of proper tools for error handling in commercial GIS software. The aim of the work described in this paper is to bundle existing techniques for error modelling and error propagation in a software-independent tool. In designing the tool the following aims as stated by Heuvelink (1993) are pursued: 1. the error tool should not replace the GIS but only add functionality; 2. the error tool should be able to handle a variety of GIS operations; 3. the error tool must have a user-friendly interface to encourage its use; 4. the error tool should employ efficient routines to minimize the numerical load.

1 Centre for Cartography and GIS, Department of Geography, Vrije Universiteit Brussel, Brussel, Belgium.
E-mail: fforier@vnet3.vub.ac.be, fcanters@vnet3.vub.ac.be

225
The process of handling uncertainty in spatial data involves three steps. The first step consists in identifying for each data layer what kind of error information is available or can be extracted from other data sources. The type of error information will determine which assumptions have to be made when error is modelled (Lanter and Veregin 1992). For example, if thematic error is merely described by a global index of classification accuracy, e.g. the PCC index or kappa statistic, no thematic or spatial differentiation of error is possible. This means that one is forced to assume that error levels do not vary from class to class and are uniformly distributed over space, even although this will probably not be the case.

The second step in the error handling process involves the definition of a conceptual model of error. The choice which is made at this point will depend to a large extent on the kind of error information that is available (step 1). With respect to the type of error information used in the model a distinction can be made between global, thematic and spatial/thematic error models. The choice of an appropriate error model will also depend on the error propagation technique that will be applied (step 3). In this connection a distinction should be made between purely descriptive and stochastic error models.

Finally an appropriate error propagation function has to be selected to verify how error is propagated through the GIS operation. In the tool which is described in this paper two techniques for error propagation are considered, i.e. formal error propagation and simulation modelling. Formal error propagation uses explicit mathematical models which describe the mechanisms of error propagation for particular GIS operations (Veregin 1994). The formal approach has the advantage of easy applicability. Starting from a number of error indices which describe the quality of the data layer(s) involved in the operation a mathematical function is applied which directly calculates one or more error indices for the output layer. The problem with formal error propagation is that very little is actually known about error propagation mechanisms. Hence error propagation functions are only available for a small number of relatively simple GIS operations and prove to be rather restrictive with respect to the type of error models that can be used. If formal error propagation cannot be applied simulation modelling is the alternative. The main advantage of simulation modelling is its overall applicability regardless of the type of GIS operator which is involved. A disadvantage, however, is the high numerical load which makes that simulation modelling is less practical in connection with real-time error propagation.

IDENTIFICATION AND EXTRACTION OF ERROR INFORMATION

One of the most important obstacles when handling uncertainty in GIS is the lack of knowledge about the error which is present in the source data. The type of error information which is available will define what kind of error modelling can be applied, whether error can be differentiated thematically and/or spatially, and as such will largely determine the quality of the whole error modelling process. Because the problem of error propagation in GIS has long been neglected information on the quality of the data is usually very sparse. Ideally, for each source layer in a GIS a map showing the spatial distribution of error should be available. Yet very often only global indicators of data quality are available providing no indication on local variations in the error field.

For categorical data information on error is usually described by a confusion matrix or classification error matrix. The matrix is constructed by cross-tabulation
of encoded and actual attribute values for a sample of locations. The element $c_{ij}$ in the matrix represents the number of cells assigned to class $i$ that actually belong to class $j$. A range of measures of accuracy such as the PCC (Proportion Correctly Classified) index and the kappa statistic (both global indices of classification accuracy), and the user's and producer's accuracy for each class can be deduced from the matrix (Congalton 1991). The confusion matrix reports error on the level of different thematic classes. Its major drawback is that it provides no information on within-class variations in the error field. The confusion matrix will therefore only support the application of thematically differentiated error models. Spatially differentiated models require extra information which is not contained in the matrix.

