Horwood Critique Article

A New Model For Handling Vector Data Uncertainty in Geographic Information Systems

Gary J. Hunter and Michael F. Goodchild

Abstract: Recently, the authors investigated the uncertainty associated with grid-cell data in geographic informa- tion systems (GIS), through the use of a model which permits a population of distorted (but equally probable) ver- sions of the same map to be generated and analyzed. While this model is easily applied to data such as digital eleva- tion models to help assess the uncertainty associated with the products derived from them, it was not directly applicable to vector data. However, the model has now been enhanced to allow for distortion of vector data through the creation of separate horizontal positional error fields in the x and y directions, which are then overlaid with the vector data to apply coordinate shifts by which new versions are created. By perturbing the data to generate dis- torted versions of it, the likely uncertainty of vector data in both position and attribute may be assessed in certain types of GIS outputs. This paper explains the background to the model and discusses its implementation and po- tential applications.

Geographic Information Systems (GIS) are now being widely implemented in the public sector for applications such as facilities management, land administration, resource monitoring and assess- ment, and emergency service command and control, whereas the private sector employs them for demo- graphic analysis, marketing and retail site selection. In April 1994, President Bill Clinton issued an executive order establishing the U.S. National Spatial Data Infra- structure in recognition of the economic, environmental

Gary Hunter is an assistant professor with the Department of Geomatics, and deputy-director of the Center for Geographic Information Systems and Modeling, at the University of Mel- bourne, Australia. He teaches undergraduate and postgradu- ate courses in land law and development, cadastral survey- ing, spatial data algorithms, and geographic information systems implementation and management. His primary re- search interest lies in the study of spatial data accuracy and he is the winner, with co-author Michael Goodchild, of the Horwood Critique Prizes at the URISA '93 and '95 confer- ences for papers dealing with this subject.

Michael Goodchild is professor of geography at the Univer- sity of California, Santa Barbara, and director of the National Center for Geographic Information and Analysis. He was edi- tor of *Geographical Analysis* for several years and serves on the board of six other journals and book series. He is well known for his interest in quality and accuracy issues in GIS, and is the editor of the well-known book, *Accuracy of Spatial Data- bases.*

and social importance of providing low-cost, accurate, and easily obtainable spatial data to all sectors of the community. Other countries, such as Australia and New Zealand, have followed the progress with keen in- terest and are considering adopting similar programs. While the issues of cost and easy access to data have been generally overcome through government policies and technological advances, the accuracy issue is still causing considerable problems within the industry.

From a historical viewpoint, as users became more experienced with GIS during the 1970s and 1980s, there gradually arose a critical awareness of the fact that of- ten they did not know how accurate their system out- puts were, and whether or not the derived information actually satisfied their accuracy requirements. This predicament has been caused not only by the false sense of security that computer technology can induce in the unwary, but also by the lack of theoretical models of spatial data error and the means to communicate it. The situation has now reached the point where, for government agencies which base their regulatory deci- sions upon spatial information or in cases where they or private companies sell data for commercial return, there is the growing risk of litigation by aggrieved par- ties seeking compensation for poor decisions based on data inaccuracies or data that had insufficient accuracy to meet their requirements. Software developers and vendors may also be affected since the algorithms they encode in their products can have the potential to in- duce additional error. Thus, the accuracy issue should

be of serious concern to all sectors of the geographic information industry.

In recent years, most international spatial data transfer standards have adopted mandatory data-quality reporting provisions (Moellering 1991), which will no doubt help address the problem by ensuring that data providers truthfully label their products in such a way that users can assess their fitness for use. However, while this approach has its merits, there is a presumption that the necessary tools for modeling and communicating spatial data error already exist. Unfortunately, this is not the case and a considerable amount of research still remains to be conducted. Goodchild (1993), for instance, suggests there are only half a dozen commonly accepted models of spatial data error. And Hunter and Beard (1992) have identified at least 150 potential error sources of which we have little or no current understanding.

