# MEAN OBJECTS: EXTENDING THE CONCEPT OF CENTRAL TENDENCY TO COMPLEX SPATIAL OBJECTS IN GIS

Michael F. Goodchild, Thomas J. Cova, and Charles R. Ehlschlaeger
National Center for Geographic Information and Analysis, and
Department of Geography, University of California
Santa Barbara, CA 93106-4060
(805) 893-8049, (805) 893-8652
good@geog.ucsb.edu, cova@geog.ucsb.edu, chuck@ncgia.ucsb.edu

## ABSTRACT

Concepts of mean, standard deviation, probability density, and confidence limits are well developed and generally familiar for scalar measurements, but fail to generalize well to objects in two dimensions. In the case of points, there are several competitors for the equivalent of the mean. We review the analogs of these scalar descriptions for points, lines, areas, and tessellations, and their implementation in GIS. Well-known statistics such as the Perkal epsilon band are placed within this general framework. We present experimental methods for robust estimation of the equivalent of the mean and standard deviation in some of these cases, and review experience with their use and associated unsolved issues. Although the mean of a distribution of a scalar measurement is almost always a possible member of the distribution, we present arguments for the general proposition that the mean of a population of complex two dimensional objects is not itself a member of the population. This paper discusses the implications of these ideas for GIS and the mapping sciences.

## INTRODUCTION

The mean or average is a well-known and well understood concept, readily calculated for any set of simple measurements, such as areas, weights, or values. It is often described as a measure of central tendency, or the central value around which a sample of measurements is distributed. It can be thought of as the single value most descriptive of the sample. One important restriction on the mean that is often relevant to GIS applications is that the measurements must be made on a measurement scale having at least interval properties; it makes no sense to try to calculate the mean of a categorical variable such as land cover class.

There are many reasons for wanting to generalize the concept of the mean to geographic features. We frequently encounter different versions of the same feature when different digital sources are available, and it may be desirable in such cases to merge the knowledge of feature geometry that is present in each source. The term "data fusion" has been used in this context. Sometimes one version is believed to be more accurate than another, suggesting that they might be weighted differently when it is desirable to combine information from both. The same concept of fusion occurs in the overlay of polygons, when lines from different layers coincide in reality, but their digital versions create slivers. In such cases we might think of sliver removal as conceptually equivalent to finding a mean.

Besides its value as a descriptive parameter, the mean is also the basis of many inferential tests, an important parameter of many statistical distributions, and thus key to many forms of simulation. For example, given a mean and standard deviation, and with the assumption that measurements have a normal distribution, it is possible to generate samples of any size, and to use them in randomization tests and other Monte Carlo experiments. The equivalent in two dimensions would be the ability, given

suitable parameters and assumptions about distributions, to simulate multiple versions of geographic features, in order to analyze the propagation of error and uncertainty, for example (Goodchild and Gopal, 1989). These two operations of estimation and simulation form a loop, so one test of these methods is that just as for univariate measurements, one should be able to take a mean and standard deviation, simulate a sample suitably dispersed about the mean, and then estimate approximations to the original mean and standard deviation from the sample. The statistics of this process are well-known for univariate measurements, particularly for the normal distribution.

Unfortunately, the familiar parameters of univariate statistics—mean, standard deviation, variance, etc.—do not generalize well to two dimensions. Maps and geographic data sets are derived from scalar measurements, but the process is often complex, and it is generally not possible to regard a geographic data set as an ensemble of independent measurements. The purpose of this paper is to explore the degree to which generalization is possible, with special reference to the mean. The paper begins with a review of previous work in this area, and extends the motivation. We then present experimental methods for estimating the mean line, based on ideas originally presented by Edwards (1994a,b). The final section of the paper inverts the approach, by using the mean together with a measure of dispersion and a stochastic process to simulate a sample of lines.

