ERROR IN CATEGORICAL MAPS: TESTING VERSUS SIMULATION

Nicholas R. Chrisman
Department of Geography DP 10
University of Washington, Seattle WA 98195 USA
CHRISMAN@MAX.bitnet or CHRISMAN@MAX.ACS.WASHINGTON.EDU

ABSTRACT

Understanding error in maps requires a combination of theory (new models) and practice (understanding how error can be measured in real applications). While other research emphasizes mathematical models to simulate error, a practical test provides a more useful judge of cartographic data quality. A comprehensive test, overlaying two categorical maps intended to be the same, can provide an estimate of separate components of error including positional and attribute accuracy along with scale effects.

MAP ERROR: A FOCUS OF EFFORT

A few years ago, cartographic data quality and map error could be called a neglected topic (e.g. Chrisman, 1983). Recent developments have placed substantial attention on data quality, but most activity has focused on recognition that there is a problem. A number of components are required for overall improvement in the treatment of cartographic data quality. At the operational level, practitioners need tools to reduce error but tools require diagnostic tests. The tests, in turn, will reflect some model of error. Some such models can be imported from other sciences, but certain forms of cartographic information will require new models. This paper outlines a procedure to test one common form of cartographic data. The result is not a full-fledged "model" of error; it does provide a taxonomy of error which can lead to a model of error.

Fundamental differences over error

A major impediment to progress has been confusion over the understanding of cartographic error. The profession seems split into a number of incompatible schools of thought. A full treatment of intellectual history would have to begin with a review of the many disciplines which combine to contribute to modern GIS developments, but such depth would occupy the full length of this paper. Instead, this paper will concentrate on providing an alternative to one dominant approach to cartographic error models.

The Simulation School. One group of researchers (exemplified by Goodchild and Dubuc 1987, but including others as well) seeks to develop a procedure which can produce a "random" map. Their approach adopts common stochastic modeling methods from mathematical statistics. Such modeling can construct some numerical procedures with results that share certain measures (topology, size distribution) with actual maps. The goal of this research seems to be developed by analogue from other sciences where a generalized random model could be developed to create analytical tools for a broad class of information. Perhaps this research track will lead to a generalized model of error, but such success is bound to be far off. Constructing a simulation that produces plausible maps does not mean that real maps arise from that process or share similar mathematical properties.

Other sciences engage in stochastic modeling from a firm foundation in measuring their phenomena of interest. In particular, the definition of error in biomedical or agricultural experiments is not a matter of controversy. The bulk of mathematical statistics depends on the concept of a "population"- a large or infinite pool of individual cases that will behave in essentially identical fashion. Error models can predict the probability of obtaining certain results from samples drawn from the population. For many sciences, the case and population paradigm summarizes potential error. By contrast, the mapping sciences have not developed a comprehensive taxonomy of what errors occur and what processes control the amount of particular kinds of error. The use of cartographic information in geographical analysis involves many properties (particularly colocation and other geometric properties) not considered in standard statistical treatment. In my opinion, it is premature to develop stochastic models for a field without a clear understanding of the fundamentals.

An Alternate Philosophy. Philosophically, stochastic modeling fits into an idealist view where the pure nature of things is clouded by an imperfect world. In other terms, the abstract model is more perfect and correct than the phenomenon it represents. While this view of the world has been held by some prominent philosophers for millenia, it is not the only possible approach. My philosophic position can be summarized by a few principles. I do not presuppose some ideal world which is more pure and correct. Observation, measurement, experiment and experience provide access to an inscrutable world. As humans we develop concepts, theories, and languages to organize our knowledge, but these human constructions are mainly useful in making further predictions of the actual operation of the world. Abstract, self-contained systems like mathematics or programming languages can be absorbing, but they prove their utility by allowing humans to manipulate the real world. While this philosophy (perhaps formally termed pragmatic realism) may sound non-controversial or banal, it leads to a different approach to cartographic error. I believe that the ultimate arbiter of cartographic error is the real world, not a mathematical formulation. I define error as the deviation of our representation from the actual state of affairs. This deviation will vary from place to place and from time to time and from technology to technology. I can only use a mathematical model to predict this error when I organize the evidence to generalize from the specific case.