To help deal with this problem, the authors have developed a model of uncertainty for dealing with spatial data; however before discussing it, some explanatory remarks are required regarding the use of the term 'uncertainty.' In the context of geographic data, it is argued that there is a clear distinction between 'error' and 'uncertainty,' since the former implies that some degree of knowledge has been attained about differences (and the reasons for their occurrence) between the results or observations and the truth to which they pertain. On the other hand, 'uncertainty' conveys the fact that it is the lack of such knowledge which is responsible for hesitancy in accepting those same results or observations without caution, and often the term 'error' is used when it would be more appropriate to use 'uncertainty.'

The uncertainty model that has evolved can be defined as a stochastic process capable of generating a population of distorted versions of the same reality (such as a map), with each version being a sample from the same population. The traditional Gaussian model (where the mean of the population estimates the true value and the standard deviation is a measure of variation in the observations) is one attempt at describing error, but it is global in nature and says nothing about local variations or the processes by which error may have accumulated.

The model adopted in this research is viewed as an advance on the Gaussian model since it not only has the ability to show local variation in uncertainty, but also has the advantage of being able to display the effects of error propagation resulting from the various algorithms and process models that have been applied—even though we do not possess propagation models *per se*. This latter point is particularly important to users, since many software vendors do not divulge the algorithms used in their packages for commercial reasons—which prevents formal mathematical error propagation analy-

sis from being undertaken. By studying different versions of the products created by the model, it is possible to see how differences in output are affected by variations in input.

The model was originally designed by Goodchild *et al.* (1992), and its use in assessing the uncertainty of products derived from grid cell data has been reported in Hunter and Goodchild (1994) and Hunter *et al.* (1994). However, a limitation of that initial version of the model was its inability to represent uncertainty in vector data, and this paper describes an enhanced version of the model which is capable of producing distorted versions of point, line and polygon data. This paper is structured such that the concepts underlying the model are first introduced, followed by the issues affecting its implementation, and then finishes with a discussion of how it might be applied in practice.

Overview of Model

Extending the original grid cell uncertainty model to cater to vector data involves the creation of two separate, normally distributed, random error grids in the x and y coordinate directions. When combined, these grids provide the two components of a set of positional error vectors regularly distributed throughout the region of the data set to be perturbed. The assumptions made by the authors are: 1) that the error has a circular normal distribution, and 2) that its x and y components are independent of each other. The grids are generated with a mean and standard deviation equal to the producer's estimate for positional error in the data set to be perturbed (which is a fundamental prerequisite for the model to be applied). These error estimates, for example, might come from the residuals at control points reported during digitizer setup, or from existing data quality statements such as those that now accompany many spatial data sets.

By overlaying the two grids with the data set to be distorted (containing either point, line or polygon features), x and y positional shifts can be applied to the coordinates of each node and vertex in the data set to create a new, but equally probable, version of it (see Figure 1). Thus, the probabilistic coordinates of a point are considered to be $(x + $ error, $y + $ error). With the new distorted version of the data, the user then applies the same set of procedures as required previously to create the final product, and by using a number of these distorted data sets the uncertainty residing in the end product is capable of being assessed. Alternatively, several different data sets may be independently distorted (each on the basis of its own error estimate) prior to being combined to assess final output uncertainty. While the new model does require an initial error estimate for creation of the two grids, it is the resultant uncertainty