Conceptually, there is some overlap between the framework of this paper and that of cartographic generalization, and its inverse. The mean of a set of features should preserve the signal present in each sample, but attempt to remove the noise, or any other aspects of a line that are inconsistent with the remainder of the sample. Thus an averaging process can be expected to smooth. In that sense, the mean line can be conceived as analogous to a cartographic generalization, and its inverse to the addition of detail (Dutton, 1981), although we would not propose the techniques in this paper as suitable for automated feature generalization.

This point has an interesting corollary. For univariate measurements, the mean is itself a possible measurement, although it may be determined with greater accuracy through the process of estimation. But for geographic line and area features, the mean feature can be expected to be more generalized than the observed samples. The mean of a set of shoreline estimates, for example, is not itself a member of the population of possible shoreline estimates. This point has profound implications for kriging, where the problems it creates for GIS suitability analysis have already been pointed out by Englund (1993).

## BACKGROUND

Geographic data can be partitioned into two broad categories, depending on the conceptual understanding of the data's meaning. Fields are defined as variables having single values at every point in the geographic plane, and can be further subdivided depending on whether the variable is measured on a nominal (categorical) or interval/ratio scale. Digital representations of fields include TINs, rasters, point grids, irregularly spaced points, digitized contours, and non-overlapping polygons (polygon coverages). Although the digital representation of a field involves discrete geometric objects (points, polylines, or polygons), these objects generally have no real-world meaning. Instead, fusion of two fields is better conceptualized as a fusion of the two field variables. In the case of interval/ratio fields, this might be done by taking a suitably weighted mean value; in the case of categorical fields, by taking the modal value. For example, the mean of two elevation fields $z_1$ and $z_2$ might be computed as a field $(z_1+z_2)/2$.

Although polygon coverages (irregular tessellations) are normally associated with the representation of fields, such as population density or land cover, it is possible to interpret the notion of mean objects in this case. It is common, for example, for states to be subdivided into regions, and for different state agencies to use different, incompatible schemes. One common consequence of this for GIS users is the need to transfer attributes from one scheme to another; for example, to transfer unemployment statistics collected by the state labor agency and tabulated for their regions, to the regions used by the state housing agency. In such cases the idea of a mean regionalization may be useful, as a basis for comparison, or as the closest approximation to a consensus.

Consider a set of N geographic building blocks, such that every regionalization is some combination of these blocks. Suppose there are M regionalizations. Count the number of times $x_{ij}$ that contiguous building blocks i and j are allocated to the same region; the largest possible value of $x_{ij}$ is M ($x_{ij}$=0 for non-contiguous blocks). Scan the N by N matrix x for the largest value; join the corresponding building blocks; zero the value; and repeat until a predetermined number of regions has been created. We have used this method to find the mean regionalization of seven California state agencies, and Monmonier has described a similar approach in a different context (Monmonier, 1982).

The primary concern in this paper is with the other class of geographic data, which originates in what has been termed the "entity view", and consists of sets of discrete point, line, or area objects, possibly overlapping, and surrounded by empty space. The objective is to explore generalizations of the mean to sets of points, polyline representations of lines, and sets of polygons.

In the case of points, several approaches can be used to generalize the mean to two dimensions. If one takes the equation defining the univariate mean, a simple generalization uses the same equation twice to calculate a mean for each coordinate, giving a point generally known as the centroid. Alternatively, the variational property of the mean, that the sum of squared distances from it to each point is minimized, can be used to define a two-dimensional generalization, and it is easy to show that this also finds the centroid. The centroid is not the point that minimizes the sum of distances, but then neither is the mean in one dimension, this property belonging to the median.

The statistics of the centroid have been explored extensively, and are used widely in surveying adjustment. Measures of dispersion about the mean, the two-dimensional equivalent of the standard deviation, are commonly used as measures of point positional accuracy. The centroid has also been used as a useful summary of the geographic position of a point set, particularly in capturing the "center of population", or the differences between two subsets of points, or two times. In summary, the concept of the mean generalizes easily to point sets.

