The Mean and Variance of Area Estimates Computed in an Arc-Node Geographic Information System

Stephen P. Prisley, Timothy G. Gregoire,* and *James L. Smith*
Department of Forestry, 319 Cheatham Hall, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061

ABSTRACT: The extensive use of GIS for deriving summary values from map analyses has created a need for an expression of the uncertainty of area estimates. Based on a few assumptions regarding the locational accuracy of point coordinates in a vector GIS, it is possible to derive the mean and variance of area estimates, and the covariance of area of adjacent polygons. The expressions obtained can prove useful in understanding and modeling the impacts of spatial errors in GIS applications.

INTRODUCTION

VIRTUALLY ALL SPATIALLY-REFERENCED DATABASES contain locational errors. Several authors have devoted much time and effort to categorizing and describing these spatial errors (Mead, 1982; Walsh *et al.*, 1987; Burrough, 1986). However, an assessment of how these errors affect decisions made using Geographic Information Systems (GIS) is lacking. These decisions are often based on values computed within the GIS, such as area and distance, rather than on the map products themselves. It is important to recognize that these derived values are estimates which contain error; i.e., they have an inherent accuracy and precision which is defined by the properties of the data and the method of calculation. Although most sample estimates, such as public opinion surveys or forest inventories, are considered incomplete without some measure of accuracy and precision, area estimates are used without the same degree of reservation. GIS have exacerbated these problems to some extent, but they have also created an environment wherein spatial values, in particular area estimates, can be treated statistically.

The concentration on summary information from GIS has focused attention on the errors in such measurements. One reason for the keen interest in error analysis is the uncertainty about map overlay products derived from different input sources. In an early treatment of overlay error, McAlpine and Cook (1971) noted the problem of map overlays which resulted in very large numbers of very small polygons which bore little or no agreement with initial map descriptions. MacDougall (1975) reported a gloomy analysis of map overlay accuracy, with which Bailey (1988) concurred. Chrisman (1987) argued for a more positive outlook: "While map error should not be ignored in map overlay, the estimates of MacDougall should be replaced by empirically derived test results. Combining information from diverse sources can actually strengthen the value of the information, not degrade it."

A model of map errors which can be used to study overlay uncertainty would be a useful tool in evaluating such concerns. Such a model could have numerous other uses, such as examining the relationships between map complexity, source map scale, polygon shapes, and line generalization. An error model could also be applied to determine the effects of various map accuracy standards on the uncertainty associated with map measures of area, length, and distance.

Several authors have examined such models. Bondesson (1986) estimated the variance of areas obtained from traverses of polygons, based upon assumptions about errors in bearings and distance measurements. A recent paper by Chrisman and Yandell (1988) examined the mean value and variance of area under an assumption of independently and identically distributed coordinate errors. Our research focuses on the development of an expression for the mean and variance of errors in polygon area under less restrictive, fairly weak assumptions about point coordinate errors.

MODEL DEVELOPMENT

FUNDAMENTAL ASSUMPTIONS

The data model used in this study was the common arc-node data structure for vector data. The fundamental feature stored in an arc-node data structure is the location of a point represented as a cartesian X,Y coordinate. All other features in a map are constructed by linking together point locations. A series of points are connected in sequence to form an arc. Arcs are then connected in sequence to form either linear or polygonal features. This data structure has been described in detail by Peuker and Chrisman (1975). Thus, it is logical that an error model for vector databases begins with the variability of point locations. We follow the notation of Chrisman and Yandell (1988), with extensions and exceptions as noted. First, we express the recorded coordinate of a point by X and Y, indexed by the point number. For example, the recorded location of point i in a polygon is given by

$$(X_i, Y_i).$$

The recorded coordinate consists of the true coordinate and an error term

$$X_i = x_i + \epsilon_i \qquad Y_i = y_i + \eta_i \tag{1}$$

where
x_i = x-coordinate of point i,
y_i = y-coordinate of point i,
ϵ_i = error in location of x-coordinate of point i, and
η_i = error in location of y-coordinate of point i.

*Presently with Westvaco Corporation Timberlands Division, P.O. Box 458, Wickliffe, KY 42087.

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 55, No. 11, November 1989, pp. 1601–1612.

0099-1112/89/5511–1601$02.25/0

As Chrisman and Yandell (1988) suggested, we consider the recorded point location to be an unbiased estimator of the true point location[2]:

$$E(X_i) = x_i \qquad E(Y_i) = y_i$$

Next, we assume that the errors have equal variance in both directions: i.e.,

$$\mathrm{Var}(\epsilon_i) = \mathrm{Var}(\eta_i) = \sigma_i^2.$$

In addition, we assume that the X and Y errors at a given point are uncorrelated: i.e.,

$$E(\epsilon_i \eta_i) = 0.$$

These assumptions are arguable but, in the absence of convincing evidence to the contrary, we feel that they provide a reasonable foundation for initial modeling.

Finally, it is reasonable to assume that the errors at points are correlated. Chrisman and Yandell (1988) examined the case where adjacent points are correlated, and the correlation is negligibly small. Neumyvakin and Panfilovich (1982) accounted for coordinate error correlation in their estimates of plot area variance through the use of a dispersion matrix. Our error model builds upon the empirical investigations of Keefer *et al.* (1988) into the covariance structure of digitizing errors. We will assume that ϵ_i is correlated with ϵ_k, and that η_i is correlated with η_k, for $k = i \pm 1$. Let ρ_i indicate the correlation between errors at points i and $i+1$: i.e.,

$$\rho_i = \frac{\mathrm{Cov}(\epsilon_i, \epsilon_{i+1})}{\sigma_i \sigma_{i+1}} = \frac{\mathrm{Cov}(\eta_i, \eta_{i+1})}{\sigma_i \sigma_{i+1}}.$$

Thus,

$$\mathrm{Cov}(\epsilon_i, \epsilon_{i+1}) = E(\epsilon_i \epsilon_{i+1}) - E(\epsilon_i) * E(\epsilon_{i+1}) = E(\epsilon_i \epsilon_{i+1}) = \rho_i \sigma_i \sigma_{i+1}$$

and

$$\mathrm{Cov}(\eta_i, \eta_{i+1}) = E(\eta_i \eta_{i+1}) - E(\eta_i) * E(\eta_{i+1}) = E(\eta_i \eta_{i+1}) = \rho_i \sigma_i \sigma_{i+1}.$$

We will further assume that the correlation between errors at adjacent points is positive. In summary, for all i, k, as noted:

$$\begin{aligned} E(\epsilon_i) &= 0 \\ E(\epsilon_i \epsilon_{i+1}) &= \rho_i \sigma_i \sigma_{i+1} \\ E(\epsilon_i \epsilon_i) &= \sigma_i^2 \\ E(\epsilon_i \epsilon_k) &= 0 \text{ for } |k-i| > 1 \\ E(e_i \eta_k) &= 0 \quad \forall\ i, k \end{aligned} \qquad \begin{aligned} E(\eta_i) &= 0 \\ E(\eta_i \eta_{i+1}) &= \rho_i \sigma_i \sigma_{i+1} \\ E(\eta_i \eta_i) &= \sigma_i^2 \\ E(\eta_i \eta_k) &= 0 \text{ for } |k-i| > 1 \end{aligned}$$

This covariance structure may be simplistic still. Yet as Chrisman and Yandell (1988) imply, a highly refined model of error covariance may not be required. In addition to the properties described above (mean error, variance, and correlation), a normal distribution for the coordinate errors might reasonably be assumed. The normal distribution is often used to model errors which are the result of a number of independent steps, each resulting in an error. However, for the derivations developed herein, the assumption of normally distributed errors is not necessary. We will later discuss some useful results that are obtained if normality is assumed.

Arcs as Collections of Points

The next structure encountered in progressing from points to polygons is an arc which delineates a boundary between two homogeneous regions. Inasmuch as an arc is defined by points, and it is likely that all points in an arc have been similarly processed, it is therefore reasonable to assume that all points in an arc will exhibit similar variability. Different arcs may have different variances, and different degrees of correlation between point errors (*cf.* Goodchild and Dubuc, 1987). Allowance for different variances among arcs is important for the verisimilitude of a model for map overlay errors. Consider a map which is derived from overlays of a soils map, a vegetative cover map, and a land ownership boundary map. The individual input maps are apt to have quite different error structures. In the overlay map, each arc can be traced back to one or more arcs in the source maps, which implies that the error structures of the source maps can be maintained as attributes of arcs in the overlay map.