FIGURE 1. The proposed model of vector data uncertainty uses normally distributed, random error grids in the x and y directions to produce a distorted version of the original point, line or polygon data.
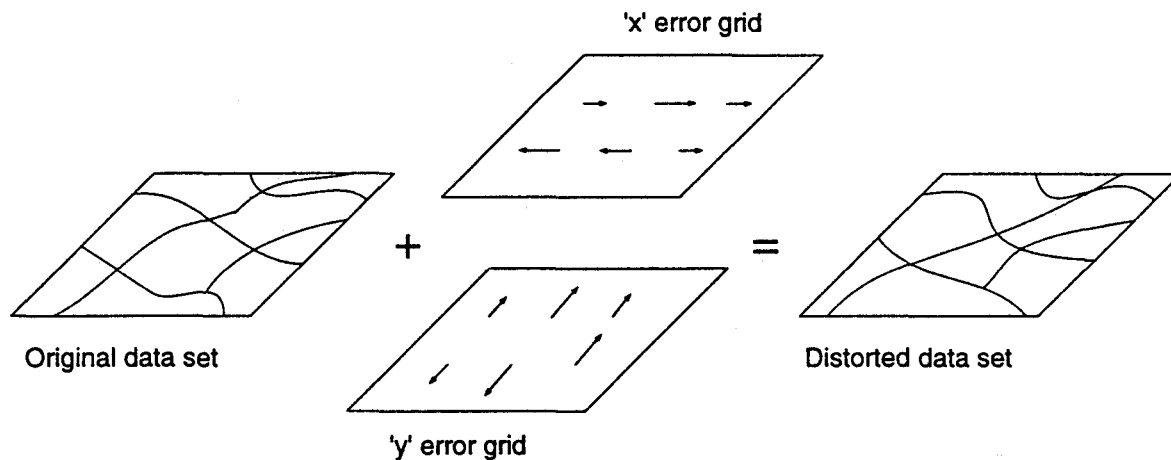
arising from the use of perturbed data due to simulation (in conjunction with the spatial operations that are subsequently applied) which is under investigation, and hence its label as an 'uncertainty' model.

Implementation Issues

Generation of the Error Grids

The two error grids are created independently by populating them with normally distributed values having a mean and standard deviation equivalent to the producer's horizontal error estimate for the data set being perturbed. Usually, the mean will equal zero and the standard deviation is the same in the x and y directions. In some GIS packages, provision already exists for creating grids that are normally distributed (such as the NORMAL function in the Arc Grid software). The two grids, which are initially assigned zero as their coordinate origin and have unit separation distance between points, are then georeferenced via a 2-dimensional coordinate transformation to achieve the required separation distance and to ensure they completely overlap the data set to be perturbed (for example, by using the SHIFT function in Arc Grid which employs the new lower left coordinates of the grid and required point spacing as its arguments).

Choice of Error Grid Separation Distance

While the choice of separation distance between points in the error grids is arbitrary, if it is larger than either the x or y components of the minimum distance between any two neighboring features in the data set to be distorted (regardless of whether they are represented by points, lines or polygons), then local positional shifts applied to those points will be equal and no longer independent—causing unwanted local spatial autocorrelation to be introduced to the data (see Figure 2). Thus, the point spacing in the error grids should be at least as small as the minimum x or y distance between observed features in the data set, although spurious data should obviously be discounted from this estimate.

As a rule of thumb, it is suggested that the spacing be equal to or less than 0.5 mm at the scale of the map from which the data originated, which is a common estimate of relative positional accuracy. This translates into 0.5m at a scale of 1:1,000; 5m at 1:10,000; and 50m at 1:100,000. Where little is known about the data set's origin, users should select a separation distance smaller than they would care to consider—given the nature of the data and the application concerned. For example, with a vegetation boundary data set it might be considered that individual boundary segments or polygon widths less than 20m, while not necessarily spurious, will not practically affect the outcome of any analysis to be conducted. Thus, selection of an error grid separation of 20m is reasonable even though the positional shifts applied to features that are less than this distance apart will be similar in magnitude, and therefore highly correlated. Alternatively, a user may decide to set the grid separation distance equal to the standard deviation of the horizontal positional error.

Preserving Topological Integrity Between Distorted Features

To preserve topological integrity between features upon distortion, some means of adjustment must be introduced to control the magnitude of positional shifts between neighboring points in the error grids. If this does

FIGURE 2. When choosing the error grid separation distance, if the spacing (d) between points in the error grids is greater than the separation or size of the smallest features, then unwanted local spatial autocorrelation between shifts at neighboring points may result.