In the case of lines, generalization is more problematic. Although they may be conceived as continuously curved in reality, lines are most commonly represented in spatial databases as polylines, or straight lines between points. One possible generalization would be to focus on the points of a polyline, treat them as repeated measurements of the same true polyline, and calculate mean positions. In some cases, where the line in reality is itself a polyline, and the truth is defined as straight mathematical lines between points, as is often the case for surveyed boundaries, each estimate of the line can be treated as an assemblage of point estimates. In other cases, such as rivers or coastlines, this model fails because it becomes generally impossible to match points between polylines. In general, the selection of points in a polyline is itself part of the sampling process, so any procedure for estimating a mean line must attempt to approximate some true, continuous line rather than its polyline representation.

The removal of polygon slivers has already been cited as a possible application of a mean line. A common approach is to assume that both edges of the sliver are equally likely estimates of the true position of the line, to select one edge arbitrarily, and to delete the other. A less arbitrary approach would be to identify the "medial axis transform", defined in this case as the set of points within the sliver polygon that are equidistant from the two edges (the medial axis transform is normally defined as the set of points equidistant from any two points on the polygon; Lee, 1982; Pavlidis, 1982). While this method guarantees a mean line inside the sliver, it has the unfortunate property that the set of points is a mix of straight line segments and parabolic curves. It would be possible to weight the lines differently, but this approach does not generalize well to more than two lines.

The most common approach to defining the dispersion of line features around a mean is often attributed to Perkal (1966), and termed the "epsilon band". Goodchild and Hunter (1995) have proposed a robust statistical implementation in which epsilon is a function of cumulative percentiles of line length. The 95th percentile, for example, is defined as a band of width $e_{95}$ about the mean line that contains 95% of the length of the sample line; as such, it can be measured for a single sample line by comparing it to some source line of higher accuracy.

The requirements of a mean line or area would seem to be as follows: 1. Any method of estimation should be robust, capable of taking a wide range of line or area inputs and producing an acceptable output that is topologically a line or area respectively; 2. Although the inputs will be polylines, the output should be conceptually closer to a true curve. Since the output must also be a polyline or polygon as it will be in digital form, this might be interpreted as meaning that the output should be denser than the inputs; 3. The procedure should emulate the unbiassedness criterion of statistical estimation; that is, the estimate should tend to the true mean as the sample size increases.

Consider, for example, the following procedure. To estimate a mean area, calculate a field whose value at any point is the proportion of sample polygons that contain the point. Then find the 50% isoline of the field (Shi and Tempfli, 1994). To estimate a mean line, first extend each polyline to form two half-planes, based on the direction of the line at each end; define a field whose value at any point is the proportion of polylines for which the point is in one of the half-planes; finally, find the 50% isoline of the field. Although the method is robust for areas (less so for lines because of the need for arbitrary extension to create half-planes), it does not satisfy the first criterion above since it is possible for the output to include isolated islands, and holes in the case of areas.

## ESTIMATING MEAN LINES

As discussed, it not difficult to conceive of an infinite population comprised of all possible polyline representations of a particular geographic feature. By analogy to the univariate case, this population would be distributed around a true or mean line, with a dispersion described by a parameter analogous to standard deviation. In this section, we discuss a technique for estimating the mean based on a method described by Edwards (1994a,b).

Each line j, j=1,n, is first given a parametric representation, by expressing the coordinates of its position (x,y) as functions of distance s along the line, $x_j(s)$ and $y_j(s)$. For convenience, the scale of s is normalized to the interval [0,1], where s=0 is the beginning node of the line and s=1 the ending node.

If we are interested in calculating a mean representation for a sample of curves, then the end points of the curves must correspond, to some degree, and be in close proximity. Areas must be handled by identifying corresponding starting and ending points, and adopting the same clockwise or anticlockwise order for each sample feature. This requirement will be relaxed in a subsequent section, but at this point we will assume that the sample of lines is taken to represent the same geographic feature, and they have a reasonably high degree of linear correspondence. Given that this is the case, and since both x(s) and y(s) are single-valued functions of s, a mean curve can be described by averaging x and y over j at each value of s. The procedure guarantees that the output is topologically a line or area, as appropriate, but it does not prevent the existence of loops.

