A THEORY OF CARTOGRAPHIC ERROR AND ITS MEASUREMENT IN DIGITAL DATA BASES

Nicholas R. Chrisman
University of Wisconsin - Madison
Madison, WI 53706

ABSTRACT

The processes of map production are necessarily approximate, and thus the resulting map contains a variety of error effects. This paper develops a theory of information content that applies to geometric details on a map, based on the epsilon distance model. Examining map production technology, the epsilon model provides a reasonable approximation of the error expected. This error can be measured, and an example is worked using data from the GIRAS digital files produced by USGS. Under conservative assumptions, 7 percent of the selected study area around Pittsburgh lies in zones of potential error.

INTRODUCTION

A model of map error based on a deductive approach can be derived from consideration of the objects measured and the processes that placed these objects on the map. By studying the sequential processes of map production and analysis, a model of error has much more coherence than the empirical results from a few specific maps. This paper concentrates maps of nominal scale data, such as land use or soils. These maps consist of polygons created by a network of lines, and the areas of these polygons are frequently calculated. These areas are then used with very little regard for their potential inaccuracies.

This study concentrates on the variability which is inherent in the map production process. For the sake of simplicity, divergence between the map and the earth's surface is termed "error". This word may be a bit harsh, but it should not imply that the map is "wrong", just that the map is limited.

INFORMATION MODELS

A fundamental model of spatial structure is captured by the topological approach to cartographic data structures (Corbett, 1975). However, topological relationships limit themselves to basic set theory and the abstract ramifications of dimensionality. The basic topological model does not consider the geometric detail of the map, so it is inadequate to model the variability of that detail.

Recursive Bands

One information model has been advanced by Thomas Poiker (formerly spelled Peucker, 1975) as a theory of the cartographic line. The theory derives from research initially developed for the removal of detail in cartographic representation (Douglas and Peucker, 1973).

Given a cartographic line represented as a set of straight line segments, the argument hinges on a definition of information content. Poiker suggests that the correct basis for using a more complex representation should be the measurement of deviation from a trend line. If a point is selected, two new trend lines are defined and the process is applied recursively. During the procedure a series of trend lines and deviations are calculated; these measurements define rectangles with one axis along the trend line, the sides displaced from the trend by the deviation.

Epsilon Distance

As a model of variability for a line, the recursive banding approach has disadvantages. In particular, the bounding rectangles often contain wide areas far from the line, while the inflection points are directly on the edge. A model of variability might start with another definition of the information content of a cartographic line, based on the theory of epsilon distance (Perkal, 1956; 1966). This theory is no more "correct" than Poiker's; it merely offers a better opportunity for this particular application.

Given a cartographic line as a straight line approximation, it might be supposed that the true line lies within a constant tolerance, epsilon, of the measured line. For a straight line segment this locus is simple, consisting of the union of a rectangle parallel to the segment, twice epsilon wide, with circles of radius epsilon centered at each end point. By union of this simple figure, more complex lines can be handled. The band can also be described as the area occupied by rolling a ball along the line.

SOURCES OF ERROR

A map is produced by a specific progression of procedures which accept, process and transmit information in various forms. The amount of error contributed depends on the technical details of each step, so error analysis cannot be fixed for all maps. Blumenstock (1953) pioneered the deductive approach to overall variability resulting from the accumulation of errors in each step. His analysis covered sampling temperature and generating an isarithmic map. Here a similar approach will be taken for cartographic features such as lines which bound land use zones.

This section will review some of the more recurrent forms of cartographic error. It does not cover all conceivable error sources, nor does it probe in detail the sources mentioned. The goal is to demonstrate the applicability of the epsilon model to a representative range of the problems contributing to actual error.

Locating Ground Position

Measurement of the earth's surface is an age-old science which has attained remarkable sophistication. Current high performance in geodesy, surveying and photogrammetry results from continual efforts, often spread over hundreds of years or more, to perfect these disciplines. It is interesting to note that these disciplines devote substantial concern to

the mathematical study of error, providing ever higher standards of performance. However, the promise of technical perfection does not mean that it is attained. Any map is a fossil, reflecting the technology used in its production. Due to the lags of production, it is very difficult to keep any map coverage (from national topography to municipal cadastre) uniformly updated to the best current technology. In addition, organizations without specialized cartographic expertise (such as planning agencies) may not have access to the latest techniques and equipment.