Just as there are inductive and deductive approaches to scientific method, there are distinct approaches to cartographic error. In some earlier papers (Chrisman, 1982b), I adopted a rather deductive strategy of assigning error to each step performed from source material to final digital product. Eventually this approach must be adopted for routine estimation of data quality, but it is not the appropriate strategy for developing a taxonomy of error behavior. This paper is based on the use of testing and empirical evidence to help structure a theory of error.

Geometry and Attributes

Perhaps the most commonplace distinction in cartography and GIS contrasts treatment of strictly spatial data from the rest of the aspatial context. The spatial elements are best termed "geometry", a term which includes both metrical position and topological components, though other terms are in common use. There is also diversity in terminology for the aspatial "thematic" components. This paper will use the term "attribute", although it often covers both geometric and thematic components.

It is a common trap in cartography and other sciences to seek finer and finer nuances of terminology as a substitute for theoretical insight. Whether geometry is simply another attribute or must be treated differently is one of the major issues dividing current GIS implementations. To date the debate between a dual "georelational" approach (Morehouse, 1985) and a unitary representation (e.g. Charlwood and others, 1988) has focused on efficient use of computers. Although such efficiency has been a primary measure for GIS, these considerations have not included treatment of data quality. Some research on error models follows the traditional division of attribute from geometry while others seek unitary models. In the simulation school, the fundamental tools deal with continuous surfaces which place the thematic component in a common metric with the horizontal position. With a surface model, there are many mathematical operations which are quite valid, but the behavior of surfaces is not the only problem confronted by cartographers. This paper will seek to show that a dual approach is required for some forms of map data.

TYPES OF CATEGORICAL DATA

Eventually, there is a need for a model of error treating all forms of spatial information. Much of the work in the mapping sciences has treated the positional accuracy of "well-defined points". Such objects can be treated separately without worrying about their context. The current interest in cartographic "feature" data adheres to this simple world where objects are surrounded by the void. It may be possible to construct an error model for feature data from more traditional mathematical statistics because features do not involve topological properties and other two-dimensional characteristics. However, I believe that feature data is often selected from a richer view of spatial relationships.

An important property of spatial information is exhaustiveness. Most analytical cartography has focused on surfaces, exhaustive fields of continuous varying attributes, but this form of data, while mathematically tractable, does not cover all of the problems faced in GIS application. The most complex problems arise when the thematic information - the attributes - are measured on a categorical scale of measurement, either nominal or ordinal. This paper is primarily concerned with one kind of two-dimensional distribution, termed a *categorical coverage* (Chrisman, 1982a). A categorical coverage is a specific type of polygon map used quite frequently for GIS applications.

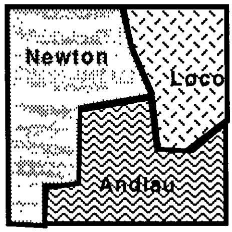
It is important to distinguish this form of polygon data from spatial collection units (Figure 1).

Figure 1
Arbitrary collection units
(one type of categorical map)

Named zones come first.

Examples:

political divisions
census tracts
. . .



Followed by attribute measurement
(nominal or ratio)

Newton	Elected mayor
Andiau	Town meeting
Loco	Monarchy
•	•
•	•
•	•

The important consideration is which component, the spatial description or the attribute, takes logical precedence (Sinton, 1978). In the pure case (administrative units such as municipalities), the positional description of the object precedes any attributes assigned. These maps are *choropleth* maps in the purest sense, because the places exist, then they are filled. [Choropleth has now come to refer to categorical maps derived from classed continuous distributions, but that does not alter the etymology.]