In reality, lines and areas will be represented as polylines and polygons, so x(s) and y(s) will be evaluated only at sampled points, and these points will be unique for each sample feature. A simple procedure is to represent both $x_j(s)$ and $y_j(s)$ as polyline functions, allowing the value of the x and y functions to be determined by linear interpolation at any value of s. Then the position of the mean line can be evaluated at every sampled value of s, ensuring that the mean line's polyline representation is denser than the inputs as required above. In general, each point on the output polyline will be the result of averaging one observed value of x(s) with n-1 interpolated values.

The procedure is depicted in Figure 1 and the steps of the procedure are given as follows:

For a given set of polylines that represent the same geographic line or area feature:

1. Identify corresponding end points s=0 and s=1.
2. Calculate the s value for each intermediate point in all the polylines.
3. Sort the points in all lines by their s value.
4. For each point (in ascending order of s):
   4a. Obtain values of $x_j(s)$ and $y_j(s)$.
   4b. Calculate the mean over j of x(s) and y(s), using weights for each line as appropriate; standard deviation in x and y directions can also be calculated.
5. Connect the resultant mean points to form a new "mean polyline".

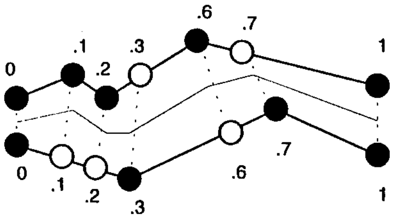

Figure 1: An example of the mean polyline interpolation process for two equally weighted polylines. The solid points are input and the white points are interpolated.

The method offers two alternatives for defining standard deviation: as an attribute of each point on the output polyline, as suggested above; or as a single estimate for the entire line. Because each input line is represented as a polyline, we can expect the resulting standard deviation to be an underestimate of the value that would have been obtained had the sampling density along the input lines been higher.

As noted earlier, the need to define matching start and stop points, s=0 and s=1, for each line can present a significant problem. Edwards (1994a,b) discusses various approaches, including matching the positions of well-defined points, such as prominences on shorelines or river crossings on roads, but this requires human intervention in most cases.

Alternatively, one might try to match the lines by searching for closest points; the location of s=0 on line 2 could be defined as the closest point to the location of s=0 on line 1. Figure 2a shows an example of this problem.

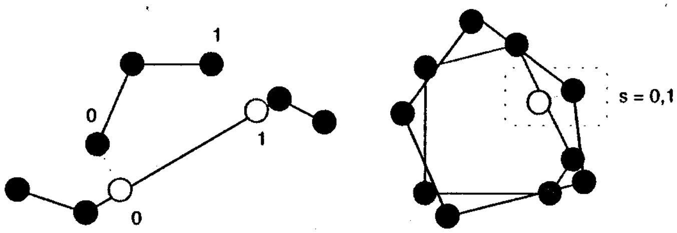

Figure 2: An example of the need to search for matching start points on lines (a) and polygons (b).

For polygons, the issue of point matching is further complicated. In this case, any point in a given polygon can be used as s=0, and a search made for the closest match in the other polygons. In this case, s=0 and s=1 coincide, and the mean polygon algorithm handles this exception accordingly. Figure 2b shows an example of how s might be defined for a set of polygons.

Figure 3a shows an example mean line (darkest line) calculated from three equally weighted versions of the same shoreline, and Figure 3b shows a mean line that is progressively weighted (grey lines) towards one of the three shorelines used to calculate it (darkest line). Note the sharply different levels of generalization of the three lines. Figure 3b depicts a problem that arises in calculating a mean line across levels of generalization. The method assumes that the process of generalization produces a uniform shortening of the line, so that points of equal s still correspond. However, many generalizations will result in deletion of features such as indentations, and thus non-uniform shortening. Figure 3b shows several instances where averaged features are offset as a result. As Edwards (1994a,b) suggests, a simple solution is to add more tie points between the lines, but whether this must be done by operator intervention or can be automated remains to be seen. In principle, tie points could be added automatically for every point based on minimum distance to the other line.