A node by definition is the endpoint of an arc, and may belong to two or more connected arcs. Therefore, the variability of a node is difficult to discern. For the purposes of modeling, one possibility is to assign the minimum of the variances of the incident arcs. This may be reasonable if an effort has been made in digitizing, editing, and overlaying to maintain the integrity of the most accurate arcs. For example, some digitizing and editing software allows a user to "snap" an arc being digitized to an existing (and presumably more accurately located) arc. In such a case, the most logical choice for the variability of the node is the variability of the existing, more accurately located, arc. In this derivation, the most general case is considered, in which all points in an arc may have different variances. In implementation of the variance expression, however, points within an arc may be assumed to have the same variance.

Arc-Sectors as Collections of Triangles

An arc by definition is shared by two polygons. Thus, each arc in an arc-node data set may contribute to the area and variability of two polygons. We will define an "arc-sector" as the two-dimensional figure composed of an arc and two line segments connecting the two nodes to the polygon centroid (Figure 1). Thus, arcs may have two arc-sectors associated with them: one for each polygon which uses the arc. External boundary arcs obviously are associated with only one arc-sector.

An arc-sector comprises individual triangles. Each triangle is composed of a single line segment of the arc (defined by a pair of adjacent points), the endpoints of which are connected to the polygon centroid. If there are $m+1$ points (including nodes) in an arc, m triangles will be defined. Each triangle may contribute positively or negatively to polygon area (Figure 2). The areas of the triangles are easily calculated based upon the coordinates of the three points which define them. It is noted that areas of

[2]$E(\bullet)$ denotes statistical expectation.

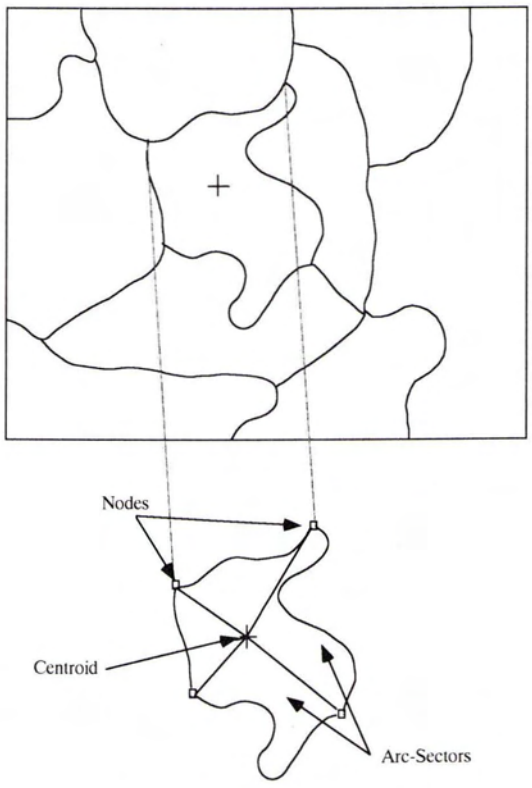

FIG. 1. Example of an "arc-sector." An arc-sector is the two-dimensional figure formed by connecting the nodes of an arc to the polygon centroid.

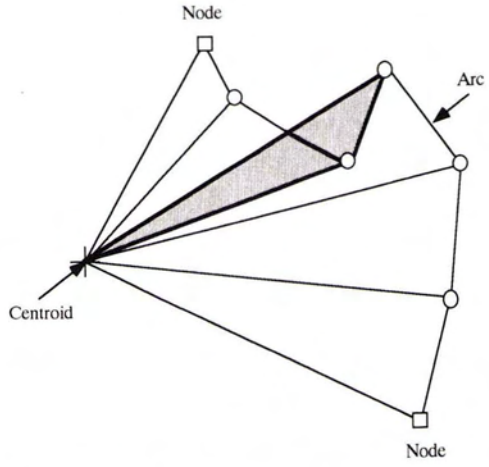

FIG. 2. Diagram of triangles within an arc-sector. Note that the area of the shaded triangle is deducted from arc-sector area, while all other triangles add to total area.

adjacent triangles will be correlated, as adjacent triangles share two points in common: the polygon centroid and a point on the arc.

POLYGONS AS COLLECTIONS OF ARC-SECTORS

The next feature to be considered is the polygon. A polygon is an accumulation of adjacent arcs-sectors. Each arc-sector may contribute positively or negatively to polygon area (like the triangles in Figure 2). As in the case with adjacent triangles, adjacent arc-sectors are correlated because they share points. The definition of arc-sectors is not necessary for the computation of the area variance for a single polygon. It will, however, be used for assessing total map error.

THE MAP AS A COLLECTION OF POLYGONS

The final consideration in the aggregation of polygons to form a choropleth map is the covariation between polygons. As the area of one polygon increases, the area of at least one of its neighbors must decrease. There is an obvious correlation, then,

between errors in area of polygons within a map. The final expression of variability of area in a choroplethic map must include both variance of, and covariance between, polygons. An example will show why covariance must be considered.

Suppose an overall estimate of some variable is desired from a map. In many cases, per-unit-area values associated with individual polygons are multiplied by polygon areas and summed over a map to obtain an overall estimate. Examples may include

- Soil erosion rates expressed in tons per hectare per year for terrain units in a map, multiplied by hectares in each unit and summed over a watershed to obtain total annual soil erosion losses;
- Timber volume per acre multiplied by area of various timber stand polygons, and summed over a given tract to obtain total tract volume;
- Land values in dollars per acre assigned to different parcels and aggregated over a tax district to obtain total assessed value.

The value being sought is a linear combination of per-unit-area values and polygon areas. In cases such as these, if the per-unit-area values are assumed to be fixed, the variance of the overall estimate is obtained by

$$\mathrm{Var}(Z) = \sum_{polygons} \left(z_i^2 * \mathrm{Var}(a_i) \right) + 2 \sum_{i<j} \left(z_i z_j * \mathrm{Cov}(a_i, a_j) \right) \tag{2}$$

where

Z = overall estimate,
z_i = per-unit-area estimate for polygon i, and
a_i = area of polygon i.

Thus, to evaluate Equation 2, the covariance of area for adjacent polygons is necessary. Ignoring a positive covariance will result in a serious underestimate of total variance, and ignoring a negative covariance will result in overestimation of total variance. The proposed model uses covariance between areas of triangles on opposite sides of arcs to obtain polygon covariances.

STATISTICAL DERIVATION

Preliminaries

Because the coordinates of a polygon boundary are random variables, polygon area is a function of random variables. By taking expected values and variances of sums and products of random variables, the variance of area is obtained. First, the variance of area of a triangle will be determined. Second, the covariance of area between adjacent triangles will be derived. By summing triangle variances and covariances, the polygon variance is obtained. To obtain covariance between polygons, the covariances between triangles in adjacent polygons will first be determined, and then summed along the arc which separates two polygons.

As we consider a single polygon first, we can subtract the coordinates of the polygon centroid from all points comprising the polygon. This effectively "centers" the polygon at the coordinate origin, but does not change the polygon area. In addition to simplifying subsequent algebra, it is helpful in a computer implementation by allowing greater precision in computation. We represent centered coordinates as $\tilde{X}$, $\tilde{Y}$: then

$$\tilde{X}_i = X_i - X_c \qquad \tilde{Y}_i = Y_i - Y_c$$

and

$$\tilde{x}_i = x_i - X_c \qquad \tilde{y}_i = y_i - Y_c$$

where (X_c, Y_c) is the centroid coordinate. Using centered coordinates in this manner introduces an additional dependency: the variance estimate depends upon the location of the centroid. This problem will be addressed in a later section.

