THE EFFECTS OF GENERALIZATION IN GEOGRAPHICAL DATA ENCODING

Michael F. Goodchild

Department of Geography
University of Western Ontario
London, Canada

Any digital map data will differ from the real world it represents, usually as the result of a number of forms of generalization. An understanding of the effects of generalization on the accuracy of digital data is essential to effective system design. This paper reviews approaches to the estimation of data errors for both raster and vector systems. In the vector case the discussion centres on the spurious polygon problem.

I. INTRODUCTION

Generalization can occur at a number of stages in the transfer of data from the real world to a digital image. In some cases it is subject to precise rules, as in the encoding of a continuous spatial variation of electromagnetic radiation by a finite number of pixels. In others the process is subjective, as for example in the smoothing of contour and boundary lines by a cartographer working from aerial photography, or by a digitizer operator following a line. It is inevitable that information will be lost in one or more of the steps between reality and encoded data.

This paper is concerned with the effects of generalization error on the subsequent use of digital map data. The most useful products of a geographical information system, or more generally any form of map data processing, are measures of some kind, such as the area of a homogeneous patch of land of certain characteristics, the length of a line or the distance between specified points. In both raster and vector encoding the accuracy of these products, with respect to the true, real world values, is a function of the generalization inherent in the capture process. The greater the generalization, the lower the accuracy. But at the same time, greater generalization usually implies a smaller volume of data, and faster processing. So an understanding of the relationship between accuracy and generalization is essential to effective system design.

The paper is divided into two sections. The first deals with raster encoding, and the effects of the generalization implicit in the choice of raster or pixel size on the products of a raster-based system. The second deals with a generalization problem normally associated with vector data. In the latter case generalization is usually considered to be a function of the density of sample points along each line.

The accuracy problem would be simple if measurements from digital data could be checked directly against their true, real world values. But this is not normally possible. In general accuracy must be predicted from the digital data alone, by making assumptions about the true data. Clearly accuracy estimates will be most useful when based on the least restrictive and most realistic assumptions.

II. RASTER ACCURACY

There are two significant estimation problems in the raster context:

i) Area: what are the confidence limits on an estimate of the area of a patch of land with a particular set of characteristics, when the patch is represented in raster form?

ii) Point: what is the probability of error in estimating the characteristics of a point from a raster?

A. Area Estimation

The first problem has received the widest recognition because of its relationship to a standard method of manual measurement of area used in forestry and photogrammetry and known as dot planimetry. A transparent sheet covered with a square array of dots is laid over the patch to be measured, and the number of dots lying within the patch is used as the basis of an area estimate.

A simple error model can be constructed by assuming that each dot has an equal and independent probability of lying within the patch. The dot count would then have a binomial distribution. For a constant area and variable dot spacing, the standard error of the area estimate would depend on the dot count N to the negative one half power (1).

However the binomial model is clearly unrealistic for most geographical data, except when patches are very small and fragmented with respect to the raster cell size. The area to be measured usually consists of one or more patches each extending over a number of contiguous raster cells. All of the measurement error is then attributable to those cells on the edges of each patch, invalidating the binomial model. Some of the boundary cells will be counted as part of a patch, but contain land which is in reality not part of the patch; these contribute to an overestimation of area. Conversely there are cells which are not counted, and yet overlap the patch; these contribute to underestimation.

Frolov and Maling (2) estimated the contribution of each boundary cell by assuming that the boundary could be regarded as a straight line drawn randomly across the cell. They found the mean square area of the cutoff portion, that is, the error variance in the area estimate for each boundary cell, to be αb^4, where b is the linear dimension of the cell, assumed to be square, and $\alpha = 0.0452$. Lloyd (3) pointed to a problem with the process used by Frolov and Maling to define a random straight line, and recomputed α as 0.0609. In an independent paper Goodchild and Moy (4) obtained 0.0619 by integrating the same expression (compare Appendix 1 of (4), p. 79, with (3), p. 25).

The error variance in the estimate of area depends on the summation of the individual errors from each of the boundary cells. Let n represent the number of cells intersected by the boundary. Then assuming the contributions of each cell to be independent, the error variance will be $n\alpha b^4$, giving a standard error of $(n\alpha)^{1/2}b^2$. Bellhouse (5) has analyzed the case where the independence assumption is invalid, and see also (6).