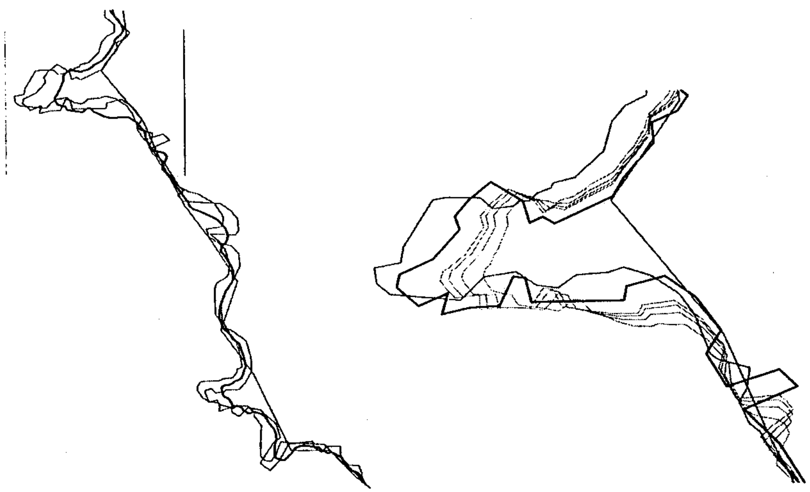

Figure 3: (a) A mean line (darkest line) calculated from a sample of three representations of the same section of shoreline. (b) A mean line progressively weighted towards one of the three lines (darker line).

## UNCERTAINTY MODELING

Earlier, we noted that the conceptual inverse of the ability to estimate a mean for complex geographic features is the ability to generate a sample of such features from knowledge of the mean. This second task is the subject of this section. Arguing again from analogy to the univariate case, the normal distribution is a commonly used basis for simulation, and is observed to achieve a good match with many empirical examples. It can be thought of as a stochastic process which generates samples given suitable parameters, specifically the mean and standard deviation. In two dimensions, our problem is to devise a similar process which when given a mean line and a suitable measure of dispersion, such as the standard deviation discussed in the previous section, or $e_{95}$ as discussed by Goodchild and Hunter (1995), will generate a sample of possible positions of the feature. The value of such a model lies in its ability to clearly depict the possible variations of a particular representation that might exist. These models have widespread practical application, as they allow an analyst to assess the possible outcomes of a spatial process by generating a sample of equally likely inputs and assessing how changes in these inputs propagate through a spatial process to impact any outputs. An example of this modeling approach might involve generating a sample of equally likely digitized representations of the mean to gain a better understanding of how the uncertainty introduced by digitizing affects the calculation of a feature's area.

In this paper, the concern is with modeling the uncertainty in vector geographic data, or more precisely, the uncertainty inherent when merging linear representations of a selected geographic feature from disparate sources. Hunter and Goodchild (1995) have developed an appropriate uncertainty model for this purpose that is capable of generating a sample of equally likely vector representations of reality. Although their model is intended to further our understanding of the effects of positional uncertainty on spatial processing, it can also be used to test the robustness of the mean line procedure discussed in the prior section.

The model is based on generating two spatially autocorrelated random fields that are then combined to create a random vector distortion field for perturbing positions. The distortion field is assumed to have a bivariate normal distribution with equal variances and zero covariance. To postpone the problems associated with grid resolution and other finite representation issues, this method can be conceptualized in continuous space.

An important concept in modeling the uncertainty in spatial data products is the inherent spatial autocorrelation of distortion. If distortion is conceptualized as spatially continuous, then certain conditions are imposed on the continuity of the vector field, to avoid rips and folds in the distorted result. The probability of violating these conditions is minimized by imposing a structure of positive spatial autocorrelation on the random fields.

There are a number of issues that must be handled when moving this model into a discrete spatial data model domain, and these are addressed systematically by Hunter and Goodchild. They include the generation of the distortion field, the spatial sampling interval of the distortion field, preserving topological integrity between distorted features, and calculating positional shifts for data points that do not coincide with cell centers in the distortion grids used to represent a distortion field.

