Dealing with Uncertainty in Categorical Coverage Maps: Defining, Visualizing, and Managing Errors

Charles R. Ehlschlaeger and Michael F. Goodchild*

Abstract

There is a considerable body of literature on techniques describing and modeling spatial database uncertainty. Unfortunately, spatial error modeling still isn't available to the general users of spatial databases. Spatial error modeling will only become viable when spatial errors are easily defined, visualized, and made available to spatial applications. Easing spatial error modeling into the user domain will be far easier if error definition, visualization, and management are based on the same fundamental statistical principles. Accordingly, this paper will describe a theoretical foundation, using many geostatistical concepts, that allows for the definition, visualization, and management of categorical coverage map errors using random fields.

1 Introduction

While some GIS practioners incorrectly treat maps as reality, maps are only representations of reality,[13] and thus databases constructed from maps are also representations. The quality of a spatial database depends on how well it represents reality, which depends in turn on numerous factors in the processes used to create the database, such as measurement error, uncertainty in interpretation, and digitizing error. In this paper we focus on errors in one particular type of spatial database, specifically categorical coverage maps, where data quality is determined by whether the category labeled at each point in the database matches what really is there in the real world.

Most of the existing GIS research on spatial errors describes the magnitude of the various types of error[7][10] while the spatial autocorrelative component of error has received less attention. The First Law of Geography states: "Everything is related to everything else, but near things are more related than distant things."[15] To put it more succinctly, spatial autocorrelation exists and is positive.

If the First Law did not apply to the real world, spatial error modeling would be a trivial task using confusion matrices.[1] Confusion matrices (error matrices or contingency tables) are used to better understand categorical data. Confusion matrices are often created by comparing the data to a ground sample or more accurate data. A confusion matrix stores the probabilities that a given category value might be any of the category values in the map. For example, a categorical coverage map with three categories might have the following confusion matrix:

$$C = \begin{pmatrix} .90 & .06 & .04 \\ .08 & .83 & .09 \\ .03 & .02 & .95 \end{pmatrix}$$

where the first category has a 90% chance of being category one, a 6% chance of being category two, and a 4% chance of being category three. Research that applies random values to each cell to get a perturbed potential reality has been done.[6] GIS practitioners would be able to choose a set of random points in the study area and stochastically compare the results of an analysis to the real world. They could see for themselves whether the application answers the questions asked. Unfortunately, spatial autocorrelation exists, and errors are spatially autocorrelated. Spatial autocorrelation of error defines the "texture" of the error.[8] Products of spatially sensitive GIS applications (e.g. buffer and neighborhood commands) need the spatial autocorrelation of source map error to understand their uncertainty. Unless map construction metadata exists, issues such as product uncertainty and sensitivity, risk analysis and others are conceptually difficult.

Most map errors are spatially autocorrelated. For example, consider a manually made landcover map. A cartographer delineates the boundaries between different landcovers from one or multiple sources. There are three major forms of error in this method: 1) Clump(s) of categories may be mislabeled. 2) Clump(s) of categories may be missing. I.e., inclusions within other categories such as clearings in a forest, or along the edge between categories such as shrubby areas between forests and grasslands. 3) Clump edges may not be accurately positioned. Each of these forms of error are spatially autocorrellated. I.e., if a point on the landcover map is incorrectly labeled, nearby points will be more likely to be incorrectly labeled than points further away.

Random fields can be used to represent spatially autocorrelated error. In the case of categorical coverage maps, random fields can be used to determine which of the possible categories reside in each cell of a map realization. If map error is understood and represented perfectly, a map realization may be a possible representation of reality. By generating a number of potential realizations (ensembles in Monte

*National Center for Geographic Information and Analysis, University of California at Santa Barbara

Presented at: ACMGIS
Gaithersburg Md

Carlo simulation terminology), a GIS practitioner can run a GIS application for each map realization, creating a set of application products. This set of products can be used to create sensitivity analysis of the application based on original map error.

Ongoing research at NCGIA is creating a suite of public domain software tools that explicitly define error in maps, propagate map errors through all spatial analyses, and visualize the implications of map error on any and all spatial analyses. This paper will discuss the implementation of error modeling software tools for categorical coverage maps developed in the GRASS GIS. First, this paper describes how map error descriptors, magnitude and spatial autocorrelation, can generate random fields that represent potential realizations of error. Second, various animation techniques will be discussed, comparing their advantages and limitations for specific problem solving goals. Finally, spatial data management issues affected by the tools described in this paper will be addressed.

2 Spatially Autocorrelated Uncertainty

There isn't an elegant method to produce the metadata necessary to determine the uncertainty of categorical coverage maps. Each source of map error (inclusions, mislabels, and boundary errors) needs to be defined separately and quantitatively. This paper presents a geostatistical paradigm of predictive statistics to characterize categorical coverage map uncertainty.

Each source of map error for each map category is then represented as a probability surface Z and its random field parameters. This section will describe the random field generation process, random field parameters for each error source, and *r.random.model*,[3] a GRASS program to generate realizations of categorical coverage maps.