Errors in locating ground position can be minimized by spending more money, but it would be foolish to increase accuracy in this phase of map production without regard for the rest of the process. It is little use to eradicate centimeter errors with laser survey equipment, only to introduce meters of imprecision with inexact linework and unstable paper (see below).

The error in surveying and geodesy applies to the points measured and, by extension, to all information interpolated between the known points. For simple point objects, a point error model is adequate, but most maps consist of more complex features. The epsilon model provides a method to apportion the uncertainty in surveying to all features on the map.

Interpretation

Studies of error in locating ground position are typically concerned with "well defined" points (Thompson, 1960), but the bulk of cartographic detail does not consist of distinct points. Most maps consist of lines which might represent entities directly, as in the case of railroads or faults, or the lines might serve as boundaries of areal units. A boundary line represents a change from one areal feature to another. The spatial accuracy of the line is dependent on more than the technology of surveying, because the line has to be perceived in the first place. Locating a line involves discrimination of the adjacent features; the difficulty of doing this depends on the particular case. Property boundaries have a very precise meaning and usually can be located quite exactly, but the border between forest types might not be a line at all, just a fuzzy zone of interpenetration and transition.

Discrimination at borders is only one possible source of interpretation error. For instance, a completely erroneous classification could be recorded. However, such an error is not really a spatial problem. Standard misclassification analysis, used in medical diagnosis and other fields (Fleiss, 1973), provides useful tools to deal with such errors. These methods differ from the deductive approach of this paper, because they rely on some form of resurvey, although the findings of the two approaches can be integrated.

Scale

Scale is more than the mathematical relationship between the earth and the map; scale implies a specific decision about generalization and aggregation. Error in procedures such as

surveying are adjusted according to the scale of output required. Scale has a similar impact on interpretation. MacDougall (1975) attributes most error to lack of "purity", but strict classification accuracy is inevitably sacrificed to scale. For example, standard practice in land use mapping lumps scattered corner stores into residential neighborhoods (Anderson and others, 1976).

In spite of the recognition of scale in most systems of classification, it is normal to assess accuracy by point sampling (Fitzpatrick-Lins, 1978). Differences in the land use detected by this procedure might be misleading. Scale-specific effects might require that a point be swallowed up in a larger zone, or the sampling point could lie near an imprecise boundary. It would be difficult to disentangle these error effects to allow sampling at different scales.

Conversion to Map Space

Projections are only a small issue in transfering measurements onto maps, although they receive generous attention in cartography. In general terms, drawing a map involves two physical objects, "pen" and "paper", controlled by a person. These three components introduce their own forms of error.

Perhaps the simplest effect is created by the "pen". Whatever writing implement might be used (pencil, scribing tool, or ink pen), a line is represented by a narrow region of more or less constant width. With the highest state of the art, line widths will be quite uniform and as narrow as .1 mm, but many maps use much wider lines. The mark made by a pen, incidentally, provides a near perfect rendition of the epsilon model with epsilon set at one half the pen width.

The error effects of drafting are not confined to the physical nature of the pen. A human operator wields the pen and attempts to record spatial information. The decisions made will generate error, depending on the nature of the information available and the technology used. A common problem, though not the only one, is registration where visual clues are used to align two images. When misregistration is less than the line width there is no easily detected visual clue. However, when the error is larger, it produces a sliver zone. The outside edge of the traced line may be up to one and one half line widths away from the center of the source line before any sliver appears.

The problem of line following and misregistration will appear in any manual phase in the production process. Degradation of map accuracy caused by graphic revisions such as retracing has been documented in a number of cases (Libault, 1961, p. 68; Harley, 1975, p. 165). Modern digital production should remove redrafting from the production flow, but it will not remove the errors already introduced in existing map series.

Traditionally, maps reside on paper. This material is cheap, flexible and durable, but it is also dimensionally unstable. Humidity can change spatial measures substantially and often permanently. A one percent change in length is demonstrated to be possible (Libault, 1961; Braund, 1980). In evaluating the whole map production process, errors of this magnitude can dominate all other effects. Although high quality map production now requires stable base material, paper maps are an unavoidable legacy of the historical record. Any data source on paper should be treated very cautiously.

Digital Handling

Production methods have evolved from the traditional hand methods to considerable reliance on computer processing of cartographic information. Automation will reduce error in so far as it avoids redrafting and similar degradation of the information. Digital systems have an aura of accuracy, but they do have inherent error effects which cannot be avoided.