Individual Polygon Area Variance

The first component of a polygon that we consider is the triangle formed by two points in an arc and the centroid. The area of the triangle formed by points X_i, Y_i and X_{i+1}, Y_{i+1}, and the centroid (X_c, Y_c) is given by

$$\begin{aligned} A_i &= \frac{1}{2} * \left((X_i - X_c)(Y_{i+1} - Y_c) - (X_{i+1} - X_c)(Y_i - Y_c) \right) \\ &= \frac{1}{2} * (\tilde{X}_i \tilde{Y}_{i+1} - \tilde{X}_{i+1} \tilde{Y}_i). \end{aligned}$$

Using Equation 1, this can be written as

$$\begin{aligned} A_i &= \frac{1}{2} * \left((\tilde{x}_i + \epsilon_i)(\tilde{y}_{i+1} + \eta_{i+1}) - (\tilde{x}_{i+1} + \epsilon_{i+1})(\tilde{y}_i + \eta_i) \right) \\ &= \frac{1}{2} * \left((\tilde{x}_i \tilde{y}_{i+1} - \tilde{x}_{i+1} \tilde{y}_i) + (\tilde{x}_i \eta_{i+1} + \tilde{y}_{i+1} \epsilon_i - \tilde{x}_{i+1} \eta_i - \tilde{y}_i \epsilon_{i+1}) + (\epsilon_i \eta_{i+1} - \epsilon_{i+1} \eta_i) \right). \end{aligned} \tag{3}$$

Now, note that the nominal area of the triangle (assuming no errors) is equal to the first two terms in Equation 3: i.e.,

$$a_i = \frac{1}{2} * (\tilde{x}_i \tilde{y}_{i+1} - \tilde{x}_{i+1} \tilde{y}_i).$$

We define the remaining six terms in Equation 3 as follows:

$$t_1 = \frac{1}{2}(\tilde{x}_i\eta_{i+1}) \qquad t_2 = \frac{1}{2}(\tilde{y}_{i+1}\epsilon_i) \qquad t_3 = -\frac{1}{2}(\tilde{x}_{i+1}\eta_i)$$

$$t_4 = -\frac{1}{2}(\tilde{y}_i\epsilon_{i+1}) \qquad t_5 = \frac{1}{2}(\epsilon_i\eta_{i+1}) \qquad t_6 = -\frac{1}{2}(\epsilon_{i+1}\eta_i)$$

And we note:

$$E(t_1) = \frac{1}{2}\tilde{x}_i E(\eta_{i+1}) = 0 \qquad E(t_2) = \frac{1}{2}\tilde{y}_{i+1}E(\epsilon_i) = 0$$

$$E(t_3) = -\frac{1}{2}\tilde{x}_{i+1}E(\eta_i) = 0 \qquad E(t_4) = -\frac{1}{2}\tilde{y}_i E(\epsilon_{i+1}) = 0$$

$$E(t_5) = \frac{1}{2}E(\epsilon_i\eta_{i+1}) = 0 \qquad E(t_6) = -\frac{1}{2}E(\epsilon_{i+1}\eta_i) = 0$$

Thus, taking the expectation of Equation 3 yields

$$E(A_i) = a_i + \left(E(t_1) + E(t_2) + E(t_3) + E(t_4) + E(t_5) + E(t_6)\right) = a_i.$$

Evidently, the mean area of a triangle with coordinate errors coincides with its nominal area. If we are willing to assume that coordinate errors are zero, on average, then the estimated area equals the true area, on average. However, an individual area estimate will deviate from the true area, and so it is important to assess the precision of area estimates.

The variance of A_i is

$$\mathrm{Var}(A_i) = \sum_{i=1}^{6} \mathrm{Var}(t_i) + 2 * \sum_{i<j} \mathrm{Cov}(t_i,t_j) \tag{4}$$

where

$$\mathrm{Var}(t_1) = \frac{1}{4}\tilde{x}_i^2\sigma_{i+1}^2 \quad \mathrm{Var}(t_2) = \frac{1}{4}\tilde{y}_{y+1}^2\sigma_i^2 \quad \mathrm{Var}(t_3) = \frac{1}{4}\tilde{x}_{i+1}^2\sigma_i^2$$
$$\mathrm{Var}(t_4) = \frac{1}{4}\tilde{y}_i^2\sigma_{i+1}^2 \quad \mathrm{Var}(t_5) = \frac{1}{4}\sigma_i^2\sigma_{i+1}^2 \quad \mathrm{Var}(t_6) = \frac{1}{4}\sigma_{i+1}^2\sigma_i^2 \tag{5}$$

and

$$\mathrm{Cov}(t_1,t_3) = -\tilde{x}_i\tilde{x}_{i+1}\sigma_i\sigma_{i+1}\rho_i$$
$$\mathrm{Cov}(t_2,t_4) = -\tilde{y}_i\tilde{y}_{i+1}E(\epsilon_i\epsilon_{i+1}) = -\tilde{y}_i\tilde{y}_{i+1}\sigma_i\sigma_{i+1}\rho_i \tag{6}$$
$$\mathrm{Cov}(t_5,t_6) = -E(\epsilon_i\epsilon_{i+1}\eta_i\eta_{i+1}) = -\sigma_i^2\sigma_{i+1}^2\rho_i^2$$

and all other covariances are zero. Substituting Equations 5 and 6 into Equation 4 yields:

$$\mathrm{Var}(A_i) = \frac{1}{4} * \left(r_i^2\sigma_{i+1}^2 + r_{i+1}^2\sigma_i^2 - 2(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1})\sigma_i\sigma_{i+1}\rho_i + 2(1-\rho_i^2)\sigma_i^2\sigma_{i+1}^2\right) \tag{7}$$

where $r_i^2 = \tilde{x}_i^2 + \tilde{y}_i^2$ and $r_{i+1}^2 = \tilde{x}_{i+1}^2 + \tilde{y}_{i+1}^2$.

The area of a polygon is the sum of n individual triangular areas: i.e.,

$$A = \frac{1}{2} * \sum_{i=1}^{n} (\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i)$$

where the sum is "circular," i.e., $\tilde{X}_{n+1} = \tilde{X}_1$ and $\tilde{Y}_{n+1} = \tilde{Y}_1$. This expression yields a positive area when the coordinates are indexed in a counter-clockwise direction. Individual triangles may be positive or negative in area, but the sum should be positive. The mean polygon area should be

$$E(A) = E\left(\frac{1}{2} * \sum_{i=1}^{n} (\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i)\right) = \sum_{i=1}^{n} a_i,$$

which is the nominal polygon area. Because errors in area of adjacent triangles *are not* independent, covariance terms will be required for adjacent triangles in order to derive the variance of A. The variance of A can be expressed as the variance of a sum: i.e.,

$$\mathrm{Var}(A) = \mathrm{Var}\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{Var}(A_i) + 2 * \sum_{i=2}^{n+1} \mathrm{Cov}(A_{i-1},A_i) \tag{8}$$

where $A_{n+1} = A_1$.

The variance of A_i is given by Equation 7. To derive $\mathrm{Cov}(A_{i-1},A_i)$, we begin by noting that

$$\mathrm{Cov}(A_{i-1},A_i) = E(A_{i-1}A_i) - E(A_{i-1}) * E(A_i) = E(A_{i-1}A_i) - a_{i-1}a_i$$

and

$$A_{i-1}A_i = \frac{1}{4} * \left((\tilde{X}_{i-1}\tilde{Y}_i - \tilde{X}_i\tilde{Y}_{i-1}) * (\tilde{X}_i\tilde{Y}_{i+1} - \tilde{X}_{i+1}\tilde{Y}_i) \right)$$
$$= \frac{1}{4} * \left(\tilde{X}_{i-1}\tilde{X}_i\tilde{Y}_i\tilde{Y}_{i+1} - \tilde{X}_{i-1}\tilde{X}_{i+1}\tilde{Y}_i^2 - \tilde{X}_i^2\tilde{Y}_{i-1}\tilde{Y}_{i+1} + \tilde{X}_i\tilde{X}_{i+1}\tilde{Y}_{i-1}\tilde{Y}_i \right).$$

It can be shown that

$$\mathrm{Cov}(A_{i-1},A_i) = \frac{1}{4} * \left(\tilde{x}_{i-1}\tilde{x}_i\sigma_i\sigma_{i+1}\rho_i + \tilde{y}_i\tilde{y}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \sigma_{i-1}\sigma_i{}^2\sigma_{i+1}\rho_{i-1}\rho_i - x_{i-1}\,\tilde{x}_{i+1}\sigma_i^2 \right.$$
$$\left. - \tilde{y}_{i-1}\tilde{y}_{i+1}\sigma_i^2 + \tilde{x}_i\tilde{x}_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + \tilde{y}_{i-1}\tilde{y}_i\sigma_i\sigma_{i+1}\rho_i + \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i \right)$$

$$= \frac{1}{4} * \left((\tilde{x}_{i-1}\tilde{x}_i + \tilde{y}_{i-1}\tilde{y}_i)\sigma_i\sigma_{i+1}\rho_i + (\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_i\tilde{x}_{i+1})\sigma_{i-1}\sigma_i\rho_{i-1} + 2\sigma_{i-1}\sigma_i{}^2\sigma_{i+1}\rho_{i-1}\rho_i - (\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1})\sigma_i^2 \right). \quad (9)$$