Frolov and Maling estimated the value of n by relating it to the perimeter length of the patch. For a constant patch, we can expect that a halving of the cell dimension b will lead to a quadrupling of N, the number of cells counted, and a doubling of the length of the perimeter and n. Thus n is proportional to $N^{1/2}$, and we can write the standard error of estimate as $kN^{1/4}\alpha^{1/2}b^2$. Since the estimate of area S is Nb^2, the standard error as a percentage of the estimate will

be proportional to $N^{-3/4}$

$$\varepsilon = k\alpha^{1/2}N^{-3/4} \qquad (1)$$

Alternatively, if we regard the variable as the cell size, b, the dependence of percentage error on b is to the power 1.5. A number of studies have verified these relationships empirically (see (4), (7) and a reanalysis of Tobler's data by Goodchild (6)).

The constant of proportionality k depends on the shape of the patch. A long, thin patch has more boundary cells than a circular patch of the same area, and hence a greater standard error. Frolov and Maling (2) computed values of k for various standard shapes, again under the independent straight line assumption. Moy (7) examined the hypothesis that the population of patches present on a map sheet of a certain type could be regarded as characterized by a single value of k.

B. Point Estimation

The point estimation problem is to determine the probability that the characteristics of some chosen point, as determined from a raster representation, are correct. For a single patch the probability π_1 that a point in the patch will be misclassified as belonging to some other, neighbouring patch is given by the total of the overestimating portions of boundary cells as a percentage of patch area. Similarly the probability π_2 that a point not in the patch will be misclassified as belonging to it is derived from the total of the underestimating portions.

The relationship between these probabilities and the previous analysis is straightforward. For those boundary cells which contribute to an overestimation of area, let δ

denote the mean fraction of cell area contributed. One half of the boundary cells are expected to contribute to over-estimation, giving a total area of $n\delta b^2/2$. Thus $\pi_1 = n\delta b^2/2S$. Using the same argument for those cells contributing to under-estimation and substituting for n gives

$$\pi_1 = \pi_2 = k^2\delta b S^{-1/2}/2 = k^2\delta N^{-1/2}/2 \quad (2)$$

Misclassification probabilities are thus inversely proportional to the square root of the number of cells counted. Moy (7) found that empirical evidence supported this analysis.

III. FRACTIONAL DIMENSIONALITY

In assuming n to be proportional to $N^{1/2}$ we assume in effect that boundaries are smooth. For contorted lines, or for cells which are large compared to the patches they represent, we can expect n to increase more rapidly. There are strong parallels here with work on the relationship between the length of geographic lines and the scale at which length is measured. Richardson (8) showed that for many real lines, of which the west coast of Britain has become the classic example, the relationship between length and scale is remarkably regular. Thus if length is measured by stepping a pair of dividers along such a line, the length estimate varies with the sampling interval, in other words, the dividers setting, raised to a negative power. The power will be zero and the length estimate constant only in the case of a smooth curve.

Mandelbrot (9) has placed Richardson's empirical results within the concept of fractional dimensionality. A curve is said to be a fractal if its length varies with the sampling interval in the manner observed by Richardson, and is said to have dimension D if the relationship is of power 1-D. D is

therefore 1 for a smooth curve, and has a theoretical limit of 2 in the case of a line so contorted that it fills the area.

In the case of a raster representation of a line, the size of the raster b corresponds to a sampling interval. Thus we expect n, the number of cells intersected by the line, to depend on b^{-D}. For a constant area, n depends on the number of cells counted to the power D/2, which gives the result used previously, $n = k^2N^{1/2}$, only for a smooth curve.

Fractional dimensionality thus provides a way of characterizing the dependence ε, π_1, and π_2 on raster size when patch boundaries are appreciably wiggly. The revised estimates in terms of cell size are as follows (6)

$$\varepsilon = k\alpha^{1/2}\, b^{2-D/2}/S \quad (3)$$

$$\pi_1 = \pi_2 = k^2\delta\, b^{2-D}/S \quad (4)$$

IV. SWITZER'S ANALYSIS

Switzer (10) presented an entirely different approach to the problem of point estimation, in terms of 'mismatch area'. The area assigned to a patch by a representative raster but in reality lying outside the patch is equal to $\pi_2 S$; similarly $\pi_1 S$ is the area misrepresented as lying in the patch. In the following the patch area is referred to as black and the surrounding area as white; counted raster cells are similarly denoted as black. Write $P_2(d)$ as the probability that a randomly chosen point in a black cell, distance d from the centre of the cell, is in fact white. Then we have:

