Effects of Point Error on Area Calculations: A Statistical Model

Nicholas R. Chrisman and Brian S. Yandell

ABSTRACT. When coordinate measurements are subject to error, areas calculated from the points will also be in error. Our statistical model describes polygons bounded by straight lines between well-defined points. Error in each point must be independent and identically distributed. This article presents two statistical findings based on this model:

(1) the standard algorithm for area is an unbiased estimator of true area

(2) a formula for the variance of area based on parameters of point error.

The paper considers the nature of developments required to handle less restrictive cases.

BACKGROUND

Area calculations are a standard product of land-related data systems, whether manual or automated. These calculations can be based on maps or field measurements of position. No matter what the source of data, there will be some amount of error in the measurement of positions. Yet it is rare to find area calculations with an estimate of error attached. The phrase "*n* acres more or less" commonly appears in legal descriptions, but the amount more or less is not specified. With the development of computer systems, area calculations are carried out to excruciating numbers of digits, but still without a measure of accuracy. None of the current sophisticated software packages sold as geographic information systems temper their area calculations with an error estimate.

This paper presents a statistical model that permits the calculation of the variance of areas under certain simple conditions. A few examples will be presented. Then more complex situations will be discussed.

Two recent publications have presented models for error in areas derived from points. Chrisman (1982a) developed an earlier version of this model, but the result was unsatisfactory in a number of ways. In particular, the error varied with the origin and orientation of the coordinate system. Neumyvakin and Panfilovich (1982) had a model that depended on the coordinate system, although they allowed an arbitrary covariance structure. Burrough (1986) reviews related literature on data quality and sources of error in geographic information systems.

Nicholas R. Chrisman is Assistant Professor in the Department of Geography at the University of Washington. His research deals with error in maps and GIS data. Brian Yandell is Associate Professor in the Department of Statistics and Horticulture at the University of Wisconsin-Madison. His research covers many topics, including aspects of spatial statistics.

ASSUMPTIONS

Area calculations apply to regions of a surface. Calculations for large regions of the earth require consideration of geoidal curvature, but many applications are well-served by a planar projection. The focus of this paper is restricted to planar polygons described by strings of points connected by straight line segments. Although some engineering drafting includes circular arcs or complex curves, many systems for digital cartography (particularly those that include advanced analytical features like polygon overlay) accept the restriction to straight line segments. A model of error in a more complex situation involving parametric curves will not be easy to develop since it will not be clear if the error is in the endpoints or in the parameters.

An N-sided polygon **P** is described by a sequence of points ($P_1, P_2, \ldots P_N$). Since there is a line between P_N and P_1, it is usual to duplicate P_1 as P_{N+1}. It will also be convenient to duplicate P_N as P_0 (See Figure 1). Each point P_i is located by a Cartesian coordinate pair (X_i,Y_i). The main assumption of this paper is that the error in the polygon is located at the points. Our model assumes that the measurements represented by (X_i,Y_i) comes from some bivariate distribution with a mean (x_i,y_i). The limitations of this assumption will be considered at the end of the paper.

The focus of this paper is on the calculation of area. If the "true" values (x,y) were available, the true area *a* could be calculated. The common trapezoid algorithm for the area of a polygon (A) can be algebraically simplified into Equation 1. Working with the more usual form of this equation lead to some of the problems in Chrisman's (1982a) previous approach to the problem. The goal of this paper is to determine how close A is to *a*.

$$A = 0.5 \sum X_i (Y_{i+1} - Y_{i-1}) \qquad (1)$$

Because the polygon is a closed loop, the storage of coordinates must be circular to permit Equation 1 to operate; $Y_{N+1} = Y_1$, and $Y_0 = Y_N$ (as explained before). This algorithm applies to simple polygons with a single exterior ring [see definition of polygon advanced as a national standard (Morrison 1988, 28)], but the result is easily extended to multiple rings, whether interior or exterior.

To develop a tractable model, we assume that the error at each point is independently and identically distributed around the local mean. We do not need to make any further assumption about the nature of the distribution (normal, log-normal . . .). Viewed in spatial terms, an independent, identical distribution creates an ellipse around each "true" value (see Figure 2). Each ellipse will have the same major and minor axis and share an angle of orientation. Viewed as a bivariate statistical distribution, the ellipses can be expressed as two variances and a correlation coefficient (Equation 2). The statistical formulation is more useful for the algebraic treatment of expected values performed below.

$$\text{var}(X_i) = \sigma^2_x;\ \text{var}(Y_i) = \sigma^2_y \text{ for } i = 1,N \qquad (2)$$
$$\text{cov}(X_i,Y_i) = \rho\, \sigma_x\, \sigma_y$$

Since the calculation of area in Equation 1 uses the X and Y axes differently, the variance terms must be tied to these axes, not the natural axes of the ellipse. Of course, the resultant model must not vary if the coordinate axes are rotated. This parametrization permits the axes to be rotated while describing the same physical errors.