To facilitate making descriptions of errors into realizations of error fields, *r.random.surface*[4] the following equation is used to generate normally distributed random fields:

$$\begin{aligned} Z(\mathbf{u}) &= \frac{\sum\limits_{v} w_{u,v}\varepsilon_v}{\sqrt{\sum\limits_{v} w_{u,v}^2}}, \\ w_{u,v} &= \begin{cases} \left(1 - \frac{d_{u,v}}{D}\right)^E & : d_{u,v} < D \\ 0 & : d_{u,v} \geq D \end{cases} \quad , u \in \mathbf{u}, v \in \mathbf{v} \end{aligned} \tag{1}$$

where $Z(\mathbf{u})$ is a random field for the coordinate vector $\mathbf{u}$, $w_{u,v}$ is the spatial autocorrelative effect between points u and v, ε_v is an independent Gaussian random deviate with a mean of zero and variance of one, $d_{u,v}$ is the distance between points u and v, D is the minimum distance of spatial independence, E is the distance decay exponent, and $\mathbf{v}$ is the coordinate vector of independent Gaussian random deviate points within D's spatial autocorrelative effect of $\mathbf{u}$. Edge effect problems do not exist using $\mathbf{v}$ (for a comprehensive discussion of edge effects, see[9]). $\mathbf{v}$ also allows a combination of D and E parameters to have a specific theoretical correlogram, independent of $\mathbf{u}$. This greatly reduces the task of determining random field parameters for known error fields.

While equation (1) generates normally distributed random fields, uniformly distributed random fields are needed to select appropriate category classes from a probability distribution. A normally distributed random field from (1) is converted to a uniform distribution by determining each $Z(u)$ percentile ranking based on the distribution of ε_v. The rest of this paper will use $\hat{Z}(\mathbf{u})$ to represent a uniformly distributed random field. Any given $\hat{Z}(u)$ will have an equal probability of being any value between 0.0 and 1.0. See [5] for a more complete description of this random field model.

There is a straight-forward way of applying the random field parameters D and E to the various categorical coverage map errors. Each form of categorical coverage map error will be treated separately below.

Conceptually, class inclusions are the simplest form of categorical coverage map error. An inclusion occurs when a cell is described as a category with a $X\%$ chance of being another class. The Canadian Geographic Information System[16] is an excellent example of a GIS incorporating inclusions directly in the database structure. The Canadian Geographic Information System allows for up to three classes in each polygon (e.g. 80% deciduous forest with 15% inclusions of coniferous forest and 5% inclusions of grasslands). Cartographers can determine the random field parameters by comparing the size and distribution of known inclusions to a representative sampling of random fields.

Figure 1 shows 12 inclusion patterns of 25% inclusions with various random field parameters. The topology of a random field $Z(\mathbf{u})$ using distance decay exponent E values greater or equal to 1.0 can be characterized by round and oval hills, valleys and texture. The patterns in Figure 1 are formed from the lowest values in uniformly distributed random fields:

$$I^{(l)}(\mathbf{u},\mathbf{z}) = \begin{cases} 1 : \hat{z}^{(l)}(\mathbf{u}) \leq \mathbf{z} \\ 0 : \text{otherwise} \end{cases} \tag{2}$$

where $I^{(l)}(\mathbf{u},\mathbf{z})$ is the realization of a category from a category coverage map for the coordinate vector $\mathbf{u}$, the vector $\mathbf{z}$ contains the probability that the category exists in that spatial location for that category coverage map,[1] and $\hat{z}^{(l)}(\mathbf{u})$ is a uniformly distributed realization of $\hat{Z}(\mathbf{u})$. Linear inclusions (e.g. creek beds and ore deposits) can be generated using the median values of the random field realizations to determine inclusions:

$$\hat{I}^{(l)}(\mathbf{u},\mathbf{z}) = \begin{cases} 1\text{: } .5 - \mathbf{z}/2 \leq \hat{z}^{(l)}(\mathbf{u}) < .5 + \mathbf{z}/2 \\ 0\text{: otherwise} \end{cases} \tag{3}$$

where $\hat{I}^{(l)}(\mathbf{u},\mathbf{z})$ is the realization of a linear category from a category coverage map, and $\hat{z}^{(l)}(\mathbf{u})$ is a uniformly distributed realization of $\hat{Z}(\mathbf{u})$.

Figure 2 shows 12 examples for linear inclusions generated by using the median values from a random field. The linear features from median values result from regions between the hills and valleys of random fields with large spatial dependencies.

Category mislabeling map errors as conceptually similar to inclusion errors. The only difference is that the spatial dependence parameter of a mislabeled "inclusion" will be larger than the regions potentially mislabeled. For example, consider a satellite image of landcover for a National Park. Rock outcroppings and parking lots will have similar signatures. A landcover map might have many 40m by 40m regions of rock outcropping where 10% of the rock outcropping might actually be parking lots. Generating random fields with the spatial dependence parameter D set to 500m

[1] If a category has multiple uncertainty spatial dependencies for a particular map (e.g. grasslands in a landcover map with multiple terrain types), then the category should have multiple probability vectors, one for each terrain type and/or spatial dependency.