Many of the users of GIS software do not rely upon collection unit sources. The layers fed into a GIS are more likely to be soil maps, vegetation maps, ownership parcels, and many more. Although the distinction is not absolute, these maps derive from a different approach (Figure 2).

Figure 2
Categorical Coverages

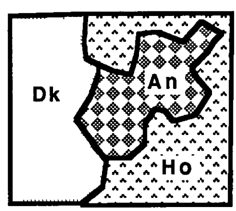
(a type of categorical map distinct from collection zones)

Taxonomy comes first.
A priori, exhaustive

An	Antigo Silt Loam
Bb	Batavia Silt Loam
Dk	Dickinson Sand
Ho	Houghton Muck
•	
•	
•	

Lines adjusted to represent the edge of the category

Followed by location in space



Both forms of data (Figures 1 & 2) may be displayed as choropleth maps, but similarity of graphic display obscures fundamental differences. In Figure 2, some system of classification (the soil taxonomy, the vegetation classes, and even the list of taxable parcels) logically precedes the map. The map results from assigning each portion of the area into one class or another. Issues of positional accuracy, scale and other cartographic concerns become much more prominent than they are in the collection zone case. The model of error implicit in collection units (spatial autocorrelation) relies on an underlying continuous distribution, aggregated into discrete and arbitrary spatial units. A model of error for categorical coverages reverses the logic. Spatial units are adjusted on a continuous space to reflect the categorical distinctions.

A FRAMEWORK FOR MEASURING ERROR

Before a complete stochastic model can be developed, the first step is to define the error to be modelled. The various disciplines involved in mapping have used widely varying concepts of error, and each should make a contribution to a comprehensive model. The fundamental issue in statistics is understanding deviations. The deviations possible in a categorical coverage involve diverse components. In particular, there are positional (geometric) issues and attribute issues. The concept of deviation used for these two are usually quite different, but, in a categorical coverage, the various error components interact. Goodchild and Dubuc (1987) reject the separation of geometry and attributes, but there are strong suggestions that parallel treatment is useful. This section describes a mechanism to deconvolve spatial error into identifiable processes, each with distinct mathematical treatment.

It is relatively easy to catalogue all of the steps used to create spatial information. Each of these steps no doubt introduces different types and amounts of error in the resultant products. But a complex model of this kind (essentially the proposal of Chrisman, 1982a) is quite difficult to verify. The amount of error can be best ascertained by a process of testing. Tests have inherent limits in their ability to distinguish errors from different sources. The existing practice of mapping sciences include a very few established testing procedures. Taken together, these tests do provide some sort of coverage for the range of problems included in the proposed US National Standard. Some of the most recent tests, like the tests of topological integrity (White, 1980), have developed from the introduction of computing to mapping, but most are longer established. The positioning sciences (geodesy, surveying, photogrammetry) have tests of positional accuracy based either on repeated measurements (internal evidence) or on tests against an independent source of higher accuracy. In both cases the tests treat "well-defined points", cartographic features taken in isolation. Photointerpretation and remote sensing use point sampling to test classification accuracy, following some relatively standard procedures to estimate proportions of a categorical variable (Rosenfield and Melley, 1980). The spatial sampling techniques outlined by Berry and Baker (1968) still provide the spatial logic for these tests.

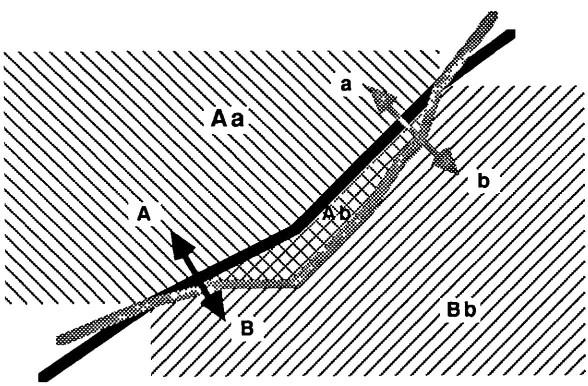
The current range of tests are each designed to treat a distinct component of overall quality. Thus they can easily be fooled by error of the other components. For instance, the emphasis on well-defined points in positional tests is to reduce the impact of classification error. While it may be correct to isolate some components for some purposes, there is a need for a comprehensive test, particularly for exhaustive categorical coverages. The common point sampling approach to classification accuracy can fial to distinguish between errors in positional and attribute components.