Substituting Equations 7 and 9 into Equation 8 obtains

$$\mathrm{Var}(A) = \frac{1}{4} * \sum_{i=1}^{n} \left(r_i^2\sigma_{i+1}^2 + r_{i+1}^2\sigma_i^2 - 2(\tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1})\sigma_i\sigma_{i+1}\rho_i + 2(1-\rho_i^2)\sigma_i^2\sigma_{i+1}^2 \right) + \frac{1}{2} * \sum_{i=2}^{n+1}$$
$$\left((\tilde{x}_{i-1}\tilde{x}_i + \tilde{y}_{i-1}\tilde{y}_i)\sigma_i\sigma_{i+1}\rho_i + (\tilde{y}_i\tilde{y}_{i+1} + \tilde{x}_i\tilde{x}_{i+1})\sigma_{i-1}\sigma_i\rho_{i-1} + 2\sigma_{i-1}\sigma_i{}^2\sigma_{i+1}\rho_{i-1}\rho_i\ \rho - (\tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1})\sigma_i^2 \right). \quad (10)$$

To simplify this expression, define

$$w_i = \tilde{x}_i\tilde{x}_{i+1} + \tilde{y}_i\tilde{y}_{i+1} \quad (11)$$
$$z_i = \tilde{x}_{i-1}\tilde{x}_{i+1} + \tilde{y}_{i-1}\tilde{y}_{i+1}$$
$$s_i = \sigma_i\sigma_{i+1}\rho_i$$

Substituting into Equation 10 obtains

$$\mathrm{Var}(A) = \frac{1}{4} * \sum_{i=1}^{n} \left(r_i^2\sigma_{i+1}^2 + r_{i+1}^2\sigma_i^2 - 2w_is_i + 2\sigma_i^2\sigma_{i+1}^2 - 2s_i^2 + 2w_{i-1}s_i + 2w_is_{i-1} + 4s_{i-1}s_i - 2z_i\sigma_i^2 \right).$$

Using the fact that this sum is "circular," we can rewrite this as

$$\mathrm{Var}(A) = \frac{1}{4} * \sum_{i=1}^{n} \left(r_i^2(\sigma_{i-1}^2 + \sigma_{i+1}^2) + 2w_i(s_{i-1} - s_i + s_{i+1}) + 2\sigma_i^2\sigma_{i+1}^2 - 2s_i^2 + 4s_{i-1}s_i - 2z_i\sigma_i^2 \right). \quad (12)$$

Note that this expression for variance of area of a polygon, in terms of the coordinates, the point variances (σ_i's), and the correlations between errors (ρ_i's), does not explicitly include representation of the arcs which comprise the polygon. Instead, individual points which made the polygon boundary are used. The identification of arcs (or arc-sectors) is not necessary; however, in a computer implementation of this formulation, some efficiencies will be noted if arcs are considered.

Definition of the Centroid

We have stated that the location of the centroid used to center the coordinates prior to variance calculations affects the value of the variance obtained. Therefore, it is reasonable to select the centroid location which *minimizes* variance, and, given the foregoing assumptions, minimizes mean square error. To determine the location of the "minimum-variance centroid" (MVC), the formula for polygon variance is written as a function of the variances of the coordinates (σ's), the correlations associated with the arcs (ρ's), the point coordinates (X_i's and Y_i's), and the centroid coordinates (X_c and Y_c) (see Appendix). Differential calculus then yields the centroid coordinates X_c, Y_c which minimize variance: i.e.,

$$X_c = \frac{\sum_{i=1}^{n} X_i(\sigma_{i-2}\sigma_{i-1}\rho_{i-2} + \sigma_{i+1}\sigma_{i+2}\rho_{i+1})}{\sum_{i=1}^{n}(2\,\sigma_i\sigma_{i+1}\rho_i)}$$

$$Y_c = \frac{\sum_{i=1}^{n} Y_i(\sigma_{i-2}\sigma_{i-1}\rho_{i-2} + \sigma_{i+1}\sigma_{i+2}\rho_{i+1})}{\sum_{i=1}^{n}(2\sigma_i\sigma_{i+1}\rho_i)}.$$

The minimum-variance centroid coordinate for a polygon is a weighted average of the polygon coordinates, in which the weights are the products of σ's and ρ's associated with adjacent coordinates.

Covariance between Polygons

Up to this point, we have been concerned with characteristics of individual polygons. However, often our interest is in an aggregation of polygons. In other words, what are the mean and variance of a linear combination of area estimates? It follows simply that the sum of unbiased estimates is itself unbiased, which makes summarized area estimates unbiased under the assumptions used here. However, variance of a total is more complex. Equation 2 indicates that total map variance is not only dependent on the variance of individual polygons, but on their covariance as well. Thus, some expression for the covariance between adjacent polygons must be developed.

First, consider polygon A as a polygon with centroid (X_a, Y_a). It shares an arc with polygon B, whose centroid is at (X_b, Y_b) (Figure 3). We consider the triangles involved in a sequence of four points on the arc: (X_{i-1}, Y_{i-1}), (X_i, Y_i), (X_{i+1}, Y_{i+1}), (X_{i+2}, Y_{i+2}). Because the sequence of indexing depends on the direction (relative to a centroid), assume the direction of indexing is that which will yield positive areas for polygon A (note direction of arrows in Figure 3). Thus, for polygon A, triangle i, the area is

$$A_i = \frac{1}{2} * \left((X_i - X_a)(Y_{i+1} - Y_a) - (X_{i+1} - X_a)(Y_i - Y_a) \right). \tag{13}$$

This implies that, for polygon B, the direction is reversed; i.e., for polygon B, triangle i, the area is

$$B_i = \frac{1}{2} * \left((X_{i+1} - X_b)(Y_i - Y_b) - (X_i - X_b)(Y_{i+1} - Y_b) \right).$$

There are three cases to consider: these involve the covariance between triangle i in A and the three triangles in B with which there is a dependency: triangles $i+1$, i, and $i-1$. Thus, we will need expressions for

$$\mathrm{Cov}(A_i, B_{i-1}), \qquad \mathrm{Cov}(A_i, B_i), \qquad \text{and } \mathrm{Cov}(A_i, B_{i+1}).$$

We have derived an expression for each of these three cases, but have shown only the first here. The remaining two follow along very similar lines[3].

Derivation of $\mathrm{Cov}(A_i, B_{i-1})$

By definition, $\mathrm{Cov}(A_i, B_{i-1}) = E(A_i B_{i-1}) - E(A_i)E(B_{i-1})$.

From Equation 3 *et seq*, we obtain $A_i = a_i + c_1$ and $B_{i-1} = b_{i-1} + c_2$

where a_i, b_{i-1} are the true polygon areas, and where

$$c_1 = \frac{1}{2} * \left(x_i \eta_{i+1} + y_{i+1} \epsilon_i + \epsilon_i \eta_{i+1} - Y_a \epsilon_i - X_a \eta_{i+1} - x_{i+1} \eta_i - y_i \epsilon_{i+1} - \epsilon_{i+1} \eta_i + Y_a \epsilon_{i+1} + X_a \eta_i \right)$$

[3]Derivations are available from the authors on request.

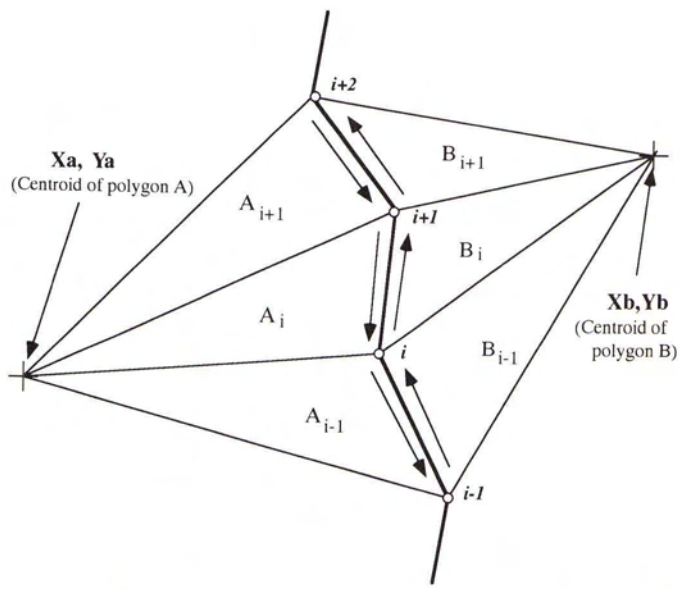

FIG. 3. Diagram of triangles in adjacent arc-sectors. Points on the arc are indexed in the direction which makes the area of polygon A positive. Triangles B_{i-1}, B_i, and B_{i+1} will have non-zero covariance with triangle A_i.