The simplest problem is that computers use finite precision for storage and calculations. The effects of roundoff introduce a uniform distribution around the coordinates stored. If the points are reasonably close, these zones merge to approximate the epsilon model. Rounding also introduces some possibility for error in calculations, but this requires careful analysis of the specific program.

The largest potential errors in digital map processing occur during digitizing, whether performed by manual devices or by automated scanning. In both cases, hardware characteristics will introduce some amount of error. Manufacturers often quote the resolution of the device (the smallest measurement produced), creating the impression that this describes its accuracy (the expected error of measurements). However, a .001 inch resolution is usually coupled with .005 inch "repeatability". Other errors can arise from the width of the spot on the cursor of a manual device.

Manual digitizing resembles drafting, so similar errors should be expected. The lack of direct visual feedback may degrade digitizer results compared to drafting. Some tentative results on the magnitude of digitizer error were obtained in an experiment by Thorpe (1981, personal communication). He measured the deviation between a known set of contours and the results of manual and laser line follower digitization. The laser device was able to keep virtually all of its measurements in a band two line widths across; the average deviation was about a third of the line width. The manual operator was five times less accurate. Additional, carefully designed studies of digitizer error are needed to establish the reliability of digital data.

Combining Effects

Each error effect relevant for a particular map can be treated as a random variable, perturbing the true line to obtain the observed line. A crucial part of this analysis hinges on combining these separate error effects. Each error effect tends to occur as the spatial information is

passed from phase to phase in the sequential process of map production. For example, any surveying error is incorporated in the data at that stage and is then treated as correct. This situation suggests that the processes can be treated as independent. In this case, sufficient results are obtained by adding the variances of the distributions (Blumenstock, 1953; Chrisman, 1982). This is a result of the calculus of probability functions, known in surveying as the Law of Propagation of Errors. An epsilon band as wide as the average deviation has the property that deviations outside are exactly balanced by those inside. Thus the area of the band is a reasonable measure of the uncertainty of area measurement.

MEASURING EPSILON BANDS

The epsilon model of map error remains a theoretical curiosity without a practical method of measurement. The epsilon band consists of all points within a distance epsilon from a line. As demonstrated above, this locus forms a band around an isolated line. However, this definition ignores one of the most important features of most maps; the lines are connected, not isolated. The error model is designed to estimate the amount of area subject to fluctuations, and no area should be counted twice.

Perimeter is an imprecise estimator of the epsilon band, but it is closely related, and adequate for very small epsilon. The most obvious modification of perimeter measurement is at the angle formed by each adjacent pair of lines. A circular section occurs on the convex side of the angle, where perimeter would undercount, while a pair of triangles on the concave side represent overcount. The triangles are bound to be larger than the circles, and hence the net effect is to subtract from the perimeter value. Since both the circles and triangles increase with epsilon squared, while perimeter effects only increase linearly with epsilon, the net effect is increasingly important with larger epsilons.

It is possible to detect other cases along a line where bands interact and overcounting occurs. Cases which are sufficiently common should be incorporated into routine epsilon measurement. Of course, a local approach, by definition, does not try to examine all possibilities. However, the simple case dominates in normal circumstances [Chrisman (1982) measured other effects and showed that they were trivial].

AN EXAMPLE OF EPSILON ERROR MEASUREMENT

In order to provide a concrete test case, the epsilon measurement program was applied to data obtained from the GIRAS digital files - the U.S. Geological Survey's Land Use/Land Cover series (Mitchell and others, 1977). Of six test cases performed (Chrisman, 1982), a single example is presented here. A rectangle of approximately 100,000 hectares around the city of Pittsburgh was extracted from the Pittsburgh sheet.

Setting Epsilon Width

Examination of the map production processes used by the Geological Survey yielded three error effects that could be quantified (Loelkes, 1977). The line width amounts to 25 meters on the ground. Thus, line width drafting error might have an average deviation of 12.5 meters under the very best circumstances. Digitizing was performed by the same hardware tested by Thorpe, giving another deviation of 8.3 meters. Roundoff contributed 2.9 meters average deviation. These error effects combine to an average deviation of 15.2 meters using the formula for adding random variables discussed above. It was decided that an epsilon band of 20 meters would be quite conservative, considering that interpretation error was not estimated and effects such as registration could not be judged.