A comprehensive test compares complete maps, not just sampled locations. Two categorical coverages purporting to map the same phenomenon are overlaid comprehensively, and the results form a test of accuracy. This test has been applied in some isolated circumstances (for instance, Ventura and others, 1986), and it has been accepted as an alternative to point sampling in the proposed US National Standard.

If one source is assumed "correct", it is a test of the other, but it could also be a test of repeatability. As in many statistical applications, a test pairs measurements. Unlike the standard applications, a spatial test pairs every point on the map by location on the ground. Such an arrangement, with an infinite number of points, requires a different error model than a "case" oriented approach. This framework is described incrementally, starting from some simple cases, then providing more complexity.

The most common form of error in overlaid maps is called a "sliver". As demonstrated in Figure 3, a simple sliver occurs when a boundary between two categories is represented slightly differently in the two source maps for the overlay. A small, unintended zone is created. Goodchild (1978) reports that some systems become clogged with the spurious entities that provide evidence of autocorrelation at different levels. These reports are a part of the unwritten lore of GIS, because most agencies are unlikely to report on failures. Some algorithms for overlay include a filter to remove the smallest of these, up to the level a user is willing to tolerate (Dougenik, 1980).

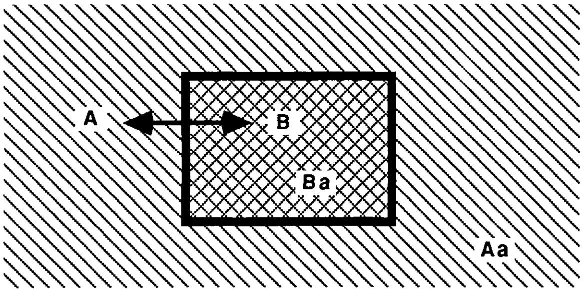
Figure 3
Slivers: the classic form of overlay error



A simple illustration: 2 categories (A,B) 2 sources (upper/lower)

Although sliver error is the most frequently mentioned, an overlay test can discover other forms of error. To follow the example described above (comparison of two maps assumed to be the same), it would be possible to have a feature on one map source which is completely missing on the other, as shown in Figure 4. While the sliver error seems to arise from *positional* error, a missing polygon is caused by classification error. Unlike the rudimentary "feature" approach, a misclassification in a coverage assigns the area to some other category. Taxonomic similarity of the two categories could be modelled in some continuous phase space - as proposed by Goodchild and Dubuc (1987), or otherwise.