$$\pi_2 = S^{-1} \sum_{h=1}^{N} \int_{A_h} P_2(|\underset{\sim}{s} - \underset{\sim}{s}_h|)\, d\mu(\underset{\sim}{s}) \quad (5)$$

where the summation is over each of the N black cells. The integral is over the area of cell h, and $\underset{\sim}{s}_h$ denotes the location of the central point in the cell. $|\underset{\sim}{s} - \underset{\sim}{s}_h|$ is therefore the distance between the centre and every other location.

Switzer proposed that $P_2(d)$ be approximated by a Taylor expansion in its derivatives with respect to d. Taking the first two terms, and recognizing that $P_2(0) = 0$ we have

$$P_2(d) = P_2' \, d + \frac{1}{2} P_2'' \, d^2 \qquad (6)$$

where P_2' and P_2'' are the first two derivatives evaluated at $d = 0$. Thus

$$\pi_2 = [P_2' \, \rho_1 + \frac{1}{2} P_2'' \, \rho_2] \qquad (7)$$

where ρ_1 and ρ_2 are the mean and mean square distances of a random point from the cell centre, respectively.

P_2' and P_2'' can be estimated from the pattern of white and black cells in the following way. Let m(1) denote the number of times a black cell can be found with a white 4-neighbour. Similarly let m(2) denote the number of times a black cell has a white 4-neighbour two cells away. A good estimate of π_2 is then given by

$$\pi_2 = [0.60 \, m(1) - 0.11 \, m(2)]/4N \qquad (8)$$

Empirical work by both Switzer and Muller (11) has supported this analysis.

The relationship between m(1), m(2) and scale can be predicted from the fractional dimension D of the boundary of the patch as follows (6). Consider one row of cells. For a given area, the number of black cells in the row is proportional to b^{-1}, and the number with white neighbours to b^{1-D} (12). Thus if the contour has dimension 1, the number of cells with white neighbours is a constant independent of the number of black cells, but for $D > 1$ it rises with increasing resolution. Since the number of rows containing black cells is proportional to b^{-1}, the total count of black cells with white neighbours,

m(1) is proportional to b^{-D}. Similarly m(2) depends on $(2b)^{-D}$. In general Switzer's estimate can be rewritten as

$$\pi_2 = K(D) \, S^{-1} \, b^{2-D} \qquad (9)$$

K(D) is given by $K'[0.60 - 0.11 \, (0.25)^{D/2}]/4$ where K' is the constant of proportionality in the equation $m(1) = K' \, b^{2-D}$. Switzer's result is thus compatible with the previous analysis for π_2, in the dependence of π_2 both on cell size b, and on the dimensionality of the patch boundary.

The Switzer and Frolov and Maling approaches are very different in application. Switzer's estimate of mismatch area is obtained from the grid representation itself. Frolov and Maling's estimate of standard error is based on assumed characteristics of the unobservable 'true' map, since k must be obtained from the patch shape.

V. VECTOR ACCURACY AND SPURIOUS POLYGONS

The determinant of generalization in a vector representation is the density of sample points along a line. If the data has been encoded by vectorization of a raster scan, the sampling density will depend on the raster size. If the data has been digitized by line following, the sampling density may have been determined implicitly by the operator, or may have been established as a system parameter.

The effect of sampling density on the accuracy of area and point estimates is obtained by a similar argument to that in the case of rasters above, and can also be related to fractional dimensionality. In particular, the expressions for the standard error of area estimate and misclassification probability have the same dependence on the mean interval between sample points as they do on cell size in the raster case.

Particular problems arise in the vector case as a result of generalization when several coverages are overlaid. Figure 1 shows an example in which census divisions for two different time periods have been overlaid. In some cases boundaries on the two maps are roughly coincident over long stretches because both represent independent samplings of the same real geographical line.