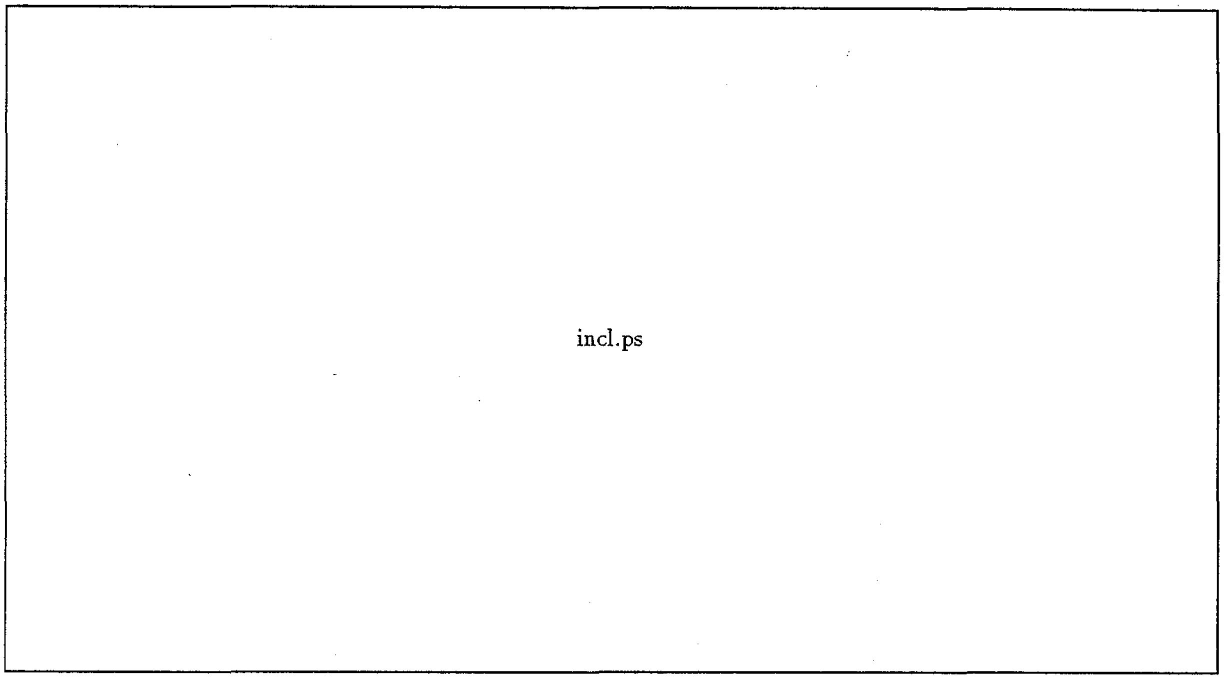

Figure 1: 12 white classes with 25% inclusion of black with varying spatial dependence D and distance decay exponents E. The D of each row is double of the row below (40, 20, & 10 cells with the resolution of maps: 61 by 50 cells). The E of each column is double of column to its left (1.0, 2.0, 4.0 & 8.0). Each random field's independent random deviates are the same for comparison purposes.