Measurement Results

The total area measured in the epsilon bands of 20 meters amounts to 7191 hectares, or about 7 percent of the total. Higher figures are obtained in more complex areas; in other rectangles studied total error reached 10.9 percent of total map area (see Chrisman, 1982). Overall these figures come within the 85 percent classification accuracy goals stated by USGS, but the error model only accounts for boundary errors, not gross classification error.

Table 1 (see separate page) contains the results provided by the measurement program applied to the Pittsburgh area. The categories of the rows and columns are the Level II land use codes defined by the U.S.G.S. (Anderson and others, 1976). In each cell of the matrix the top figure is in hectares, while the lower figure is a row percentage. The rows of this square matrix represent the land use as mapped; the sum of the row is the total area on the map. The columns represent the same land use categories, but as possible recepients of error effects. The diagonal contains the 93 percent of the map which is not affected by the epsilon bands.

Bounds on Area Measurement

In many fields, such as engineering or physics, measurements are usually presented with the best estimate "plus or minus" one standard deviation. This allows an open statement of reliability that should be provided with any scientific measurement.

The figures in Table 1 provide the raw material for placing bounds on area measurements. For a given category, the sum of its row (disregarding the diagonal) is the amount it might lose, while the column sum is the amount it might gain. These two figures provide separate "plus" and "minus" estimates, which are slightly different due to shape effects. The asymmetry of the error bands provides valuable information reflecting the spatial structure of the map. The separate bounds could be stated separately (e.g. 41 : 34284 hectares + 1977 - 2016), but I think that would be confusing. A more readable presentation uses the sign "<" (less than) to present a lower bound and an upper bound on either side of the measured area: lower bound < area < upper bound (see Table 2). These bounds should be

Table 1: Cross-tabulation of error measurements (Pittsburgh rectangle)

by Anderson code, each cell: area in hectares on top, row percent on bottom, diagonal:no error

	11	12	13	14	15	16	17	21	41	42	43	51	53	74	75	76
11	37230	210	197	80	5	13	203	117	1041	27	91	64	7	0	58	30
	94.6	0.5	0.5	0.2	0.0	0.0	0.5	0.3	2.6	0.1	0.2	0.2	0.0	0.0	0.1	0.1
12	196	2167	23	14	1	1	18	12	65	-	-	9	-	-	8	2
	7.8	86.1	0.9	0.5	0.1	0.0	0.7	0.5	2.6			0.4			0.3	0.1
13	192	23	2599	21	-	4	5	6	68	-	5	131	-	-	9	6
	6.3	0.8	84.6	0.7		0.1	0.2	0.2	2.2		0.2	4.3			0.3	0.2
14	77	14	21	1105	4	-	6	10	97	-	3	14	-	-	13	-
	5.6	1.0	1.5	81.0	0.3		0.4	0.7	7.1		0.2	1.0			1.0	
15	5	1	-	4	154	-	-	-	2	-	-	2	-	-	3	2
	2.8	0.8		2.3	88.6				1.2			1.2			1.9	1.2
16	13	1	4	-	-	191	-	2	14	-	1	-	-	-	-	-
	5.6	0.3	1.7			84.6		1.0	6.3		0.5					
17	192	18	5	6	-	-	2355	14	78	1	3	2	-	-	5	2
	7.2	0.7	0.2	0.2			87.8	0.5	2.9	0.0	0.1	0.1			0.2	0.1
21	120	12	7	10	-	2	15	8432	363	-	30	-	-	-	34	3
	1.3	0.1	0.1	0.1		0.0	0.2	93.4	4.0		0.3				0.4	0.0
41	1054	70	71	98	2	15	82	368	32268	-	-	68	4	-	136	51
	3.1	0.2	0.2	0.3	0.0	0.0	0.2	1.1	94.1			0.2	0.0		0.4	0.1
42	27	-	-	-	-	-	1	-	-	390	-	-	-	-	-	-
	6.5						0.3			93.3						
43	92	-	5	3	-	1	3	31	-	-	2432	6	-	3	18	-
	3.6		0.2	0.1		0.0	0.1	1.2			93.7	0.2		0.1	0.7	
51	64	9	132	14	2	-	2	-	67	-	6	1898	-	2	3	1
	2.9	0.4	6.0	0.6	0.1		0.1		3.1		0.3	86.3		0.1	0.2	0.0
53	6	-	-	-	-	-	-	-	3	-	-	-	30	-	-	1
	14.5								8.5				75.7			1.3
74	0	-	-	-	-	-	-	-	-	-	3	2	-	41	-	-
	0.4										7.3	3.9		88.3		
75	57	9	10	13	3	-	5	32	128	-	16	3	-	-	1824	1
	2.7	0.4	0.5	0.6	0.2		0.3	1.5	6.1		0.8	0.2			86.8	0.1
76	28	2	6	-	2	-	2	3	48	-	-	1	1	-	1	531
	4.5	0.4	1.0		0.3		0.3	0.5	7.7			0.2	0.1		0.2	84.9

interpreted as standard deviations, not absolute limits.