Small polygon (P₁) undergoing distortion

Combined error grid showing shifts at each grid point
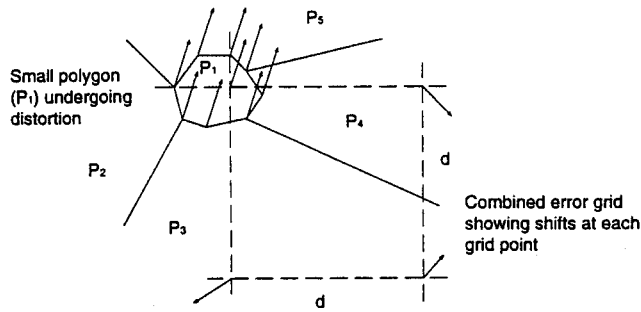
not happen, there is a likelihood that neighboring points in the original data set may be transposed in position—causing unwanted "fractures" in the feature structure (see figure 3a). This could also happen with neighboring contour lines, or with points on either side of a long, narrow polygon—causing contours to overlap and a "figure 8" transposition to occur with new polygons being formed. Intuitively, these alterations to feature structure should not happen and are considered unacceptable. Originally, the authors considered using spatial

autocorrelation to constrain the shifts in the x and y error grids, but this would have had the effect of unnecessarily altering all shifts in the two grids when in all probability only a relatively small number of "fractures" actually needed to be adjusted. Accordingly, a localized filter was seen as the preferable solution.
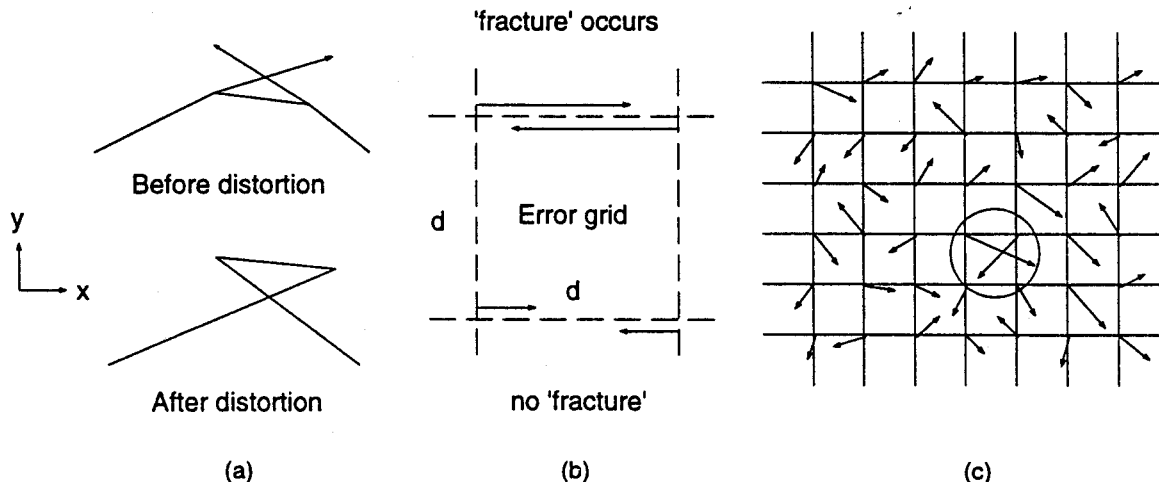
Therefore, a routine was developed to test the difference between consecutive pairs of points (in horizontal or row sequence for the x grid, and vertical or column sequence for the y grid) to determine whether the absolute value of the difference between them was greater than the chosen grid separation distance (Figure 3b). If so, then a "fracture" is possible at that location (if two data points are nearby) and a filter must be applied to average out the shift values on the basis of neighboring grid points. This testing between neighbors in the x and y grids is iterative and proceeds until no "fractures" exist in either error grid.