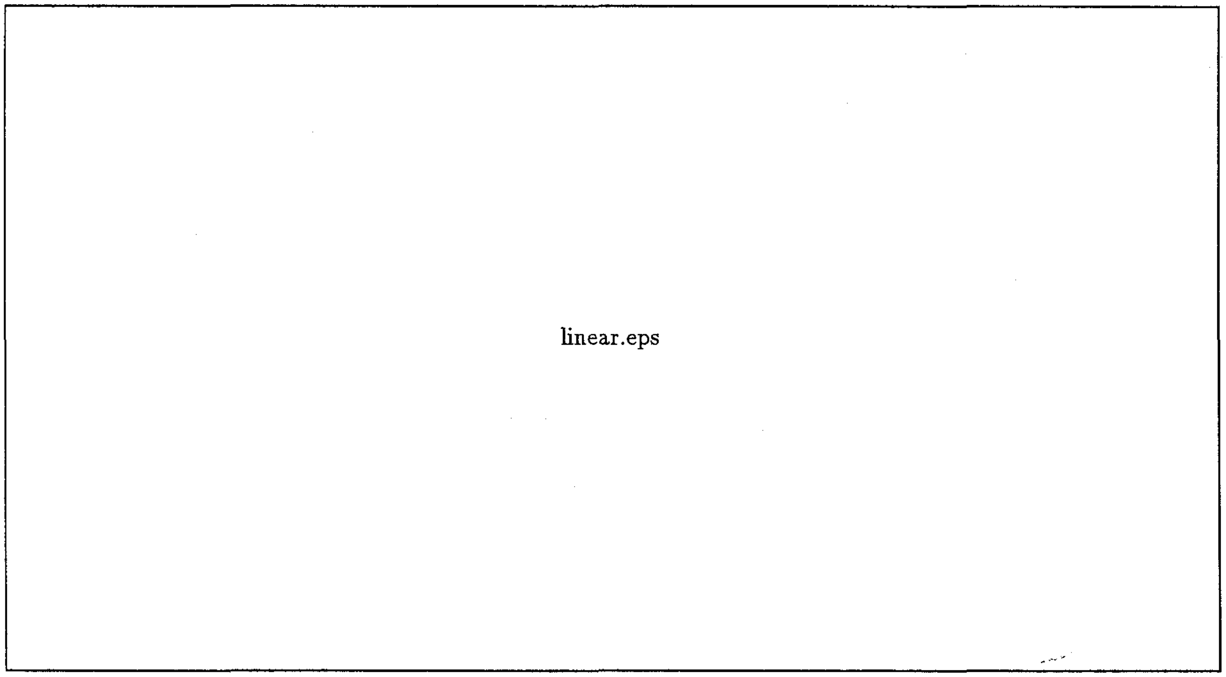

Figure 2: 12 white classes with 25% inclusion of black linear features with varying spatial dependence D and distance decay exponents E. The D of each row is double of the row below (40, 20, & 10 cells with the resolution of maps: 61 by 50 cells). The E of each column is double of column to its left (1.0, 2.0, 4.0 & 8.0). Each random field's independent random deviates are the same for comparison purposes.

or greater (and the distance decay exponent E set to 1.0, minimizing texture) make it likely that if a cell within a rock outcropping region is set to parking lot, all cells in that region will be set to parking lot. The actual choice of D will depend on the spatial autocorrelation of mislabeled regions (i.e., if mislabeled regions are clustered, the size of clustered regions should determine D).

Positional errors are the third form of categorical coverage map error. Sources of these errors include rectification error and uncertainty of category boundaries. Region boundaries may be represented as a transition zone between two or more categories.[12] Each transition zone will have a probability of 100% of being a class at close edge of the transition zone to 0% at the far edge. Figure 3 shows 12 transition zones between two classes. The left edge of each zone has a 2% chance of being black and the right edge has a 98% chance of being black. Figure 3 transition zones are linearly interpolated between 0% and 100%. Ideally, the transition zones should represent a normal distribution of expected boundary locations. Varying the parameters of D and E allows the cartographer to determine the texture of the edge line. As E increases, edge line complexity increases. As D increases, the edge lines develop more gradual curves. As D approaches infinity, the boundary lines will become parallel to the transition zone, but each realization's boundary will be distributed throughout the transition zone. While positional error can be represented in this manner, research continues at NCGIA to develop more elegant solutions for perturbing positional errors.

With the major forms of categorical coverage error defined, *r.random.model* provides a process to generate map realizations. To run properly, *r.random.model* requires a set of probability maps, at least one for each category, and a set of random fields, one for each probability map (built using the category's spatial autocorrelative parameters) except the final map. For each cell, *r.random.model* compares the probability of each category with its random field map in the order the category maps are given. The order is somewhat important as the error shapes of earlier categories will affect the shapes of later categories. This approach is similar to geostatistic's sequential simulation approach,[2] and is represented by:

$$
\begin{gathered}
\mathrm{z}_1 + \mathrm{z}_2 + \ldots + \mathrm{z}_n = 1.0 \\
\mathrm{c}_1 = \begin{cases} 1\text{: } \hat{z}_1^{(l)}(\mathrm{u}) \le \mathrm{z}_1 \\ 0\text{: otherwise} \end{cases} \\
\mathrm{c}_2 = \begin{cases} 1\text{: } \hat{z}_2^{(l)}(\mathrm{u}) \le \frac{\mathrm{z}_2}{1-\mathrm{z}_1} \\ 0\text{: otherwise} \end{cases} \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathrm{c}_n = \begin{cases} 1\text{: } \hat{z}_n^{(l)}(\mathrm{u}) \le \dfrac{\mathrm{z}_n}{1-\sum\limits_{i=1}^{n-1} \mathrm{z}_i} \\ 0\text{: otherwise} \end{cases} \qquad (4) \\
M^{(l)}(\mathrm{u}) = \begin{cases} C_1 : \mathrm{c}_1 = 1 \\ C_2 : \mathrm{c}_1 = 0, \mathrm{c}_2 = 1 \\ \cdots\cdots\cdots\cdots\cdots \\ C_n\text{: } \sum\limits_{i=1}^{n-1} \mathrm{c}_i = 0, \mathrm{c}_n = 1 \end{cases}
\end{gathered}
$$

where $M^{(l)}(\mathrm{u})$ is a potential realization of a category coverage map of vector u, z_1 through z_n are probability maps of categories C_1 through C_N, and $\hat{z}_1^{(l)}(\mathrm{u})$ through $\hat{z}_n^{(l)}(\mathrm{u})$ are potential realizations using spatial uncertainty parameters particular to each category.