Two significant issues to consider in generating the distortion grids are the magnitude of the random field and its degree of spatial autocorrelation. Ehlschlaeger and Goodchild (1994) have implemented a method for generating random fields with inherent spatial autocorrelation within the GRASS environment called r.random.surface. The generated field corresponds to a normal distribution with a theoretical mean of 0 and a standard deviation of 1. Hunter and Goodchild (1995) recommend generating an distortion field with a standard deviation equivalent to the producer's horizontal distortion estimate for the data set, and, if this information is available, a given random field can be multiplied by a scalar to adjust its standard deviation to correspond to this value. In terms of autocorrelation, r.random.surface takes two parameters: a distance decay exponent and a minimum distance of spatial independence. Both are used to control the form of the autocorrelogram of the simulated random field; the distance decay exponent controls the rate of decrease in spatial autocorrelation with distance, and the minimum distance parameter controls the autocorrelogram's range.

Spatial autocorrelation is essential in generating the perturbation field as the topological integrity of features must be preserved. If a line's neighboring points are perturbed by varying magnitudes, then there is the possibility of introducing a fold, where one point "overtakes" another point resulting in a change in the topological relationship between the two points. In this case, the signed difference in the perturbation's components between neighboring points is important and not the absolute difference. A point to the right of another point can move right by more than the point to the left without causing a topological problem. In general, if u is the x distortion field and dx is the cell size, then $u(x+dx) - u(x) < -dx$ will cause a fold. In terms of u as a continuous function, $\partial u/\partial x$ or $\partial u/\partial y$ must be less than -1 for a fold to occur. Hunter and Goodchild (1995) suggest using bilinear interpolation to determine the value of the vector perturbation field at points other than grid points. Given the geometric properties of the surface between grid points when interpolated in this way ($\partial u/\partial x$ independent of y; $\partial u/\partial y$ independent of x), a test of the difference in perturbation vectors at neighboring grid points is sufficient to determine the presence of folds in the interpolated surface.

Hunter and Goodchild (1995) recommend a distortion grid resolution smaller than the minimum distance between any two points in a polyline or polygon. A possible rule of thumb is a grid resolution of 0.5mm at the scale of the map from which the data originated. If this information is not available, the grid resolution should be set to some very small value.

Figure 4 shows the results of simulation of from a mean shoreline and standard deviation. The mean shoreline was generated using the method described for three sample shorelines. Standard deviations for the x and y directions were obtained by summing the squared distances between each point in each sample shoreline and the corresponding point in the mean line. Distortion fields (x and y components) were generated at a resolution of 3m, which is less than the minimum distance between points in the mean line. The nine simulations in Figure 4 are the result of varying the distance decay exponent from 1 to 0.5 to 0.1, and varying the minimum distance to independence from 200m to 400m to 600m. The x and y components of the distortion field were multiplied by the magnitude of the standard deviation for x and y (approximately 30m and 90m, respectively). The distortion fields were evaluated between grid points using bilinear interpolation.

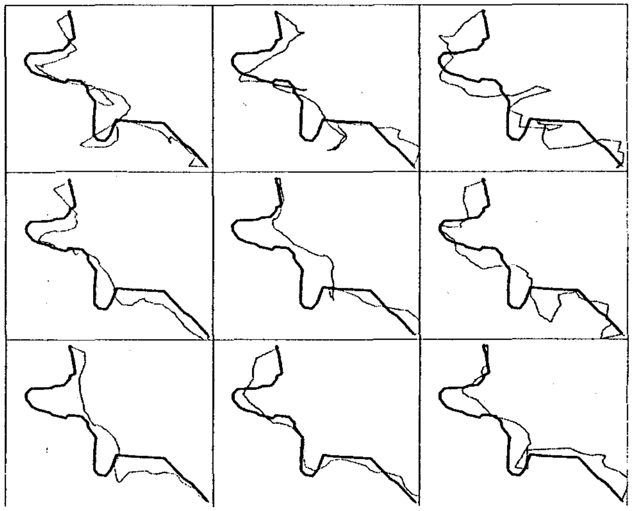

Figure 4: Sample shorelines (grey lines) generated by distorting a mean shoreline with an random, autocorrelated vector field. The darker line is the actual shoreline from one of the sources and is shown for comparison. The distance exponents (columns, from left) are 1.0, 0.5, and 0.1; the minimum distance to independence (rows, from top) are 200m, 400m, and 600m. The cell size is 3m and the area shown is 250 by 333 cells.