For example, given consecutive shifts Δx_0, Δx_1, Δx_2 and Δx_3 in any row of the x error grid (and testing the difference between the two middle shifts), then there is potential for a "fracture" to occur if $\Delta x_1 - \Delta x_2 > d$ (where d is the grid separation distance). The adjusted values of Δx_1 and Δx_2 are computed by equation (1):

$$\Delta x_1' = \frac{\Delta x_0 + \Delta x_1 + \Delta x_2}{3}$$

$$\text{and } \Delta x_2' = \frac{\Delta x_1 + \Delta x_2 + \Delta x_3}{3} \tag{1}$$

FIGURE 3. In (a), unconstrained shifts between neighboring error grid points may cause unacceptable "fractures" or transposition of features—thereby damaging the topological integrity of the data set. In (b), "fractures" occur when the difference between neighboring error grid points is larger than their separation distance (d). In (c), a "fracture" is circled in which case the x and y shift values must be filtered on the basis of neighboring grid point values.

'fracture' occurs

Before distortion

y

x

After distortion

d   Error grid

d

no 'fracture'

(a)   (b)   (c)

The justification for the filtering process is that although the shifts are intended to be normally distributed, it should be remembered that the choice of the normal distribution is only an assumption and we must accept that there may be some faults in its selection as a theoretical model. For example, when statisticians analyze human intelligent quotient (IQ) data using a normal distribution with a mean of 100 and standard deviation of 20, the distribution is deliberately truncated at zero to disallow values less than zero that intuitively should not occur—even though they will be predicted by the model. Similarly, the filtering of neighboring error shift values to avoid 'fractures' is simply a truncation of the distribution to cater for a minority of extreme cases.

Calculation of Positional Shifts for Data Points not Coinciding with the Grids

Inevitably, it is expected that few if any nodes or vertices in the observed data set will coincide with the error grid points, and a technique is required for calculating x and y shifts based on the values of neighbors in the error grid. Accordingly, a simple bilinear interpolation procedure (Watson 1992) is proposed in which the x and y shifts assigned to each data point are calculated on the basis of the respective shifts of the four surrounding grid points (Figure 4).

Transfer of Positional Shifts to the Data Points

The final implementation issue concerns the means by which the positional shifts are transferred to the data points. At this stage, the likely solution is to take each point in the data set to be distorted and label its position as the "From" coordinates, to which the x and y shifts are added to give the "To" coordinates. Where topology

FIGURE 4.   The x and y shifts for a data point not coinciding with an error grid intersection point are calculated based on the four surrounding grid values using bilinear interpolation.
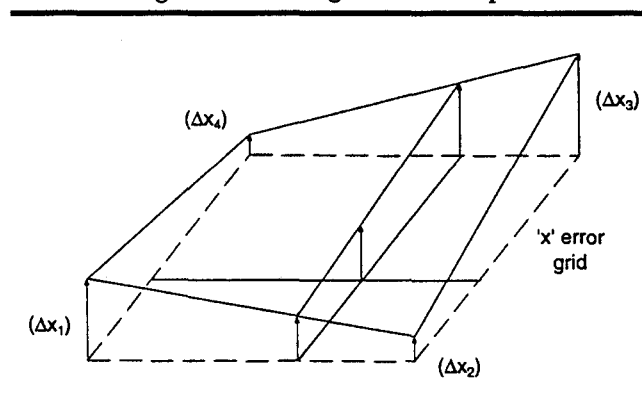
is present, it will need to be reconstructed after distortion. At this time, the transfer procedure has not been implemented and testing will be required to develop efficient computational algorithms. Clearly, the need may arise for high-performance computers to be employed, particularly where many thousands of nodes and vertices are to be perturbed, and current research into GIS and supercomputers (such as described in Armstong, 1994) is being investigated by the authors.