Figure 4 shows two different realizations of a section

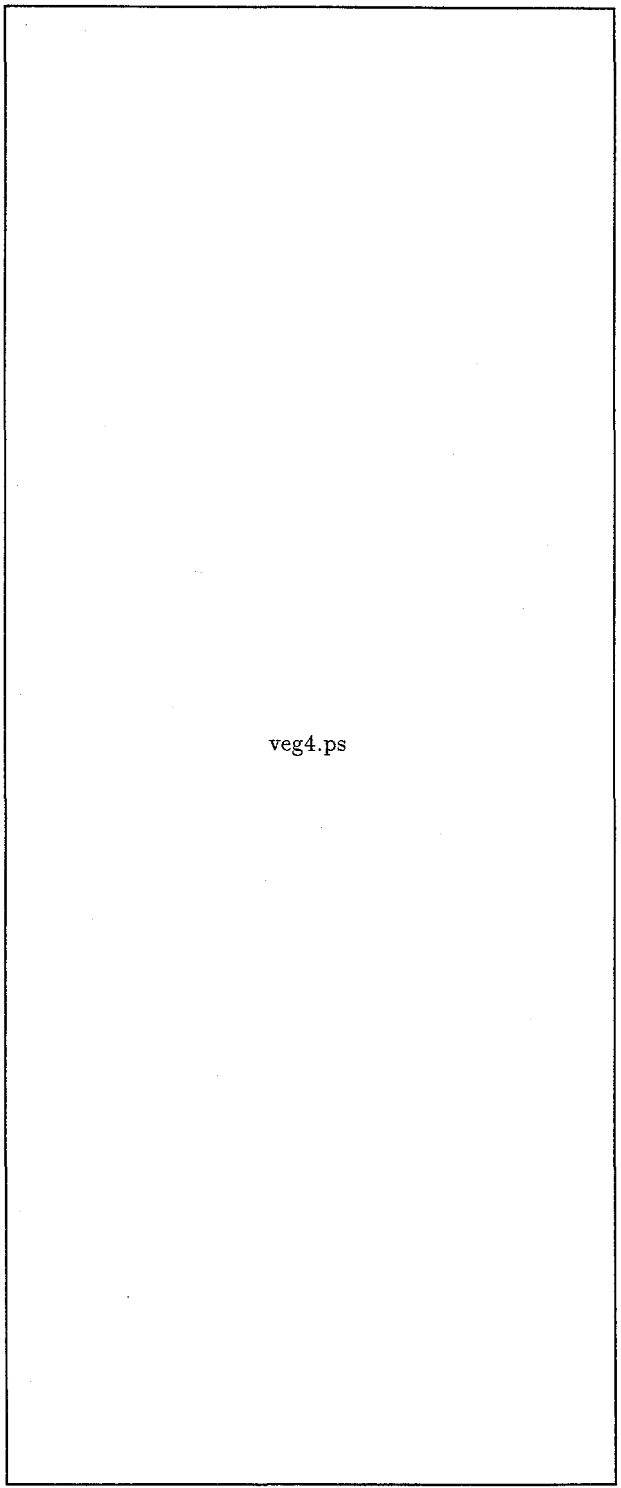

Figure 4: One section of GRASS's Spearfish landcover map (top) along with two realizations of landcover.

ramps.eps

Figure 3: 12 transition zones with varying spatial dependence D and distance decay exponents E. The left edge of each transition zone has a 2% chance of being black. The right edge has a 98% chance of being black. The D of each row is double of the row below (40, 20, & 10 cells with the resolution of maps: 61 by 50 cells). The E of each column is double of column to its left (1.0, 2.0, 4.0 & 8.0). Each random field's independent random deviates are the same for comparison purposes.

of Spearfish's landcover map along with the original section (the Spearfish database is available via ftp along with GRASS). Except for 15% inclusions of decidous forest (thin /// lines) and 10% inclusions of rangeland (skinny vertical lines) in coniferous forest (white), the realizations only take positional uncertainty into account. The positional uncertainty is represented by transition zones 400m wide, with a linearly interpolated probability between adjacent categories. Note that the error parameters chosen for this study are for illustrative purposes only. NCGIA is currently exploring data collection methods that generate spatial autocorrelation parameters along with the probability distributions of category classes.

3 Visualizing Uncertainty

Even if the explicit description of a map's error is definable, visualizing the map and its potential errors is not a trivial task. For the sake of discussion, consider the following problem: A landscape architect wants to locate a campground near the town of Spearfish, SD. Wishing to take advantage of existing landcover, the landscape architect wants to identify regions containing small clusters of deciduous forest, coniferous forest, and rangeland in close proximity to each other. The landscape architect's goal, as well as many GIS applications, may not be easily quantified.[2] In this case, the landscape architect needs more than a "the category most likely" or even probability surfaces of potential categories.

Showing the effects of map uncertainty and the influence of uncertainty's spatial autocorrelative effects on applications is a more difficult task. Depending on the application, it can be difficult to see how a map will affect that application. It is difficult to imagine the resulting shapes of a landcover map by viewing the random fields modifying it. It is even more difficult to imagine the potential shapes of a random field by knowing the specific error descriptors necessary to construct a representative random field. Unless the error is so minute as to be irrelevant, it is probably impossible to see how the error descriptors will modify the shapes and results of applications.

Considering the infinite variety of potential map error shapes and the lack of a systematic way of cataloging them, the only practical way to demonstrate map error is to show an ensemble of potential maps. This ensemble of map realizations will likely be formed with Monte Carlo sampling techniques due the potentially thousands of random variables needed to create one map realization.