For the purposes of describing the model, it is assumed that the three parameters of Equation 2 are known. In actual practice, they must be estimated from internal evidence, repeated measurement or testing (for a full discussion of these alternatives see the *Proposed Standard for Digital Cartographic Data,* Morrison 1988). The design of tests based on independent sources of higher accuracy has been the subject of protracted discussion in developing the American Society for Photogrammetry and Remote Sensing (Merchant 1983; 1987; Committee for Standards and Specifications 1985) proposed specification for large scale maps. The procedures sug-

gested for performing the ASPRS test will produce adequate information (Petersohn and Vonderohe 1982) to provide an estimate of the parameters of this model (variances and correlation).

A SIMPLE MODEL FOR ERROR IN AREA

For notational clarity, we define new random variables that describe the deviations between the observed coordinates and the mean for each point.

$$\epsilon_i = X_i - x_i \; ; \; \eta_i = Y_i - y_i \qquad (3)$$

Given the above assumptions, these two variables will have a mean of zero, and respective variances of σ^2_x and σ^2_y. Notice that adjacent points along the boundary should not show any particular correlation. This is a major restriction of this model. Less restrictive forms are considered below.

Given the assumptions and notation presented above, the measured area A can be decomposed through algebraic steps into Equation 4; the true area *a* with two error terms.

$$A = a + B + C \qquad (4)$$

where

$$B = 0.5 \sum \epsilon_i (y_{i+1} - y_{i-1}) - \eta_i (x_{i+1} - x_{i-1})$$

and

$$C = 0.5 \sum \epsilon_i (\eta_{i+1} - \eta_{i-1})$$

The first inspection of Equation 4 should concern the first moment; does A have *a* as its central tendency or expected value? Terms B and C have an expected value of zero, given the assumptions (specifically that ϵ_i and η_i have a zero expected value). Hence A is an unbiased estimator of *a* under these conditions. This is a fortunate result because A is the value normally reported. Of course, expected value is a property of a whole distribution, not a single realization.

Bias is not the only component of accuracy and reliability. The second moment of the distribution (variance, though often reported as standard deviation) is required. The variance of A can be stated as an expected value of the variables defined above:

$$\text{var}(A) = E(B^2) + E(C^2) + 2E(BC) \qquad (5)$$

This last product has a zero expectation. The other terms expand to:

$$E(B^2) = 0.25 \sum \epsilon_i^2 (y_{i+1} - y_{i-1})^2 - \eta_i^2 (x_{i+1} - x_{i-1})^2 - 2\epsilon_i \eta_i (y_{i+1} - y_{i-1})(x_{i+1} - x_{i-1}) \qquad (6)$$

and

$$E(C^2) = 0.25 \sum \epsilon_i^2 (\eta_{i+1} - \eta_{i-1}) - 2\, \epsilon_i \epsilon_{i+1} \eta_i \eta_{i+1} \qquad (7)$$

Using the identities established, the expected values of B^2 and C^2 can be rewritten to derive the variance of A (Equation 8). This equation is invariant to rotations of the coordinates. The process of generating this equation involves algebraic substitution and deletion of terms with zero expected value not necessary to include in the paper.

$$\sigma_A^2 = 0.5\, N \sigma^2_x \sigma^2_y (1 - \rho^2) + 0.25 \sum [\sigma^2_y (x_{i+1} - x_{i-1})^2 - \sigma^2_x (y_{i+1} - y_{i-1})^2] - 0.5 \sum \rho\, \sigma_x \sigma_y (y_{i+1} - y_{i-1})(x_{i+1} - x_{i-1}) \qquad (8)$$

SAMPLE APPLICATIONS

Equation 8 can be applied when the spatial data fits the point model, the errors are not intercorrelated and the error parameters are available. This section will demonstrate what such results can show.

The preconditions used for the statistical model (polygons formed from discrete points and error only at the points) apply to a few realistic applications. In particular, property

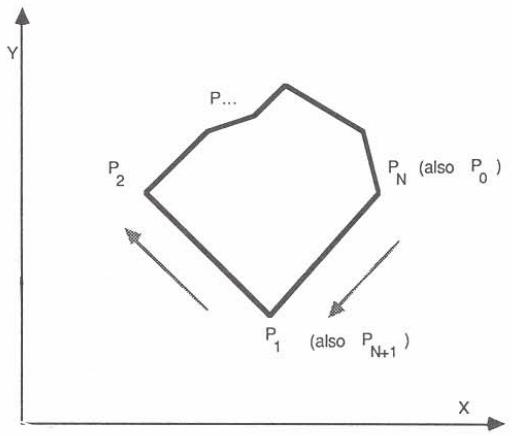

Figure 1: A simple polygon, showing duplication of first and last point.

parcel maps and similar features created by human laws and institutions are often specified by specific points on straight line boundaries. Cartographic compilation usually introduces some form of correlated error, but some procedures for point dictionary databases could fit this model.