Summary of the Model's Operation

The following steps summarize the considerations that have been discussed with respect to implementing the model in practice:
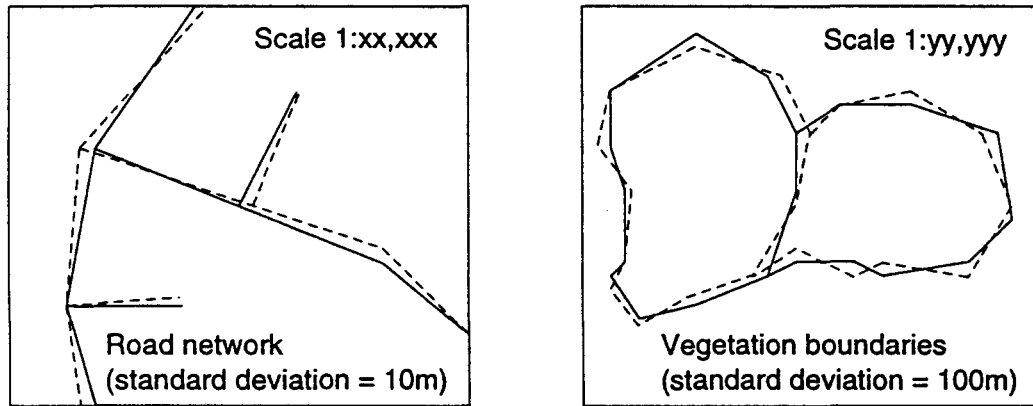
Step 1:   Determine the separation distance required for the error grids and the coordinate extent that the grids will need to cover.

Step 2:   Generate two error grids populated with independent, normally distributed values.

Step 3:   Adjust the dimensions of the grids to give the required spacing between points and transform their coordinate origins to agree with the data set being perturbed.

Step 4:   Test the grids for "fractures" and filter error shift values as necessary. Repeat until no "fractures" exist.

Step 5:   Taking each point in the observed data set in turn, calculate the positional shifts in x and y to be applied based on the neighboring error grid values.

Step 6:   Update the coordinates of each point using the shifts calculated in Step 5.

Step 7:   Reconstruct the topology of the distorted grid, to produce a distorted version of the original data set.

Potential Applications

The potential applications of the model are considered to lie in several areas. First, from an educational perspective, a family of distorted but equally probable versions of the same data set could be used to illustrate the uncertainty that might be expected when interpreting the data. This could be a useful approach for data producers to adopt when explaining the meaning of their error statistics to potential users who may not necessarily understand the statistics provided—especially now that the use of GIS has become so diverse and there are many new users without sound analytical backgrounds in this field. These visual statements could form part of the data quality reports (Figure 5).

For operational purposes, the model could be used to determine the uncertainty of attributes derived from area and length estimates due to positional error in the original data. For instance, different versions of a road centerline database could be created to test the uncertainty in travel time and routing applications by run-

FIGURE 5. In an educational role, the model could be used by data producers to generate visual samples of perturbed data sets for inclusion in their data quality reports to help convey the meaning of their error statistics to users.

Scale 1:xx,xxx

Road network
(standard deviation = 10m)

Scale 1:yy,yyy

Vegetation boundaries
(standard deviation = 100m)

ning the same problem with perturbed data sets to assess which solutions have the best overall results. In other applications, more than one data set may need to be perturbed. For example, if the problem is to determine the economic effects of flooding based on different land uses, property values, soil types and flood zone ratings, then each vector data set could be perturbed according to its own positional error estimate before being overlaid to identify land which is at highest risk of flooding and the potential monetary loss. By running the model a number of times, the resultant variation in the financial amounts involved can be determined under conditions of uncertainty.

Finally, the model may also be useful as a means of densifying or "ungeneralizing" data, if we assume that for certain types of data not only might the endpoints of lines be distorted, but also the positions of intermediate

points along them. While this approach would be unsuitable for perturbing, for example, land parcel boundaries which are necessarily defined by their endpoints, on the other hand it might provide a more realistic method of representing the boundaries of data subject to natural variation. This could be achieved by first placing additional vertices along each line segment (the usual means of densifying lines), and then applying the model to provide shifts at all node and vertex locations (Figure 6).