Monte Carlo sampling has been proposed in error modeling as an analytical tool[14] because it can find the range of an application's variability easily. Openshaw suggests that a limited number of application runs (as few as thirty) is enough to give a statistical understanding of the application. For visualization purposes, thirty potential realizations are more than enough to show the general shapes caused by the spatial autocorrelation of errors.

4 Managing Uncertainty

These visualization techniques are conceptually compatible with Monte Carlo sampling as well as easy to understand. They allow the viewer to see how the errors in maps affect the quality of their application's results without requiring the viewer to explicitly understand the individual errors. By visualizing the application's results realizations and the data realizations, the viewer can make more informed manage-

[2] This situation can be placed in the "I can't define it, but I'll know it when I see it" category of GIS use.

ment decisions about the likelihood that their application's results are representative of reality.

The main drawback to this visualization technique occurs when the time required to show a representative sampling of realizations exceeds the attention span of the viewer. This can be a real problem if maps contain complex error descriptors or if unlikely potential realities can significantly affect the analyses of an application. For example, suppose a significant event only has a 1% chance of occurring. If the Monte Carlo simulation only has 100 realizations, there is $.99^{100}$ or 36.6% chance that the significant event will not be in the sampling. If the viewer is seeing one realization every second, he or she might not to wait for the nearly two minutes to view the relevant realizations. There are work-a-rounds to these problems, but they require the GIS practitioner to understand the ramifications of the potential misuse of information presented from a statistical perspective.

The worse thing that can happen is that the user can fall into a false sense of security due to the "pretty graphics" of animations technique. The authors hope that these visualization techniques are used to understand error, not hide it.

5 Concluding Remarks

This paper has discussed spatial data uncertainty issues, especially as they relate to categorical coverage maps. The most difficult part in defining spatial data uncertainty is determining the spatial autocorrelation and being able to represent this in a GIS. Random fields can be used to represent the various sources of spatial error, allowing realizations of data to be simulated. Applying the realizations of data to a GIS application creates an ensemble of application results. A set of ensembles for a GIS application gives the user a statistical perspective on the application's possibilities. The user can get a visual understanding of their application's uncertainties by displaying the ensembles in an animation where time dimension represents the likelihood of occurrence. Finally, the advantages and drawbacks of this Monte Carlo scheme were discussed.

6 Acknowledgements

The authors wish to thank the following people and organizations: The reviewers for their valuable suggestions, the Institute for Computational Earth Systems Science at UCSB for the access to their Video Imaging Lab, and Steve Miley for his assistance with the computer-to-video hardware and software.

References

[1] Congalton, R., Oderwald, R. & Mead, R. 1983, Assessing Landsat Classification Accuracy using discrete multi-variate analysis statistical techniques, *Photogrammetric Engineering and Remote Sensing*, Vol. 49, pp. 1671-8.

[2] Deutsch, C. & Journel, A. 1992, *Geostatistical Software Library and User's Guide*, Oxford University Press, New York.

[3] Ehlschlaeger, C. 1994a, r.random.model man page, *GRASS 4.2 User Manual*, Office of GRASS Integration, Champaign, IL.

[4] Ehlschlaeger, C. 1994b, r.random.surface man page, *GRASS 4.2 User Manual*, Office of GRASS Integration, Champaign, IL.

[5] Ehlschlaeger, C. & Goodchild, M. 1994, Uncertainty in spatial data: Defining, visualizing, and managing data errors, *Proceedings of GIS/LIS 1994*, pp. 246-53, Phoenix AZ.

[6] Fisher, P. 1991, Modeling soil map-unit inclusions by Monte Carlo simulation, *International Journal of Geographical Information Systems*, Vol. 5, No. 2, pp. 193-208.

[7] Fisher, P. 1992, First experiments in viewshed uncertainty: Simulating the fuzzy viewshed, *Photogrammetric Engineering & Remote Sensing*, Vol. 58, pp. 1345-52.

[8] Goodchild, M., Sun, G. & Yang, S. 1992, Development and test of an error model for categorical data, *International Journal of Geographical Information Systems*, Vol. 6, No. 2, pp. 87-104.

[9] Griffith, D. 1983, The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, Vol. 23, pp. 377-87.

[10] Hunter, G. & Goodchild, M. (in press), Dealing with uncertainty in spatial databases: a simple case study, *Photogrammetric Engineering and Remote Sensing.*

[11] Isaaks, E. & Srivastava, R. 1989, *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

[12] Mark, D. & Csillag, F. 1989, The nature of boundaries on 'area class' maps, *Cartographica*, Vol. 26, pp. 65-78.

[13] Monmonier, M. 1975, Maps, Distortion, and Meaning, *Association of American Geographers Resource Paper No. 75-4*, Washington D.C.

[14] Openshaw, S. 1989, Learning to live with errors in spatial databases, *Accuracy of spatial databases*, pp. 263-76, Taylor & Francis.

[15] Tobler, W. 1979, Cellular Geography, *Philosophy in Geography*, Dordrecht, pp. 379-86.