and

$$c_2 = \frac{1}{2} * \Big(x_i\eta_{i-1} + y_{i-1}\,\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1} - \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i \Big).$$

As shown earlier, $E(c_1) = E(c_2) = 0$ and, consequently,

$$E(A_i) = a_i$$
$$E(B_{i-1}) = b_{i-1}$$
$$E(A_iB_{i-1}) = a_ib_{i-1} + E(c_1c_2)$$

Thus,

$$\begin{aligned}
\mathrm{Cov}(A_i,B_{i-1}) &= E(c_1c_2)\\
&= \frac{1}{4} * E\Big((x_i\eta_{i+1} + y_{i+1}\epsilon_i + \epsilon_i\eta_{i+1} - Y_a\epsilon_i - X_a\eta_{i+1} - x_{i+1}\eta_i - y_i\epsilon_{i+1} - \epsilon_{i+1}\eta_i + Y_a\epsilon_{i+1} + X_a\eta_i)\\
&\quad * (x_i\eta_{i-1} + y_{i-1}\epsilon_i + \epsilon_i\eta_{i-1} - Y_b\epsilon_i - X_b\eta_{i-1} - x_{i-1}\eta_i - y_i\epsilon_{i-1} - \epsilon_{i-1}\eta_i + Y_b\epsilon_{i-1} + X_b\eta_i)\Big)\\
&= \frac{1}{4} * \Big(-x_{i-1}x_iE(\eta_i\eta_{i+}) + X_bx_iE(\eta_i\eta_{i+1}) + y_{i-1}y_{i+1}E(\epsilon_i^2) - Y_by_{i+1}E(\epsilon_i^2) - y_iy_{i+1}E(\epsilon_{i-1}\epsilon_i) + Y_by_{i+1}E(\epsilon_{i-1}\epsilon_i)\\
&\quad - E(\epsilon_{i-1}\epsilon_i\eta_i\eta_{i+1}) - Y_ay_{i-1}E(\epsilon_i^2) + Y_aY_bE(\epsilon_i^2) + Y_ay_iE(\epsilon_i\epsilon_{i-1}) - Y_aY_bE(\epsilon_{i-1}\epsilon_i) + X_ax_{i-1}E(\eta_i\eta_{i+1})\\
&\quad - X_aX_bE(\eta_i\eta_{i+1}) - x_ix_{i+1}E(\eta_{i-1}\eta_i) + X_bx_{i+1}E(\eta_{i-1}\eta_i) + x_{i-1}x_{i+1}E(\eta_i^2) - X_bx_{i+1}E(\eta_i^2)\\
&\quad - y_{i-1}y_iE(\epsilon_i\epsilon_{i+1}) + Y_by_iE(\epsilon_i\epsilon_{i+1}) - E(\epsilon_i\epsilon_{i+1}\eta_{i-1}\eta_i) + Y_ay_{i-1}E(\epsilon_i\epsilon_{i+1}) - Y_aY_bE(\epsilon_i\epsilon_{i+1})\\
&\quad + X_zx_iE(\eta_i\eta_{i-1}) - X_aX_bE(\eta_{i-1}\eta_i) - X_ax_{i-1}E(\eta_i^2) + X_aX_bE(\eta_i^2)\Big)\\
&= \frac{1}{4} * \Big(-x_{i-1}x_i\sigma_i\sigma_{i+1}\rho_i + X_bx_i\sigma_i\sigma_{i+1}\rho_i + y_{i-1}y_{i+1}\sigma_i^2 - Y_by_{i+1}\sigma_i^2 - y_iy_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1}\\
&\quad + Y_by_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} - \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i - Y_ay_{i-1}\sigma_i^2 + Y_aY_b\sigma_i^2 + Y_ay_i\sigma_{i-1}\sigma_i\rho_{i-1}\\
&\quad - Y_aY_b\sigma_{i-1}\sigma_i\rho_{i-1} + X_ax_{i-1}\sigma_i\sigma_{i+1}\rho_i - X_aX_b\sigma_i\sigma_{i+1}\rho_i - x_ix_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1}\\
&\quad + X_bx_{i+1}\sigma_{i-1}\sigma_i\rho_{i-1} + x_{i-1}x_{i+1}\sigma_i^2 - X_bx_{i+1}\sigma_i^2 - y_{i-1}y_i\sigma_i\sigma_{i+1}\rho_i + Y_by_i\sigma_i\sigma_{i+1}\rho_i\\
&\quad - \sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i + Y_ay_{i-1}\sigma_i\sigma_{i+1}\rho_i - Y_aY_b\sigma_i\sigma_{i+1}\rho_i + X_ax_i\sigma_{i-1}\sigma_i\rho_{i-1} - X_aX_b\sigma_{i-1}\sigma_i\rho_{i-1}\\
&\quad - X_ax_{i-1}\sigma_i^2 + X_aX_b\sigma_i^2\Big)\\
&= \frac{1}{4} * \Big(\sigma_i^2(y_{i-1}y_{i+1} - Y_by_{i+1} - Y_a\,y_{i-1} + Y_aY_b) + \sigma_i^2(x_{i-1}x_{i+1} - X_bx_{i+1} - X_ax_{i-1} + X_aX_b)\\
&\quad - \sigma_{i-1}\sigma_i\rho_{i-1}(y_iy_{i+1} - Y_by_{i+1} - Y_ay_i + Y_aY_b) - \sigma_{i-1}\sigma_i\rho_{i-1}(x_ix_{i+1} - X_bx_{i+1} - X_ax_i + X_aX_b)\\
&\quad - \sigma_i\sigma_{i+1}\rho_i(x_{i-1}x_i - X_bx_i - X_ax_{i-1} + X_aX_b) - \sigma_i\sigma_{i+1}\rho_i(y_{i-1}y_i - Y_by_i - Y_ay_{i-1} + Y_aY_b)\\
&\quad - 2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big)\\
&= \frac{1}{4} * \Big(\sigma_i^2\Big((y_{i+1} - Y_a)(y_{i-1} - Y_b) + (x_{i+1} - X_a)(x_{i-1} - X_b)\Big) - \sigma_{i-1}\sigma_i\rho_{i-1}\Big((y_{i+1} - Y_a)(y_i - Y_b)\\
&\quad + (x_{i+1} - X_a)(x_i - X_b)\Big) - \sigma_i\sigma_{i+1}\rho_i\Big((y_i - Y_a)(y_{i-1} - Y_b) + (x_i - X_a)(x_{i-1} - X_b)\Big) - 2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i\Big).
\end{aligned}$$

The derivation of $\mathrm{Cov}(A_i,B_i)$ follows similarly, yielding

$$\begin{aligned}
\mathrm{Cov}(A_i,B_i) &= \frac{1}{4} * \Big(-\sigma_y^2\Big((y_{i+1} - Y_b)(Y_{i+1} - Y_a) + (x_{i+1} - X_b)(x_{i+1} - X_a)\Big)\\
&\quad - \sigma_{i+1}^2\Big((x_i - X_b)(x_i - X_a) + (y_i - Y_b)(y_i - Y_a)\Big) + \sigma_i\sigma_{i+1}\rho_i\Big((x_i - X_a)(x_{i+1} - X_b) + (x_i - X_b)(x_{i+1} - X_a)\Big)\\
&\quad + \sigma_i\sigma_{i+1}\rho_i\Big((y_i - Y_b)(y_{i+1} - Y_a) + (y_{i+1} - Y_b)(y_i - Y_a)\Big) - 2\sigma_{i+1}^2\,\sigma_i^2(1 - \rho_i^2)\Big).
\end{aligned}$$

And we further obtain

$$\begin{aligned}\mathrm{Cov}(A_i, B_{i+1}) = \frac{1}{4} * \Big(& \sigma_{i+1}^2 \Big((x_i - X_a)(x_{i+2} - X_b) + (y_i - Y_a)(y_{i+2} - Y_b) \Big) \\ & - \sigma_i \sigma_{i+1} \rho_i \Big((y_{i+1} - Y_a)(y_{i+2} - Y_b) + (x_{i+1} - X_a)(x_{i+2} - X_b) \Big) \\ & - \sigma_{i+1} \sigma_{i+2} \rho_{i+1} \Big((x_i - X_a)(x_{i+1} - X_b) + (y_i - Y_a)(y_{i+1} - Y_b) \Big) \\ & - 2\sigma_i \sigma_{i+1}^2 \sigma_{i+2} \rho_i \rho_{i+1} \Big) \Big).\end{aligned}$$