Categorical maps are usually not accompanied by a description of error in the form of a confusion matrix. When attempting to model error in GIS source layers that are derived from existing maps it is therefore often necessary to extract error information from a variety of other sources. Sometimes information on error can be found in reports which are delivered with map documents. Fisher (1991), for example, showed how information on soil map-unit inclusions from soil survey reports can be used to build a confusion matrix. Information on uncertainty can also be derived from field inventories (Aspinall and Pearson 1994) or from more detailed maps. Overlaying the original map with a more detailed spatial reference not only allows to derive a confusion matrix, but also makes it possible to obtain useful information on the spatial structure of uncertainty, for example, by calculating the mean size of inclusions of each class in the reference map within each class in the original source map. This kind of information is of utmost importance for the application of spatially differentiated error models, as will be shown later. Interesting information may also be provided by related spatial data. Fisher (1989) demonstrated how expert knowledge about the relation between soil and other physical data can be used to predict the spatial distribution of error in soil maps from the mapped distribution of these related variables.

In developing the tool which is referred to in this paper special effort has been made to provide the user with a set of functions which allow to derive error information for categorical maps from more detailed or related map documents. If no information on error is available or can be derived from other sources the user is offered the opportunity to introduce uncertainty in the data layers. This allows to carry out sensitivity analysis enabling the user to examine to what extent the output of a sequence of GIS operations is sensitive to specified levels of uncertainty in the source data.

**CHOOSING AN APPROPRIATE ERROR MODEL**

Depending on the error information which is available three types of error modelling can be applied: global, thematic and spatial/thematic modelling. For global and thematic modelling the parameterization of the error models is based on well-known error indices that can be derived from the confusion matrix. Some error propagation techniques use the entire confusion matrix as a model of error (Veregin 1995). Spatial/thematic modelling requires additional information on the distribution of error within each thematic class.

In global error modelling thematic error is described by one single-valued index which is used as a model parameter. The best known error index that can be derived from the confusion matrix is the PCC (Proportion Correctly Classified) index which represents the probability that a location selected at random is correctly
classified. The PCC index is defined as the sum of the diagonal elements in the confusion matrix divided by the total number of samples. It has a value between 0 and 1, corresponding with maximum and no error respectively. Alternatively the kappa statistic may be used (Congalton 1991). The kappa index allows to account for change agreement and will be equal to zero if agreement is equal to what is expected in a random case. If agreement is less it will be negative, if agreement is perfect it will be equal to 1. The PCC index and the kappa statistic provide an overall measure of error. They do not describe how error may vary from class to class. Hence in global error modelling one is obliged to assume that error levels are the same for all classes and that confusion is equal among each pair of classes. If real error conditions largely deviate from these assumptions error propagation results based on global error modelling may be of little use.

Thematic error modelling accounts for variations in the level and type of error for different classes. Depending on the error propagation technique which will be used specific information on thematic error for different classes will be derived from the confusion matrix. For some GIS operations formal error propagation functions have been defined which propagate the entire confusion matrix of the source layer(s) through the operation to produce a confusion matrix for the output layer (Veregin 1995). This means that all error information which is present in the confusion matrix is actually used in the modelling. Hence the confusion matrix itself can be considered as the error model. Unfortunately at present formal error propagation functions are only available for a small set of simple GIS operations. If other operations are involved simulation modelling is the only alternative.

Simulation modelling requires the definition of a stochastic error model which allows to produce perturbed versions of the original data layer using available error information. Stochastic error models for categorical data randomly assign each cell in the data layer to one of the classes according to a priori defined class membership probabilities (Goodchild and Wang 1989, Fisher 1991). For all cells which have been assigned to a given class in the original source data membership probabilities for all classes can be derived from the confusion matrix by dividing the elements in the corresponding row of the matrix by the total number of sample locations that have originally been assigned to the class. Using the obtained class membership probabilities the most obvious way of introducing uncertainty into the source layer is to assign classes independently in each cell. As class membership probabilities for a cell always add to one each class can be matched to a sub-interval in the range 0-1 with its length proportional to the probability of that class. For each cell a random number in the range 0-1 is drawn and the cell is assigned to the class that corresponds with the interval including this random number. Applying a sequence of GIS operations to the perturbed data layer will yield an outcome that differs from the result obtained with the original source data. Differences in outcome which are obtained by repeating this process a large number of times will indicate how uncertainty is propagated through the analysis.