[16] Tomlinson, R., Calkins, H. & Marble, D. 1976, *Computer Handling of Geographical Data*, UNESCO, Paris.

[17] Veregin, H. 1989, Error Modeling for the Map Overlay Operation, *Accuracy of Spatial Databases*, Taylor & Francis, pp. 3-18.

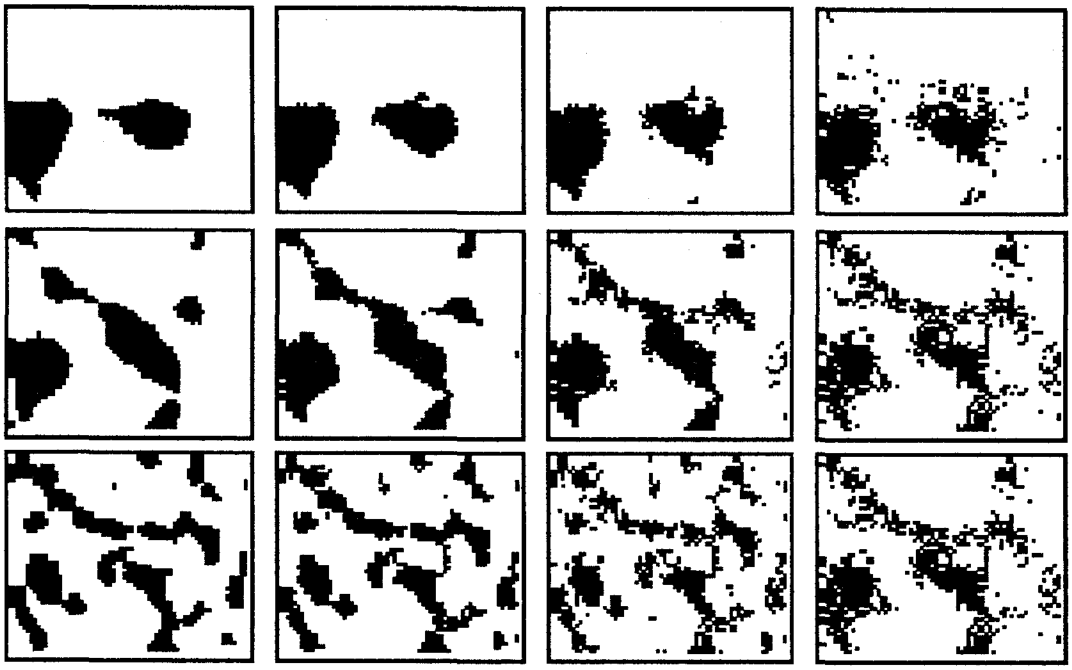

Fig 1 incl.ps

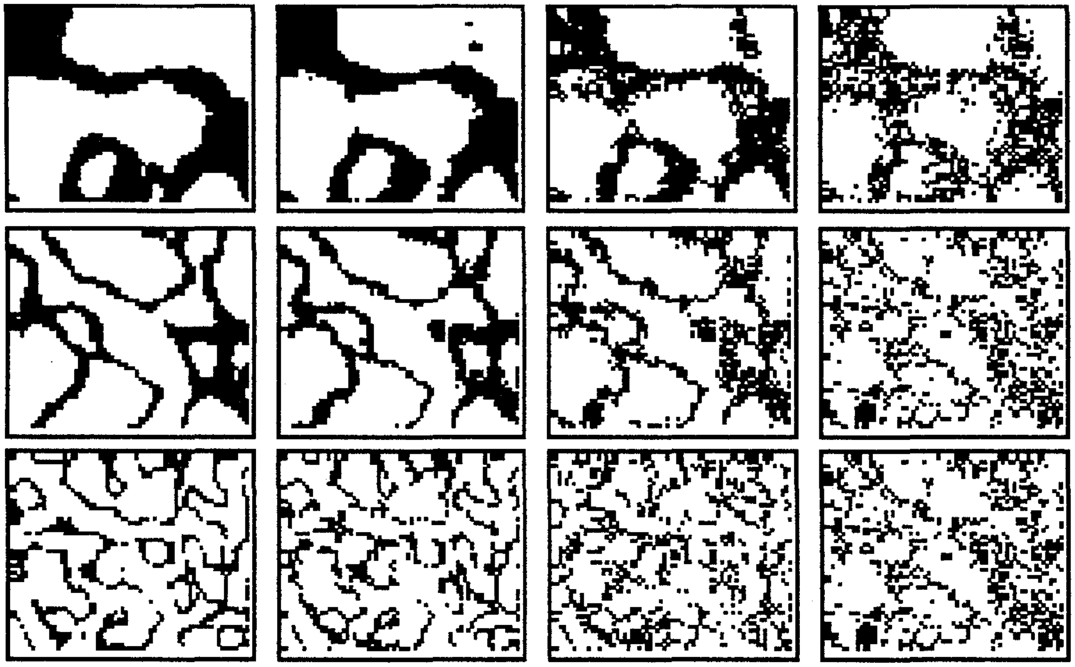

Fig 2 /inear.eps

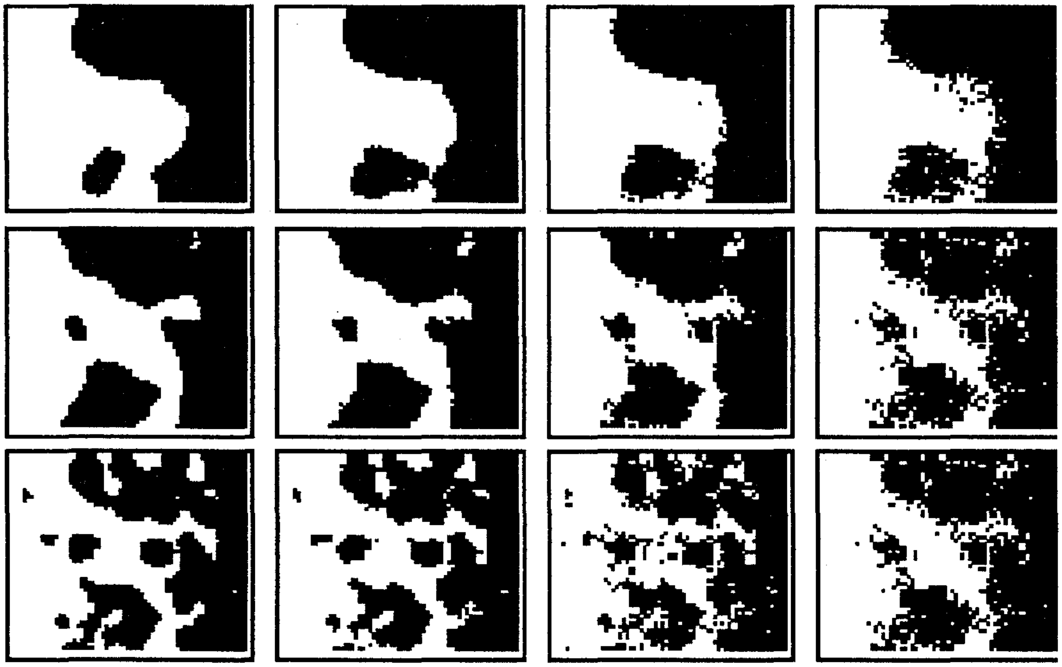

Fig 3 ramps.eps

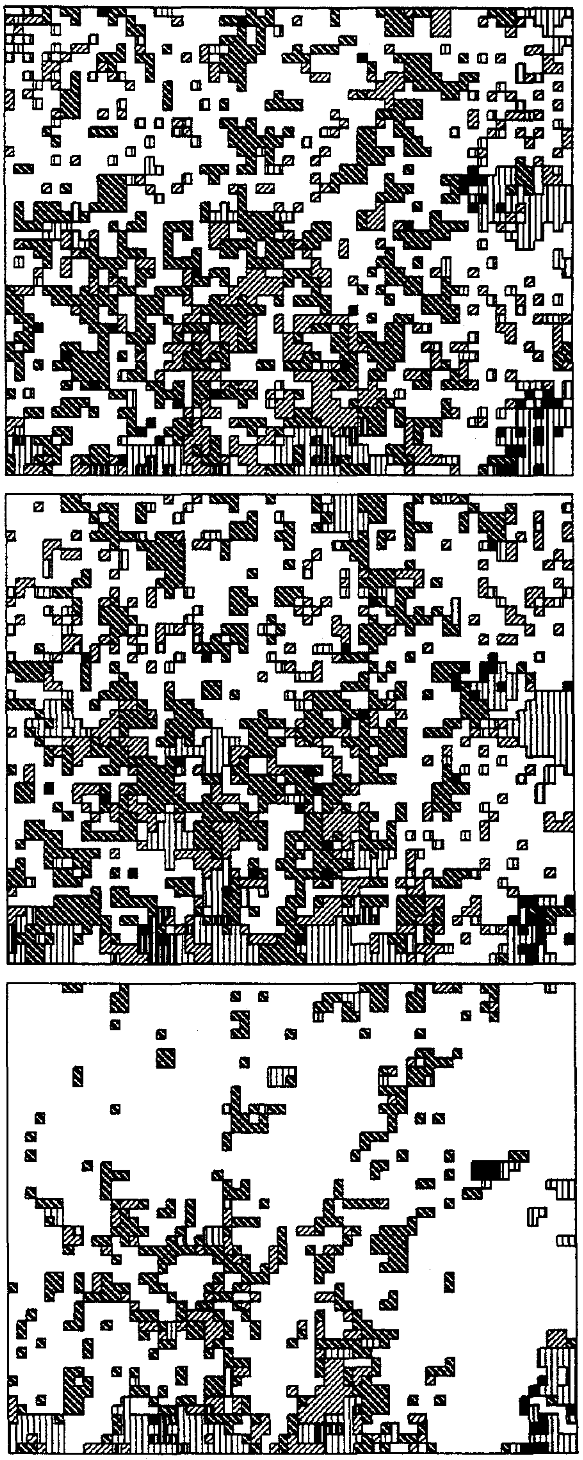

Fig 4 veg4.ps