Summing Triangles in an Arc

We have developed expressions for the three cases of covariance between a triangle in one polygon and the triangles in an adjacent polygon which touch it. The next step is to sum the covariances for all traingles formed by an arc. Assume that an arc which separates polygons A and B has $m+1$ points. There will be m triangles in the arc-sector in polygon A (A_i, $i=1 \ldots m$) and m triangles in the arc-sector in polygon B (B_j, $j=1 \ldots m$). The covariance between polygons A and B is the sum of the triangle covariances: i.e.,

$$\begin{aligned}\mathrm{Cov}(A,B) = {} & \mathrm{Cov}(A_1,B_1) + \mathrm{Cov}(A_1,B_2) + \mathrm{Cov}(A_2,B_1) + \mathrm{Cov}(A_2,B_2) + \mathrm{Cov}(A_2,B_3) + \mathrm{Cov}(A_3,B_2) + \mathrm{Cov}(A_3,B_3) + \mathrm{Cov}(A_3,B_4) \\ & + \ldots \ \ldots \ \ldots \mathrm{Cov}(A_i,B_{i-1}) + \mathrm{Cov}(A_i,B_i) + \mathrm{Cov}(A_i,B_{i+1}) + \ldots \ \ldots \ \ldots \mathrm{Cov}(A_{m-1},B_{m-2}) + \mathrm{Cov}(A_{m-1},B_{m-1}) + \mathrm{Cov}(A_{m-1},B_m) \\ & + \mathrm{Cov}(A_m,B_{m-1} + \mathrm{Cov}(A_m,B_m).\end{aligned}$$

Or, if we define $\mathrm{Cov}(A_1,B_0) = 0$ and $\mathrm{Cov}(A_m,B_{m+1}) = 0$, we can use the summation

$$\mathrm{Cov}(A,B) = \sum_{i=1}^{m} \Big(\mathrm{Cov}(A_i,B_{i-1}) + \mathrm{Cov}(A_i,B_i) + \mathrm{Cov}(A_i,B_{i+1}) \Big). \tag{14}$$

Now, define

$$\begin{aligned} \tilde{x}_{ai-1} &= x_{i-1} - X_a & \tilde{x}_{bi-1} &= x_{i-1} - X_b \\ \tilde{x}_{ai} &= x_i - X_a & \tilde{x}_{bi} &= x_i - X_b \\ \tilde{y}_{ai-1} &= y_{i-1} - Y_a & \tilde{y}_{bi-1} &= y_{i-1} - Y_b \\ \tilde{y}_{ai} &= y_i - Y_a & \tilde{y}_{bi} &= y_i - Y_b \end{aligned}$$

etc...

Then, substituting the individual covariance terms into Equation 14 and rearranging yields

$$\begin{aligned}\mathrm{Cov}(A,B) = \frac{1}{4} * \sum_{i=1}^{m} \Big(& \sigma_i^2 (\tilde{x}_{ai+1}\tilde{x}_{bi-1} + \tilde{y}_{ai+1}\tilde{y}_{bi-1}) - \sigma_{i-1}\sigma_i\rho_i (\tilde{x}_{ai+1}\tilde{x}_{bi} + \tilde{y}_{ai+1}\tilde{y}_{bi}) - \sigma_i\sigma_{i+1}\rho_{i+1}(\tilde{x}_{ai}\tilde{x}_{bi-1} + \tilde{y}_{ai}\tilde{y}_{bi-1}) \\ & - 2\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_i\rho_{i+1} \Big) \\ & - \Big(\sigma^2{}_i(\tilde{x}_{ai+1}\tilde{x}_{bi+1} + \tilde{y}_{ai+1}\tilde{y}_{bi+1}) + \sigma_{i+1}^2(\tilde{x}_{ai}\tilde{x}_{bi} + \tilde{y}_{ai}\tilde{y}_{bi}) - \sigma_i\sigma_{i+1}\rho_i (\tilde{x}_{ai}\tilde{x}_{bi+1} + \tilde{y}_{ai}\tilde{y}_{bi+1}) \\ & - \sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{ai+1}\tilde{x}_{bi} + \tilde{y}_{ai+1}\tilde{y}_{bi}) + 2\sigma_i^2\sigma_{i+1}^2 (1 - \rho_i^2) \Big) + \Big(\sigma_{i+1}^2(\tilde{x}_{ai}\tilde{x}_{bi+2} + \tilde{y}_{ai}\tilde{y}_{bi+2}) \\ & - \sigma_i\sigma_{i+1}\rho_i(\tilde{x}_{ai+1}\tilde{x}_{bi+2} + \tilde{y}_{ai}\tilde{y}_{bi+2}) - \sigma_{i+1}\sigma_{i+2}\rho_{i+1}(\tilde{x}_{ai}\tilde{x}_{bi+1} + \tilde{y}_{ai}\tilde{y}_{bi+1}) - 2\sigma_i\sigma_{i+1}^2\sigma_{i+2}\rho_i\rho_{i+1}) \Big)\end{aligned}$$

$$\begin{aligned} = \frac{1}{4} * \sum_{i=1}^{m} \Big(& \sigma_i^2 \Big(\tilde{x}_{ai+1}(\tilde{x}_{bi-1} - \tilde{x}_{bi+1}) + \tilde{y}_{ai+1}(\tilde{y}_{bi-1} - \tilde{y}_{bi+1}) \Big) + \sigma_{i+1}^2 \Big(\tilde{x}_{ai}(\tilde{x}_{bi+2} - \tilde{x}_{bi}) + \tilde{y}_{ai}(\tilde{y}_{bi+2} - \tilde{y}_{bi}) \Big) \\ & - s_{i-1}(\tilde{x}_{ai+1}\tilde{x}_{bi} + \tilde{y}_{ai+1}\tilde{y}_{bi}) - s_{i+1}(\tilde{x}_{ai}\tilde{x}_{bi+1} + \tilde{y}_{ai}\tilde{y}_{bi+1}) - s_i \Big(\tilde{x}_{ai}(\tilde{x}_{bi-1} - \tilde{x}_{bi+1}) + \tilde{y}_{ai}(\tilde{y}_{bi-1} - \tilde{y}_{bi+1}) \Big) - s_i \Big(\tilde{x}_{ai+1}(\tilde{x}_{bi+2} - \tilde{x}_{bi}) \\ & + \tilde{y}_{ai+1}(\tilde{y}_{bi+2} - \tilde{y}_{bi}) \Big) \Big) - \frac{1}{2} * \sum_{i=1}^{m} \Big(s_i(s_{i-1} - s_i + s_{i+1}) + \sigma_i^2\sigma_{i+1}^2 \Big)\end{aligned}$$

where s_i is as in Equation 11. While this expression is rather imposing, considerable simplification has been obtained through the use of matrix notation, the development of which is beyond the scope of this paper, but which is available from the authors on request.

DISCUSSION

Potential Applications

Some mention of potential applications of the variance expression may help to encourage further research in this area. Analysis of the variance of polygon areas may provide answers to such questions as

- What is the effect of digitizing technique (point versus stream mode) on polygon area variance?
- What is the relationship between polygon complexity or shape and polygon area variance?

- What effect does a reduction in positional accuracy of points have on polygon area variation?
- How does the degree of area variation in a map compare to the degree of attribute variation? In other words, which would best improve the accuracy of a cartographic modeling application: improved sampling for more precise attributes or improvements in mapping for more precise boundaries?

As stated earlier, one of the most important uses of a model of polygon variance may be to further study the propagation of errors in overlay analysis. Sensitivity studies can be performed, which assign values of σ based upon source map scale in order to assess the effect of overlaying maps of varying scale and precision. Knowledge of area variance alone may be informative and helpful to the cartographic modeler. For instance, suppose a number of sites are being analyzed for recreational development potential, and they are compared on the basis of a score resulting from a cartographic model. The variance of the score could be calculated as in Equation 2. If two sites have very similar scores, it would appear that an arbitrary choice of one or the other could be made. However, if one site were to have a considerably lower variance associated with its score, it may be a more desirable choice because of the reduction in uncertainty.

Determining Model Parameters

The expressions for polygon variance and covariance depend upon the coordinates and the σ's and ρ's which indicate the variability and correlation of points in an arc. Several possibilities exist for selecting values of σ. In Chrisman's (1982) work in this area, close examination of the steps involved in producing a map provided deductive estimates of individual error components, which were propagated to arrive at an overall estimate of positional accuracy. Such an approach has been suggested in the Digital Cartographic Data Standard (Morrison, 1988). Another technique is possible for maps with an accuracy standard stated in terms of an error distance and probability (e.g., 90 percent of tested points fall within 1/20 inch at map scale). If a normal distribution is assumed for points, then the probability and distance will imply a standard deviation for points (Keefer *et al.*, 1988). A more costly but more reliable procedure would involve comparison of features on the map with phenomena in the field. Such verification is usually restricted to testing well-defined points, which may exhibit different error structures from poorly-defined features.