To obtain the error parameters, some accuracy tests have been performed to determine the positional error associated with digital databases for parcel and related maps (see for example, Crossfield and Mezera 1982; Petersohn and Vonderohe 1982; Vonderohe and Chrisman 1985). The tests referenced were conducted according to a draft proposed standard prepared by the American Society of Photogrammetry (Merchant 1983), which defined accuracy in terms of the bias and precision (standard deviation). These figures assist in deriving the variance of area calculations because the bias component of the distribution is irrelevant to the area calculation. In the more recent proposal, the American Society for Photogrammetry and Remote Sensing (Merchant 1987) simplified its test to Root Mean Square Error, a figure which mixes together the bias and precision component. This may be proper if the goal is determination of positional accuracy standards. However, for use in derivations like area, the strategy should be to divulge more information about the distribution of the test results.

As an example, this study will report on the Digital Line Graph (DLG) data distributed by the U.S. Geological Survey. Tests determined the accuracy of the Public Land Survey System (PLSS) section corner information (Vonderohe and Chrisman 1985). This test compared the location of section corners in the DLG data with the corresponding coordinates determined by a completely independently conducted ground survey. The survey covered the survey township of Oregon, Wisconsin, which falls into two quadrangles (Oregon and Attica). On the Oregon quad sheet (based on 34 matching points), the test estimated $\sigma_x = 5.44$ meters and

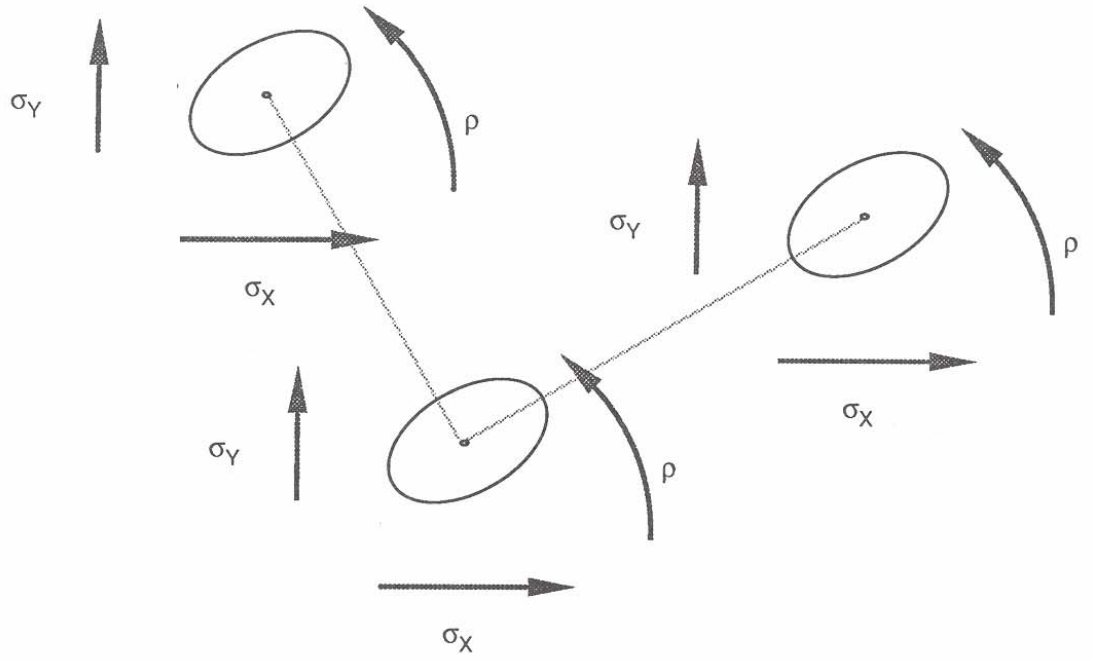

Figure 2: Parameters of point error.

σ_y = 3.46 meters; ρ = 0.026. The error ellipse was nearly oriented along the easting axis of the UTM projection. This error distribution describes the estimated error in positions for coordinates of other well-defined points digitized from this particular 1:24,000 quadrangle sheet. Any extension from the specific case to a more general one must be done cautiously. It would be unjustified nationally, but from the other sheets tested in Dane County, it may apply beyond the single quad sheet.