Whether formal error propagation techniques or simulation modelling is applied, the use of an error model which is solely based on class-specific information derived from the confusion matrix ignores the possible presence of autocorrelation in the error field. In fact a model which only permits thematic differentiation of error implicitly assumes that error distribution is purely random, in other words, that cells which are assigned to class i but actually belong to class j do not tend to cluster in space. In most cases this assumption will not be valid. For categorical data only one formal error propagation method is known to these authors which accounts for the presence of autocorrelation in the error field. The method allows to
estimate the accuracy of a buffer operation from a number of characteristics describing the source data and the error which is associated with it (Veregin 1994). Goodchild et al. (1992) presented a stochastic error model for categorical data which allows to control the level of spatial dependence by inducing correlation between outcomes in neighbouring cells. The mechanism of class assignment is again based on the cell's class membership probabilities. Yet instead of drawing a random number in the range 0-1 for each cell independently a spatially correlated random field is generated by means of a first order autoregressive model defined on a regular grid:

\[ z_{ij} = q (z_{i-1,j} + z_{i+1,j} + z_{i,j-1} + z_{i,j+1}) + e_{ij} \]  

(1)

where \( z_{ij} \) is the value of the random error field at position \( i,j \), \( q \) (\( < 0.25 \)) is a parameter that indicates the level of autocorrelation and where \( e_{ij} \) is a spatially independent Gaussian noise field with mean 0 and standard deviation unity. Heuvelink (1992) presented a simple and robust iterative method for the generation of random fields of type (1) which avoids the computational overhead associated with the traditional inversion method (Haining et al. 1983). Since the marginal distribution of the correlated random field is of the Gaussian type with zero mean and a standard deviation which is function of the parameter \( q \) a simple transformation allows to obtain a uniform distribution in the range 0-1. This guarantees that class assignment for each cell is proportional to the a priori defined class membership probabilities.

Although it takes spatial correlation into account Goodchild's model has two important drawbacks. First of all, it is only applicable to two classes. If the model is used on more than two classes the inclusions which are simulated in the original map units will be spatially nested (figure 1a). Secondly there is no simple way of defining a proper value for the parameter \( q \). Although the average size of patches in the simulated layer is clearly related to the value of \( q \) it is not possible to define a one-to-one correspondence between both as class membership probabilities disturb the relationship. Based on Goodchild's ideas a method is proposed which allows to simulate uncertainty in categorical maps with more than two classes. For each class in the source map the method takes into account the type and the mean size of inclusions as well as their proportion within the source map units. Instead of simulating inclusions of different classes simultaneously using one random field inclusions are generated sequentially (one class at a time), each time using a different random field. The advantage of sequential simulation is that adjacent inclusions will occur for each pair of classes, independent of the sequence of simulation (figure 1b).

The use of different random fields also allows to control the level of autocorrelation for each category and therefore provides a mechanism to adjust the size of inclusions. Unfortunately, as has been mentioned already, it is very difficult to define a clear relationship between the level of autocorrelation in the random field and the mean size of simulated inclusions. When sequential simulation is applied the relationship will not only be influenced by the class membership probabilities, yet also by the exact sequence of class simulation. As the level of autocorrelation which is required to generate inclusions of a given size cannot be defined a priori, an iterative method is used. Starting with a random field with a low level of correlation inclusions of one class are simulated. If the mean size of the inclusions is smaller than the proposed size the process is repeated with a random field which has a higher level of correlation, until the proposed size of inclusions is reached.
The major drawback of this procedure is its computational load since dealing with one class already involves the use of different correlated random fields. On-line generation of these fields would be impossible. That is why a large set of correlated random fields was generated for different levels of autocorrelation and permanently stored on disk. During the simulation process random fields with a required level of correlation are randomly picked from this set. At present the prototype tool for error modelling which has been implemented uses 15 sets of 50 correlated random fields, each set corresponding to a different level of autocorrelation. The fact that only 50 random fields are defined for each level has to do with limitations in storage capacity. The size of the fields was chosen large enough (1500 x 1500) to allow for their use in a large variety of applications.

![Simulation](image)

**Figure 1.-** Simulation of classes B and C into class A with probabilities $p_A = 0.5$, $p_B = 0.3$, $p_C = 0.2$ and level of autocorrelation $q = 0.2450$ (a) simulation according Goodchild's model with 1 random field, (b) sequential simulation involving 2 random fields.