Table 2: Probable bounds on area measurements

11	37230	<	39374.1	<	41498
12	2167	<	2516.0	<	2885
13	2599	<	3070.9	<	3552
14	1105	<	1363.7	<	1626
15	154	<	174.2	<	194
16	191	<	226.2	<	262
17	2355	<	2681.9	<	3024
21	8432	<	9028.2	<	9622
41	32268	<	34284.7	<	36261
42	390	<	418.4	<	447
43	2432	<	2594.6	<	2753
51	1898	<	2199.0	<	2500
53	30	<	40.1	<	51
74	41	<	46.2	<	52
75	1824	<	2101.7	<	2391
76	531	<	625.8	<	724

The figures in Table 2 provide a summary of Table 1 while losing some of the structure of spatial interdependence. Area estimates derived from fallible sources, such as maps, should be presented with bounds such as these so that a user does not take a measurement too literally.

CONCLUSIONS

The epsilon model of information content suits the purposes of understanding the variability of cartographic lines. By deduction, the uncertainties introduced in the steps of cartographic production can be estimated. This paper set forth a simple triginometric procedure to measure the area of the epsilon bands. The resulting error estimates provide a more realistic description of area measurements obtained from maps. This paper presented the theory and practicalities of error measurement, but there is a need for further inductive studies to calibrate the model. In addition, this work should lead to new statistical procedures for treating spatial information which are sensitive to the structure of spatial interdependence described by the error procedures.

ACKNOWLEDGEMENTS

The research for this paper was supported in part by National Science Foundation Grant SES 7909370 and the Laboratory for Computer Graphics at Harvard University.

REFERENCES

Anderson, J.R., Hardy, E.E., Roach, J.T. and Witmer, R.E. 1976, A land use and land cover classification system for use with remote sensor data: Professional Paper 964, Washington DC, U.S. Geological Survey

Blumenstock, D.I. 1953, The reliability factor in the drawing of isarithms: Annals of the Association of American Geographers, Vol.43,pp.289-304

Braund, M. 1980, An analysis of the effects of temperature and humidity on a variety of drawing media: Bulletin of the Society of University Cartographers, Vol.14,pp.25-35; reprinted from Journal of the Association of Surveyors of Papua New Guinea

Chrisman, N.R. 1982, Methods of spatial analysis based on error in categorical maps, unpublished Ph.D. thesis, University of Bristol

Corbett, J.P. 1975, Topological principles in cartography: Proceedings AUTO-CARTO II, pp.61-65

Douglas, D. and Peucker, T.K. 1973, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature: Canadian Cartographer, Vol.10,pp.112-122

Fleiss, J.L. 1973, Statistical methods for rates and proportions, New York, John Wiley.

Harley, J.B. 1975, Ordnance Survey maps: a descriptive manual, Southampton, Ordnance Survey

Libault, A. 1961, Les mesures sur les cartes et leur incertitude, Paris, Presses Universitaires

Loelkes, G.L., 1977, Specifications for Land Use and Land Cover and Associated Maps: Open File Number 77-555, Reston VA, Geological Survey

Mitchell, W.B., Guptill, S.C., Anderson, K.E., Fegeas, R.G. and Hallam, C.A. 1977, GIRAS, a geographic information retrieval and analysis system for handling land use and land cover data: U.S. Geological Survey Professional Paper, 1059, Reston VA, U.S. Geological Survey

Perkal, J. 1956, On epsilon length: Bulletin de l'Academie Polonaise des Sciences, Vol.4,pp.399-403

Perkal, J. 1966, On the length of empirical curves: Discussion paper 10, Ann Arbor MI, Michigan Inter-University Community of Mathematical Geographers.

Peucker, T.K. 1976, A theory of the cartographic line: International Yearbook of Cartography, Vol.16,pp.134-143

Thompson, M.M. 1960, A current view of the National Map Accuracy Standards: Surveying and Mapping, Vol.16,pp.449-457

Thorpe, L.W. 1981, personal communication.