Conclusions

In this paper the authors have presented a model for handling uncertainty in vector data which is an enhancement of their earlier work that dealt with grid cell data only. The model permits different, but equally

FIGURE 6. The model might also be employed to randomly densify features, but its use is not suited to all data types. For example, it would not be applicable to cadastral boundaries, as in (a), where only the endpoints of lines should be distorted, however for boundaries subject to natural variation, as in (b), independent distortion of intermediate points along a line may provide a useful means of densifying features.

New lines

Original lines

(a)

New lines

Original lines

(b)
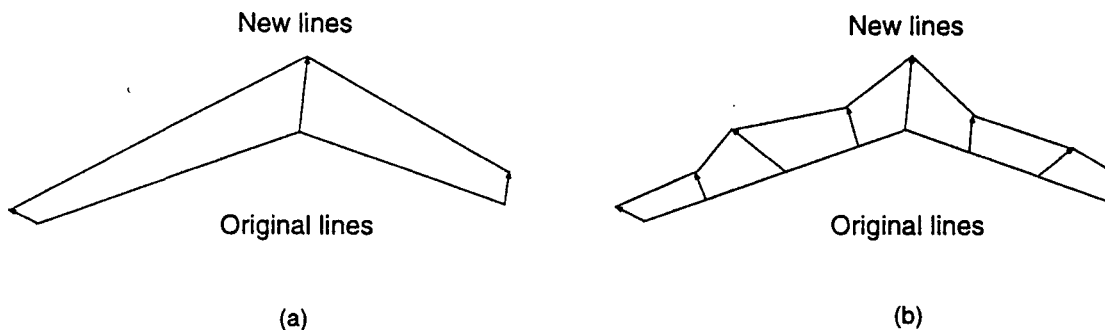
probable, versions of point, line and polygon data sets to be created via the introduction of distortion grids in the x and y directions, which in turn are used to apply coordinate shifts to every feature in the data to be perturbed. The paper discusses the underlying concepts, the implementation issues affecting the model, and the way in which it may be applied in practice. In particular, it might be used to show positional uncertainty effects alone, or else their subsequent effect upon derived attributes based (for instance) on area and length calculations. It might also prove useful as a non-linear means of randomly densifying line and polygon features. While the model does not purport to be able to deal with all cases of uncertainty in vector data, the authors believe it nevertheless provides a useful advance on the current body of knowledge and techniques in this field until such time as formal models of spatial data error are more widely developed and accepted.

Acknowledgement

References

Armstrong, M.P. 1994. "GIS and High-Performance Computing," In *Proceedings of the GIS/LIS 94 Conference*, Vol. 1:13, Phoenix, Arizona.

Goodchild, M.F., G. Sun and S. Yang. 1992. "Development and Test of an Error Model for Categorical Data," *International Journal of Geographical Information Systems*, Vol. 6(2): 87–104.

Goodchild, M.F. 1993. "Data Models and Data Quality: Problems and Prospects," In *Environmental Modeling with GIS*, M.F. Goodchild, B.O. Parks and L.T. Steyaert (eds), pp. 94–103, Oxford University Press: New York.

Hunter, G.J. and K. Beard. 1992. "Understanding Error in Spatial Databases," *The Australian Surveyor*, Vol. 37(2): 108–119.

Hunter, G.J and M.F. Goodchild. 1994. "Design and Application of a Methodology for Reporting Uncertainty in Spatial Databases," *Proceedings of URISA '94* , Vol. 1: 771–785, Milwaukee, Wisconsin.

Hunter, G.J., M.F. Goodchild and M. Robey. 1994. "A Toolbox for Assessing Uncertainty in Spatial Databases," *Proceedings of AURISA '94* , Vol. 1: 367–379, Sydney, Australia.

Moellering, H. (ed.). 1991. *Spatial Database Transfer Standards: Current International Status* 320 pp., Elsevier: New York.

Watson, D.F. 1992. *Contouring: A Guide to the Analysis and Display of Spatial Data*, Pergamon/Elsevier Science: Tarrytown, New York.