**METHODS FOR ERROR PROPAGATION**

When source data are input to a GIS operation the uncertainty which is present in the data will be propagated through the operation and will introduce uncertainty in the outcome. Error propagation will continue when the output of one operation becomes the input to another. The network of input and output relations between spatial data layers can be schematized in a data flow diagram which is usually called a cartographic model (Tomlin 1990). Evaluating the accuracy of the output of a cartographic model is only feasible if the propagation of error can be modelled for each operation which is part of it. Although most GIS users may be aware that errors propagate through spatial analysis, in practice they will usually not pay any attention to it. An important reason for this is the lack of functionality for handling error propagation in commercial GIS.

Lanter and Veregin (1992) have proposed an automated system for error propagation that is based on the use of formal error propagation functions which transform the PCC index of the source layer(s) through the cartographic model of an application. The problem with formal error propagation methods, however, is that they are only available for a small number of GIS operations. Hence the use of tools which are strictly based on formal error propagation methods will be limited to
applications which only involve a few simple GIS operations. That is why some authors (Openshaw 1989) plead in favour of simulation modelling which puts no restrictions on the complexity of the application, thereby arguing that the computational load which is the major drawback of simulation modelling will become less a problem with the advent of increasingly powerful means of processing.

Although it may be questioned if the use of simulation modelling will allow enough runs to be performed in real-time to obtain a statistically valid quantitative assessment of uncertainty (Heuvelink 1993) it will give a clear indication of the level of uncertainty in the output which is to be expected when source data with known accuracy are subjected to a particular spatial analysis. While formal error propagation methods always produce a tabular report of error, i.e. a single error index or a confusion matrix, simulation modelling has the surplus advantage of providing interesting ways of visualizing uncertainty (see below). Hence it is the opinion of the present authors that most advantage can be derived from the complementary use of both error propagation techniques. In the following a short description is given of the techniques which have been implemented. Some reference is also made to their practical use.

**Formal Error Propagation Methods**

Formal methods of error propagation can be defined as mathematical functions describing how errors in the source data are modified by a particular GIS operation. Each error propagation function is based on a set of assumptions about the nature of the errors present in the source data (a particular model of error) and is specific to one GIS operation. Assumptions about the nature of error will be based on the type of error information that is available for the source data, i.e. a global error index, class-specific indices of error or the entire confusion matrix. At present the prototype tool only includes functions which propagate the entire confusion matrix through the GIS operation. Functions are available for overlay (AND, OR, XOR) and for the reclass operation (Veregin 1995). If the entire confusion matrix is not known it is reconstructed from available error indices by making some simplifying assumptions about the distribution of error among different classes. So far no assumptions are made on thematic dependence, i.e. errors in different source layers are considered statistically independent.

For many GIS operations little is known about the mechanisms of error propagation and formal error models remain to be developed. Recently Veregin (1994) reported on a method to derive a formal mathematical expression describing how error is propagated through the buffer operation using simulation modelling. The method demonstrates how simulation modelling can add to a better understanding of error propagation mechanisms and to the development of formal expressions describing how error is propagated. It provides a good example of the complementary use of both error propagation techniques and of the benefits of integrating both in the same error modelling tool. Further research in this direction may lead to the development of formal error propagation functions for other more complex GIS operations.

Although the use of formal methods for error propagation in categorical data is still limited to a few GIS operations the technique is easy to apply. When performing an analysis where multiple source layers are subjected to a sequence of GIS operations it may be advantageous to combine simulation modelling with
formal error propagation in order to decrease the computational load. The combined use of both error propagation techniques has been a major concern in designing the tool which is described in this paper.

Monte Carlo Simulation Modelling

Simulation modelling starts with the definition of a stochastic error model which allows to produce a set of perturbed versions of the original source data. As already explained the stochastic model will be based on a number of assumptions with regard to the type, the magnitude and the spatial distribution of error. The method for stochastic modelling which has been presented in the previous section and which allows to deal with autocorrelation in the error field has been implemented in the prototype version of the error handling tool. If no information on the spatial structure of error is available it is assumed that no correlation is present in the error field. If the mean size of inclusions in each class is known sequential simulation as described above can be applied.