The choice of the correlation coefficient, ρ, may be more difficult. In an analysis of digitizing error, Keefer *et al.*, (1988) used time series analysis to detect serial correlation of errors. Similar techniques may be used to evaluate correlation of overall coordinate errors. One advantage of an algorithm for calculating polygon variance is the capacity for performing sensitivity analyses in order to determine the impact of correlation on the resulting variance.

In some cases, a single σ and ρ may suffice for all arcs in a map. In other cases, knowledge about the ability to locate various boundaries may suggest the use of different parameters for different arcs. For example, in maps derived from interpretation of color-infrared photographs, some boundaries (such as those between water and land) may be discernible with much greater precision than other boundaries (such as those between vegetation types with similar spectral signatures). In these instances, assumption of different σ's for the different arcs may be justified. In any case, it is logical to consider the values for σ and ρ to be attributes of an arc, and to be maintained as such when overlaying polygons. Then, the values may be made available to computer programs which could calculate polygon variances for the resulting overlay map.

Distribution Assumptions

Until now, we have not suggested a statistical distribution for polygon area; derivations of variance and covariance have been made without assuming any specific distribution for point location errors. However, knowledge about variance of polygon area could be more useful if the distribution is known. Then, the probability of certain events occurring may be inferred. For example, "sliver" polygons which arise from overlay and intersection of similar arcs present a problem in interpretation. Do such polygons represent significant features on the ground, or are they artifacts of the map overlay process? Generally, such polygons are small in size. In fact, some software modules provide for the arbitrary elimination of polygons smaller than some threshold area, on the assumption that they must be insignificant. If the distribution of polygon area were known, a p-value could be obtained which would indicate the probability of getting a sliver of the observed size when, in fact, none exists. Such a statement could be useful in the determination of which sliver polygons to eliminate.

An assumption of normal errors in point location has been suggested by Chrisman (1982), and seems quite reasonable. We could express this as

	$\epsilon_i \sim N(0,\sigma_i^2)$	$\eta_i \sim N(0,\sigma_i^2)$.
Then,	$X_i \sim N(x_i,\sigma_i^2)$	$Y_i \sim N(y_i,\sigma_i^2)$
and	$(X_i - X_c) \sim N(\tilde{x}_i,\sigma_i^2)$	$(Y_i - Y_c) \sim N(\tilde{y}_i,\sigma_i^2)$.

Now, the formula for area of a triangle (Equation 13) can be rewritten as a function of a random determinant: i.e.,

$$A_i = \frac{1}{2} * \begin{vmatrix} (X_i - X_a) & (X_{i+1} - X_a) \\ (Y_i - Y_a) & (Y_{i+1} - Y_a) \end{vmatrix}$$

If we assume that adjacent coordinates are independent and normally distributed, we can apply the findings of Nicholson (1958) and Nyquist *et al.* (1954), who reported on the mean and variance of random determinants, and noted that they could be approximated by a normal distribution. Indeed, if $\rho = 0$, our expression for variance of triangle area in Equation 7 agrees with the variance of a 2 by 2 random normal determinant described by Nicholson (1958). Thus, in the absence of correlation between coordinate errors, triangle areas are approximately normally distributed and the polygon area is the sum of n near-normal random variables. Because these triangles are certainly not independent, even if the coordinate errors are, the Central Limit Theorem is not strictly applicable. However, we propose that a normal distribution may be a reasonable approximation. In fact, preliminary simulations of errors in polygon boundaries have resulted in distributions of area that are statistically indistinguishable from the normal. Caution is needed when making such assumptions about the distribution of polygon area; it is possible that polygons composed of only a few points (as most sliver polygons are) may be poorly modeled by the normal distribution. Further evaluation of the distribution of area through simulation is being conducted.

SUMMARY

Analysis of the errors in GIS systems has been a high research priority in recent years. The desire for a means of characterizing the uncertainty of estimates derived from GIS analyses has been expressed by a number of authors (Bennett, 1977; Aronoff, 1982; Chrisman, 1984). While positional accuracy statements have been discussed at length, statements regarding the uncertainty of area estimates have been rare.

The objective of this work has been to characterize the mean and variance of polygon areas computed in a vector GIS. These values are vital if GIS are to continue to be used in everyday decision-making. The arc-node data structure was chosen as a framework due to its current popularity and topological orientation, which allows locational precision to be stored as an arc attribute. Based on a few fundamental assumptions regarding the positional accuracy of point locations, it has been possible to derive expressions for variance of polygon area as a function of random variables. Because choroplethic maps involve a number of contiguous polygons, total map variance requires an expression for covariance of area of adjacent polygons. Because both of these expressions required the use of a polygon centroid, a consistent means of locating a centroid location which minimizes the variance of polygon area was described. A centroid-insensitive expression would be appealing; however, it may be more difficult to develop covariance expressions in the absence of such reference points. The undesirability of a centroid-oriented expression may be mitigated by evidence that it reliably predicts the variability of area observed in simulations of error-influenced polygons.

The availability of expressions for the variance of polygon area presents an opportunity for considerable further research in GIS errors and their implications for decision-making. Some examples include investigations into error propagation in map overlay; effect of polygon size, shape, and complexity on area errors; evaluation of map accuracy standards; and sensitivity analysis for determining the impact of individual map error components.

Additional work is needed to properly apply the expressions developed herein. The equations require the specification of point variability and the correlation between coordinate errors at adjacent points. More study is needed to evaluate the sensitivity of these assumptions. Methods for determining reasonable values for σ and ρ will need to be examined. Finally, before the expression of polygon variance can be widely used, software must be developed to incorporate accuracy assumptions as arc attributes, and calculate and report polygon variances and covariances in a useful form.

While the expression for polygon variance may be useful to those concerned with the reliability of GIS analyses, additional expressions are needed. Statistical characterization of errors in distances (between points and between a point and an arc) and lengths are being developed using the techniques reported here. However, as with most developments in information processing, the beneficial application of these techniques will require users who are aware of, and concerned about, the quality of information being used and produced by geographic information systems.

REFERENCES

Aronoff, S., 1982. Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing* 48(8):1309–1312.

Bailey, Robert G., 1988. Problems with using overlay mapping for planning and their implications for geographic information systems. *Environmental Management* 12(1):11–17.

Bennett, H.C., 1977. The cartographic data base-reliability or chaos? *Proceedings of American Congress on Surveying and Mapping*, pp. 675–680.

Bondesson, Lennart, 1986. *Estimation of Standard Errors of Area Estimates of Forest Compartments Obtained by Traversing*. Swedish University of Agricultural Sciences, Section of Forest Biometry, S-901 83 Umea, Sweden, Report 24, 49 p.

Burrough, P.A., 1986. *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford University Press, New York, 193 p.

Chrisman, Nicholas R., 1982. *Methods of Spatial Analysis Based on Error in Categorical Maps*. PhD dissertation, University of Bristol.

———, 1984. The role of quality information in the long-term functioning of a geographic information system. *Cartographica* 21:79–87.

———, 1987. The accuracy of map overlays: a reassessment. *Landscape and Urban Planning* 14:427–439.

Chrisman, Nicholas R., and Brian S. Yandell, 1988. Effects of point error on area calculations: a statistical model. *Surveying and Mapping* 48(4):241–246.

Goodchild, Michael, and Odette Dubuc, 1987. A model of error for choroplethic maps, with applications to geographic information systems. *Auto Carto 8*, pp. 165–174.

Keefer, Brenton J., James L. Smith, and Timothy G. Gregoire, 1988. Simulating manual digitizing error with statistical models. *GIS/LIS '88 Proceedings*, pp. 475–483.

MacDougall, E.B., 1975. The accuracy of map overlays. *Landscape Planning* 2:23–30.

McAlpine, J.R., and B.G. Cook, 1971. Data reliability from map overlay, *Proceedings of the 43rd Congress of the Australian and New Zealand Association for the Advancement of Science*, Brisbane, Australia.

Mead, D.A., 1982. Assessing data quality in geographic information systems, *Remote Sensing for Resource Management* (Johannsen and Sanders, ed.), Soil Conservation Society of America, pp. 51–62.

Morrison, Joel, ed., 1988. The proposed standard for digital cartographic data. *The American Cartographer* 15(1).