Figure 4 clearly demonstrates the importance of the spatial autocorrelation parameters in controlling the incidence of loops and other topological problems resulting from distortion. In general, more linear distance decay (values closer to 1.0) and longer minimum distances of independence are less likely to create problems. Because the dispersion parameter was obtained by comparing representations at three different levels of generalization (Figure 3) the distortions are larger than they would likely be in modeling other sources of uncertainty, such as digitizing error. Note also the importance of matching, made obvious here by the use of one source line rather than the mean line in the comparison.

## CONCLUSION

We have suggested that the concepts of mean, standard deviation, estimation from a sample, and simulation from a distribution can be usefully extended from the usual scalar context to two-dimensional objects in GIS. Applications range from data fusion to sliver removal and enhancement of line detail. We have explored several methods, noting some of the problems associated with each. While it is robust, the principal problem with the method based on parametric line representations is its sensitivity to the choice of matching points, and methods of automatic matching need to be explored in greater detail. Of the other methods, averaging based on binary field representations of half planes is robust and attractive, but requires that the user accept the possibility of holes and islands in the result. The "medial axis transform", as interpreted above, is also attractive, but will require extension to more than two lines. We have also noted that the combination of estimation and simulation can be used as the basis for a rigorous test of these methods, and will be continuing this line of research in the future.

## ACKNOWLEDGMENT

The National Center for Geographic Information and Analysis is supported by the National Science Foundation under Cooperative Agreement SBR 88-10917.

## REFERENCES

Dutton, G. (1981) Fractal enhancement of cartographic line detail. *The American Cartographer* 8(1): 23-40.

Dutton, G. (1992) Handling positional uncertainty in spatial databases. *Proceedings, 5th International Symposium on Spatial Data Handling*, pp. 460-469.

Edwards, G. (1994a) Characterising spatial uncertainty and variability in forestry data bases. *Proceedings, International Symposium on Spatial Accuracy of Natural Resource Data Bases*, pp. 88-97.

Edwards, G. (1994b) Characterising and maintaining polygons with fuzzy boundaries in geographic information systems. *Proceedings 6th International Symposium on Spatial Data Handling*, pp. 223-238.

Ehlschlaeger, C., and M.F. Goodchild (1994) Uncertainty in spatial data: defining, visualizing, and managing data errors. *Proceedings, GIS/LIS, Phoenix, Arizona*, pp. 246-253.

Englund, E. (1993) Spatial simulation: environmental applications. In M.F. Goodchild, B.O. Parks, and L.T. Steyaert, editors, *Environmental Modeling with GIS*. Oxford University Press, New York, pp. 432-437.

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

Goodchild, M.F.,and G.J. Hunter (1995) A simple positional accuracy measure for linear features. Submitted to *International Journal of Geographical Information Systems*.

Hunter, G.J., and M.F. Goodchild (1995) A new model for handling vector data uncertainty in geographic information systems. *Proceedings, URISA, San Antonio, Texas*, pp. 410-419.

Lee, D.T. (1982) Medial axis transformation of a planar shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(4): 363-369.

Monmonier, M.S. (1982) *Computer-Assisted Cartography: Principles and Prospects*. Prentice-Hall, Englewood Cliffs, NJ, p. 158.

Pavlidis, T. (1982) *Algorithms for Graphics and Image Processing*. Computer Science Press, Rockville, MD.

Perkal, J. (1966) On the length of empirical curves. Discussion Paper 10, Inter-University Community of Mathematical Geographers, Ann Arbor, MI.

Shi, W., and K. Tempfli (1994) Modelling positional uncertainty of line features in GIS. *Proceedings ASPRS/ACSM '94, Reno.*