Each perturbed version of the source data is subjected to the sequence of GIS operations involved in the analysis and will produce a different output. If the output of the analysis is a set of numerical values, e.g. the total area allocated to different types of land-use, the mean output over M realizations gives an estimate of the most likely outcome of the analysis taking the uncertainty in the source data into account. The standard deviation over M realizations provides a measure of the uncertainty in the output. If the outcome of the analysis is an entire coverage, e.g. a visibility map, simulation results allow to generate a variety of products describing output uncertainty (see below).

The implementation of the stochastic modelling procedure is based on scripting. When the analysis is executed for the first time the cartographic model is stored in a script file. Monte Carlo analysis is automatically performed by running the script a fixed number of times. This strategy, of course, assumes that the GIS allows scripting and that the script language includes control structures for iteration.

VISUALIZATION OF UNCERTAINTY

An important issue in the development of a tool for error handling is how to visualize the quality of the source data and the reliability of the output of spatial analysis in an easy and straightforward manner. To give the user a visual impression of the error present in the source data it may be interesting to visualize the result of one stochastic simulation of error. When dealing with classified satellite data class membership probabilities produced by statistical classification techniques can be used to visualize uncertainty which is due to the classification process. Based on these probabilities not only the first, but also the second, third, ... most likely class can be shown. To evaluate the reliability of pixel assignment a probability image can be constructed which shows the highest class membership probability for each pixel. Confusion in pixel assignment can be measured by calculating the relative probability entropy (Maselli et al. 1994). This index indicates for each pixel if probability tends to concentrate in one class or is spread over several classes.

When error is propagated using the Monte Carlo approach and the outcome of the analysis is an entire coverage the same techniques as described for the visualization
of uncertainty in classified satellite data can be applied since for each pixel a vector indicating the frequency of occurrence of each class can be constructed from the M coverages. If error is propagated using formal error propagation methods uncertainty cannot be visualized since output is in tabular form.

IMPLEMENTATION

The above described techniques for error modelling and error propagation have been implemented in the C programming language in a UNIX workstation environment. The prototype version of the tool is linked to GRASS 4.1. The choice of GRASS has been based on the fact that it is public-domain software, written in C and running under all common UNIX operating systems and that it provides a powerful language for scripting. The most important advantage of GRASS, however, is its open structure, meaning that developers can directly add new functionality to the system. The approach that has been followed during the implementation of the error handling tool was to consider the GIS as an external module which is used to carry out spatial analysis and to display maps. Hence the tool operates independently from the GIS and should be easy-to-link with other GIS software.

CONCLUSION

This paper reviews existing methods for error modelling and error propagation in raster-formatted categorical data. The aim of the work reported here was to bundle these methods in a user-friendly and independent tool for error handling which can easily be linked with existing GIS software. A prototype version of the tool has been implemented using GRASS 4.1 for raster-based spatial analysis and display.

The tool combines the use of formal error propagation methods with simulation modelling. Both techniques should not be considered as two opposites but rather as complementary techniques. When spatial analysis involves the use of GIS operations for which no formal error propagation function exists simulation modelling can be applied. At present formal methods have been implemented for overlay analysis (AND, OR, XOR) and for the reclass operation. They allow to propagate the entire confusion matrix through the operation. Simulation modelling is based on a method for sequential simulation which allows to deal with spatial autocorrelation in the error field. In the implementation of the method specific measures were taken to reduce computational load.

ACKNOWLEDGMENTS

The work reported in this paper is supported by the Flemish Institute for Scientific and Technological Research for the Industry (IWT) and by the Belgian Scientific Research Programme on Remote Sensing by Satellite - phase three (Federal Office for Scientific, Technical and Cultural Affairs), contract Telsat/III/03/004. The scientific responsibility is assumed by its authors.
REFERENCES


BIOGRAPHICAL SKETCH

Both authors are research assistants in the Department of Geography of the Vrije Universiteit Brussel, Belgium. Their main research interests lie in spatial database accuracy, error modelling and error propagation in GIS.