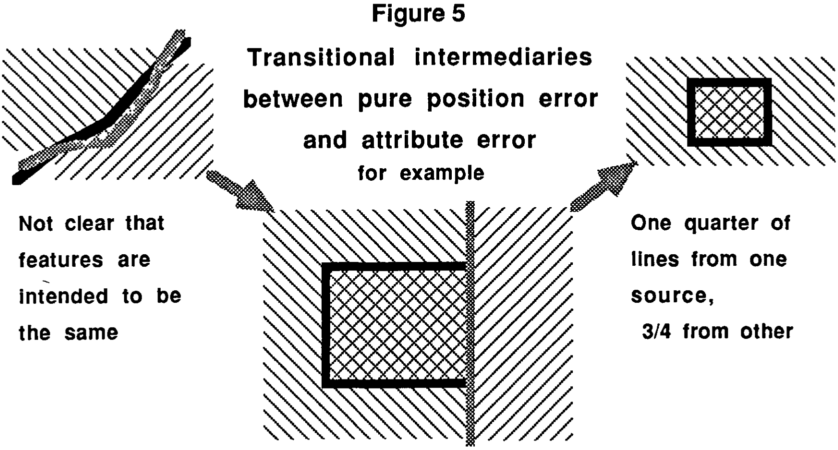
Figure 4
Another case of overlay error



As extremes, the positional sliver and the attribute classification error seem perfectly distinct. But the two are quite difficult to disentangle in practice. For instance, a sliver error might arise from an interaction of positional error and difficulty in discriminating the classifications (more of an attribute problem than an error in positioning technology). Chrisman (1982a) proposes a division of classification error for categorical coverages into components of *discrimination* (essentially the sliver effects) and *identification* (essentially the subject of Figure 4). This test seeks to build this distinction into a larger framework.

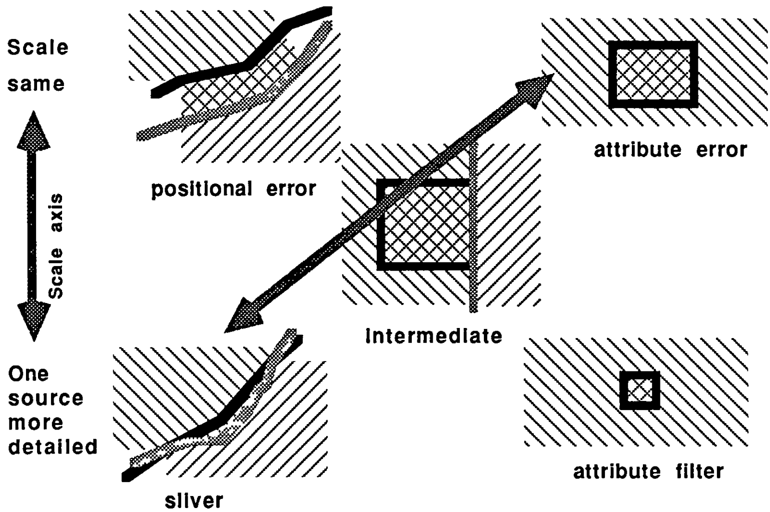
As an additional complication, not all error falls perfectly into the two cases presented in Figures 3 and 4. The sliver involves roughly the same contribution of linework from each source, while the classification error has all the linework from one source. As Figure 5 shows, there is a continuum possible between the two extremes which might be hard to classify. While it is easy to develop anecdotes about this kind of error, there is no workable theory in common use.

Figure 5



The previous argument deals with the existence of both positional and attribute error, but ignored the issue of scale. In spite of the power of modern GIS software, the basic information is still strongly dependent on scale. Positional accuracy of lines is expected to be linked to scale, but the amount is rarely specified or measured. Even more so, attribute error is linked to scale. At some scales, features like farmsteads are consciously removed from land use maps. Scale involves a distortion of the information, but a distortion that is tolerated and expected. To develop a framework, Figure 6 shows how positional and attribute error might interact schematically with scale issues.

Figure 6
Scale adds a dimension to the transition
(previous diagram now the diagonal)



The framework presented above is more than a diagram. It provides the basis to construct a mathematical model where the total error is decomposed into a set of stochastic processes operating simultaneously. The stochastic process for boundary error will have to reflect the geometric impact of cartographic representation and processing, while the attribute error will have to reflect taxonomic similarity of classes. For the technical "process" errors that simply degrade the positional accuracy, the epsilon model (Perkal, 1966; Chrisman, 1982a; 1982b) may provide a useful start. For errors in identification or misclassification, some modification of Goodchild's phase spaces may be developed, depending on the basic science for the particular information. To further complicate affairs, these two processes will operate inside scale-dependent rules that can be modelled as filters and other constraints. It is extremely unlikely that we can expect a single overarching scheme to treat error in geographic information, but the constraints of testing must influence our ability to discern and differentiate such errors.