Neumyvakin, Yu. K., and A.I. Panfilovich, 1982. Specific features of using large-scale mapping data in planning construction and land farming. *Proceedings AUTO-CARTO 5*, pp. 733–738.

Nicholson, W.L., 1958. On the distribution of 2 × 2 random normal determinants. *Annals of Mathematical Statistics*, 29:575–580.

Nyquist, H., S.O. Rice, and J. Riordan, 1954. The distribution of random determinants. *Quarterly of Applied Mathematics* 12(2):97–104.

Peuker, T., and N. Chrisman, 1975. Cartographic data structures. *The American Cartographer* 2(1):55–69.

Walsh, Stephen J., D.R. Lightfoot, and David Butler, 1987. Recognition and assessment of error in geographic information systems. *Photogrammetric Engineering and Remote Sensing* 53(10):1423–1430.

APPENDIX

The derivation of the minimum-variance centroid begins with an expression for polygon area variance as a function of coordinates, σ's and ρ's. Rewriting the formula for the variance of a polygon of n points (Equation 12) yields

$$\begin{aligned} V = \frac{1}{4} * \sum_{i=1}^{n} \Big(& \sigma_{i+1}^2\big((X_i - X_c)^2 + (Y_i - Y_c)^2\big) + \sigma_i^2\big((X_{i+1} - X_c)^2 + (Y_{i+1} - Y_c)^2\big) \\ & + 2\sigma_i^2\sigma_{i+1}^2(1-\rho_i^2) - 2\sigma_i\sigma_{i+1}\rho_i\big((X_i - X_c)(X_{i+1} - X_c) + (Y_i - Y_c)(Y_{i+1} - Y_c)\big) \\ & + 4\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i + 2\sigma_i\sigma_{i+1}\rho_i\big((X_{i-1} - X_c)(X_i - X_c) + (Y_{i-1} - Y_c)(Y_i - Y_c)\big) \\ & + 2\sigma_{i-1}\sigma_i\rho_{i-1}\big((X_i - X_c)(X_{i+1} - X_c) + (Y_i - Y_c)(Y_{i+1} - Y_c)\big) \\ & - 2\sigma_i^2\big((X_{i-1} - X_c)(X_{i+1} - X_c) + (Y_{i-1} - Y_c)(Y_{i+1} - Y_c)\big)\Big) \\ = \frac{1}{4} * \sum_{i=1}^{n} & (\phi_x + \phi_y + \phi) \end{aligned} \tag{15}$$

where

$$\begin{aligned} \phi_x = {} & \sigma_{i+1}^2(X_i^2 - 2X_iX_c + X_c^2) + \sigma_i^2(X_{i+1}^2 - 2X_{i+1}X_c + X_c^2) - 2\sigma_i\sigma_{i+1}\rho_i(X_iX_{i+1} - X_iX_c - X_{i+1}X_c + X_c^2) \\ & + 2\sigma_i\sigma_{i+1}\rho_i(X_{i-1}X_i - X_iX_c - X_{i-1}X_c + X_c^2) + 2\sigma_{i-1}\sigma_i\rho_{i-1}(X_iX_{i+1} - X_iX_c - X_{i+1}X_c + X_c^2) \\ & - 2\sigma_i^2(X_{i-1}X_{i+1} - X_{i-1}X_c - X_{i+1}X_c + X_c^2) \end{aligned}$$

$$\begin{aligned} \phi_y = {} & \sigma_{i+1}^2(Y_i^2 - 2Y_iY_c + Y_c^2) + \sigma_i^2(Y_{i+1}^2 - 2Y_{i+1}Y_c + Y_c^2) - 2\sigma_i\sigma_{i+1}\rho_i(Y_iY_{i+1} - Y_iY_c - Y_{i+1}Y_c + Y_c^2) \\ & + 2\sigma_i\sigma_{i+1}\rho_i(Y_{i-1}Y_i - Y_iY_c - Y_{i-1}Y_c + Y_c^2) + 2\sigma_{i-1}\sigma_i\rho_{i-1}(Y_iY_{i+1} - Y_iY_c - Y_{i+1}Y_c + Y_c^2) \\ & - 2\sigma_i^2(Y_{i-1}Y_{i+1} - Y_{i-1}Y_c - Y_{i+1}Y_c + Y_c^2) \end{aligned}$$

$$\phi = 2\sigma_i^2\sigma_{i+1}^2(1 - \rho_i^2) + 4\sigma_{i-1}\sigma_i^2\sigma_{i+1}\rho_{i-1}\rho_i.$$

The derivative with respect to X_c is

$$\begin{aligned} \frac{\partial V}{\partial X_c} = \frac{\partial \phi_x}{\partial X_c} = \frac{1}{2} * \sum_{i=1}^{n} (& -\sigma_{i+1}^2X_i + \sigma_{i+1}^2X_c + \sigma_i^2X_{i-1} - \sigma_i^2X_c + \sigma_i\sigma_{i+1}\rho_iX_{i+1} \\ & - \sigma_i\sigma_{i+1}\rho_iX_{i-1} - \sigma_{i-1}\sigma_i\rho_{i-1}X_i - \sigma_{i-1}\sigma_i\rho_{i-1}X_{i+1} + 2\sigma_{i-1}\sigma_i\rho_{i-1}X_c). \end{aligned}$$

Setting the derivative to zero and rearranging yields

$$\frac{1}{2} * \sum_{i=1}^{n} X_c(-\sigma_{i+1}^2 + \sigma_i^2 - 2\sigma_{i-1}\sigma_i\rho_{i-1}) = \frac{1}{2} * \sum_{i=1}^{n} X_i(-\sigma_{i+1}^2 - \sigma_{i-1}\sigma_i\rho_{i-1})$$

$$+ \frac{1}{2} * \sum_{i=1}^{n} X_{i-1}(\sigma_i^2 - \sigma_i\sigma_{i+1}\rho_i) + \frac{1}{2} * \sum_{i=1}^{n} X_{i+1}(\sigma_i\sigma_{i+1}\rho_i - \sigma_{i-1}\sigma_i\rho_{i-1}).$$

By making the following substitutions:

$$\sum_{i=1}^{n} X_{i-1}\sigma_i^2 = \sum_{i=1}^{n} X_i\sigma_{i+1}^2; \qquad \sum_{i=1}^{n} X_{i-1}\sigma_i\sigma_{i+1}\rho_i = \sum_{i=1}^{n} X_i\sigma_{i+1}\sigma_{i+2}\rho_{i+1};$$

$$\sum_{i=1}^{n} X_{i+1}\sigma_i\sigma_{i+1}\rho_i = \sum_{i=1}^{n} X_i\sigma_{i-1}\sigma_i\rho_{i-1}; \sum_{i=1}^{n} X_{i+1}\sigma_i - 1\sigma_i\rho_{i-1} = \sum_{i=1}^{n} X_i\sigma_{i-2}\sigma_{i-1}\rho_{i-2}$$

we get $X_c * \left(\sum_{i=1}^{n} - 2\,\sigma_i\sigma_{i+1}\rho_i\right) = \sum_{i=1}^{n} -X_i(\sigma_{i+1}\sigma_{i+2}\rho_{i+1} + \sigma_{i-2}\sigma_{i-1}\rho_{i-2})$.

Thus, we have

$$X_c = \frac{\sum_{i=1}^{n} X_i\,(\sigma_{i-2}\,\sigma_{i-1}\,\rho_{i-2} + \sigma_{i+1}\,\sigma_{i+2}\,\rho_{i+1})}{\sum_{i=1}^{n} (2\sigma_i\,\sigma_{i+1}\,\rho_i)}.$$

The procedure for ϕ_y follows similarly, yielding

$$Y_c = \frac{\sum_{i=1}^{n} Y_i(\sigma_{i-2}\sigma_{i-1}\rho_{i-2} + \sigma_{i+1}\sigma_{i+2}\rho_{i+1})}{\sum_{i=1}^{n}(2\sigma_i\sigma_{i+1}\rho_i)}.$$

To verify that this is a minimum, we can take the second partial derivative of Equation 15: i.e.,

$$\frac{\partial^2 V}{\partial X_c^2} = \frac{1}{2} * \sum_{i=1}^{n} (\sigma_{i+1}^2 - \sigma_i^2 + 2\sigma_{i-1}\sigma_i\rho_{i-1}) = \sum_{i=1}^{n} (\sigma_{i-1}\sigma_i\rho_i).$$

The fact that this expression (and the similar one for $\frac{\partial^2 V}{\partial Y_c^2}$) is positive (for $\rho_i > 0$) indicates that the solution obtained is a minimum.