Equation 8 depends on the specific polygons tested. This demonstration uses a square mile section, the object tested in the Oregon quad. A section should have a nominal area of 2,588,881 m^2. If it is cartographically defined by four corner points, σ_A is 10,337 m^2. If it is defined, in better correspondence with the legal system, by section corners and quarter corners (assumed to be mathematically set for the purposes of the model), σ_A is 8,979 m^2. The lower figure comes from the shorter runs on the boundaries of the polygon. This reduces the second term in Equation 8 much more than the first term grows. In fact, the second term is five orders of magnitude larger than the first term and dominates the result in the cases examined.

The figures determined above are rather small, much below one percent of the area estimate. The positional accuracies were above one percent of the length of a section. This seems to indicate that area measurements are not overly sensitive to inaccuracies in positional data. This finding must be restricted to the types of polygon data modeled here, namely polygons constructed with well-defined points. More effort must be exerted to extend this model to polygons where the points are more arbitrary samples to indicate the general trend of the boundary. Such data is common in natural resource inventory and other circumstances where boundaries are formed by continuous curves. In these cases there will be much more likelihood of correlated error between adjacent points from the serial process of line following.

CORRELATED ERRORS

The previous analysis depends on simplifying terms due to the assumption that errors are independent. It has to be remembered that statistical independence differs from causal linkage. Errors can be correlated without a direct connection. In mapping, a drafter, or digitizer or stereoplotter operator might have certain predictable types of error, such as lagging behind a curve due to inertia. In the field as well, measurements in one area might depend on common sources of geodetic control or similar effects. Furthermore, any errors that vary with different terrain violate uniformity and create correlated errors. The first requirement of further research is better understanding of error distributions. A model incorporating correlated error could be refined in a number of different ways, but only a few possibilities might be required to model actual circumstances. Neumyvakin and Panfilovich (1982) developed their model of error in area measurements to include autocorrelation among all pairs of coordinates, although the only case they consider in detail has constant autocorrelations.

We present one type of correlated error under simple conditions. We assume that error is circular $\{\sigma_x = \sigma_y; \rho = 0\}$, and that the correlation occurs between adjacent points $\{\text{var}(X_i, X_{i+1}) = \text{var}(Y_i, Y_{i+1}) = \tau\sigma^2\}$. Under this model we assume that the number of sides to the polygon (N) is large and that τ is small. In this case A will still be an unbiased estimator of a, at least asymptotically over many polygons. The variance of A is considerably more complex and has not been fully determined. This approach suggests the use of Markov processes as the basis for further work.

A further problem with the simple model is that polygons do not always float independently in the void. Most polygon maps exhaustively partition the total study region. The errors on a particular border contribute to exactly two specific polygons though near the nodes the error involves other polygons.

Thus, although the formulas presented here might even be correct, errors will correlate to some extent between adjacent polygons. A comprehensive error model must accommodate the topology of the map. Chrisman (1982b) has suggested a different method to describe errors in area calculations that tries to deal with the effects of shared boundaries. Perhaps further research will unify the topological approach with the more rigorous statistical foundation presented in this article.

REFERENCES

Burrough, Peter A. (1986), *Principles of Geographical Information Systems for Land Resources Assessment,* Oxford, Clarendon Press.

Chrisman, Nicholas R. (1982a), "Methods of Spatial Analysis Based on Error in Categorical Maps," unpublished Ph.D. thesis, University of Bristol UK.

Chrisman, Nicholas R. (1982b), "A Theory of Cartographic Error and its Measurement in Digital Data Bases," *Proceedings AUTO-CARTO 5,* pp. 159–168.

Committee for Standards and Specifications (1985), "Accuracy Specifications for Large-Scale Line Maps," *Photogrammetric Engineering and Remote Sensing,* Vol. 51, no. 2, pp. 197–199.

Crossfield, James K. and David F. Mezera (1982), "The Westport Section Line Accuracy Test," *Surveying and Mapping,* Vol. 43, no. 1, pp. 47–52.

Merchant, Dean (1983), "Spatial Accuracy Standards for Large-Scale Line Maps," in H. Moellering, editor, *Papers from the Joint ACSM/ASP Session on Digital Cartographic Data Standards,* Columbus, OH, National Committee for Digital Cartographic Data Standards, Report #1, pp. 27–36.

Merchant, Dean (1987), "Spatial Accuracy Specification for Large Scale Topographic Maps," *Photogrammetric Engineering and Remote Sensing,* Vol. 53, no. 7, pp. 958–961.

Morrison, Joel, editor (1988), "The Proposed Standard for Digital Cartographic Data," *The American Cartographer,* Vol. 15, no. 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.

Petersohn, Cliff and Alan P. Vonderohe (1982), "Site-Specific Accuracy of Digitized Property Maps," *Proceedings AUTO-CARTO 5,* pp. 607–619.

Vonderohe, Alan P. and Nicholas R. Chrisman (1985), "Tests to Establish the Quality of Digital Cartographic Data: Some Examples from the Dane County Land Records Project," *Proceedings AUTO-CARTO 7,* pp. 552–559.