The framework developed above may explain the results obtained from empirical accuracy experiments. The concept of an exhaustive test through polygon overlay has been accepted as a component of the US proposed national standard, but few tests have been performed using this approach (for example, Ventura and others, 1986). Empirical results measure the total error from all processes, and there is no guaranteed mechanism to deconvolve them. Each of the individual components above will be easier to model in isolation, then the error components can be combined.

CONCLUSIONS

Considering the public and private investment in geographic information systems, additional research on the error of overlaid maps is required. This paper sketches a preliminary taxonomy of error that can be used as the basis for research. With substantial development, a new set of analytical procedures may be developed, perhaps even a "geographical analysis of variance" (Warntz, 1966).

ACKNOWLEDGEMENTS

The research in this paper was supported in part by National Science Foundation Grant SES 87-22084.

REFERENCES

Berry, B.J.L. and Baker, A.M. 1968: Geographic sampling. In Berry and Marble (ed.) *Spatial Analysis,* 91-100.

Charlwood, G., Moon, G. and Tulip, J. 1987: Developing a DBMS for geographic information: a review, *Proc. AUTO-CARTO 8,* 302-315.

Chrisman, N.R. 1982a: Methods of spatial analysis based on error in categorical maps. PhD thesis, U. of Bristol.

Chrisman, N.R. 1982b: A theory of cartographic error and its measurement in digital data bases. *Proceedings AUTO-CARTO 5,* 159–168.

Chrisman, N.R. 1982c: Beyond accuracy assessment: correction of misclassification. *Proc. International Society of Photogrammetry and Remote Sensing Commission IV, 24-IV,* 123–132.

Chrisman, N.R. 1983: The role of quality information in the long-term functioning of a GIS. *Proceedings AUTO-CARTO 6, 2,* 303-321.

Chrisman, N., Mezera, D., Moyer, D., Niemann, B., Vonderohe, A. 1984: Modernization of routine land records in Dane County, Wisconsin: implications to rural landscape assessment and planning, *URISA Professional Paper 84-1.*

Dougenik, J.A. 1980: WHIRLPOOL: a processor for polygon coverage data. *Proc. AUTO-CARTO IV,*304–311.

Goodchild, M. 1978: Statistical aspects of the polygon overlay problem, in vol. 6 G. Dutton, ed., *Harvard Papers on Geographic Information Systems,* Addison Wesley.

Goodchild, M. and Dubuc, O. 1987: A model of error for choropleth maps with applications to geographic information systems. *Proc. AUTO-CARTO 8,* 165-174.

Harding, E.F. and Kendall, D.G. 1974: *Stochastic Geometry,* John Wiley.

Morehouse, S. 1985: Georelational data structure for GIS, *Proc. AUTO-CARTO 7.*

Perkal, J. 1966: On the length of empirical curves. *Discussion Paper 10,* Michigan InterUniversity Community of Mathematical Geographers.

Peucker, T.K. 1976: A theory of the cartographic line. *International Yearbook of Cartography, 16,* 134–143.

Rosenfield, G. and Melley, M. 1980: Applications of statistics to thematic mapping. *Photogrammetric Engineering and Remote Sensing, 46,* 1287–1294.

Sinton, D. 1978: The inherent structure of information as a constraint to analysis: mapped thematic data as a case study , in vol. 7 G. Dutton, ed., *Harvard Papers on Geographic Information Systems,* Addison Wesley.

Ventura, S. Sullivan, J.G. and Chrisman, N. 1986: Vectorization of Landsat TM land cover classification data. *Proceedings URISA, 1,* 129-140.

Warntz, W. 1968: On the nature of maps. *Discussion Paper 12,* Michigan InterUniversity Community of Mathematical Geographers.