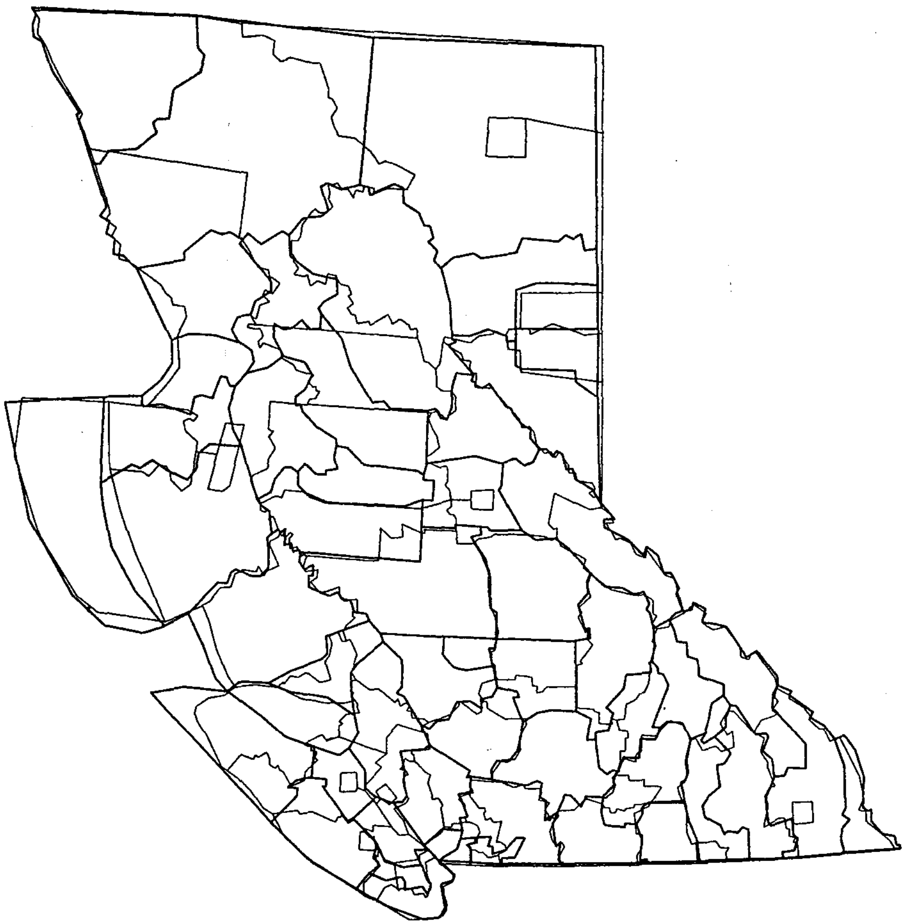

FIGURE 1. Overlay of Census Divisions from two time periods for British Columbia

TABLE I: Polygons by Area for Five CGIS Coverages and Overlays

Acreage	924C	925C	926C	927C	928C	UTES	UTE4	URIS
0-1	0	0	0	1	2	2640	27566	77346
1-5	0	165	182	131	31	2195	7521	7330
5-10	5	498	515	408	10	1421	2108	2201
10-25	1	784	775	688	38	1590	2106	2129
25-50	4	353	373	382	61	801	853	827
50-100	9	238	249	232	64	462	462	413
100-200	12	155	152	158	72	248	208	197
200-500	21	71	83	89	92	133	105	99
500-1000	9	32	31	33	56	39	34	34
1000-5000	19	25	27	21	50	27	24	22
>5000	8	6	7	6	11	2	1	1
Total	88	2327	2394	2149	487	9558	39188	90599

924C Recreation Capability
925C Land Use, 1964
926C Land Use, 1968
927C Land Use, 1973
928C Soil Capability for Agriculture

UTES 924C + 926C + 928C
UTE4 UTES + 927C
URIS UTE4 + 925C

Note: 1 acre = 0.405 hectare

The number of polygons produced in an overlay depends not so much on the number of polygons overlaid, but on the complexity of each boundary. Two polygons of n_1 and n_2 vertices can produce from 3 to $n_1n_2 + 2$ overlay polygons, classified into 16 logical combinations (13). Frequently both maps in an overlay will contain digital representations of the same real line. Since the maps have usually been subjected to independent generalization, the overlay will contain large numbers of 'spurious' polygons. Table I shows the result of overlaying five coverages in the Canada Geographic Information System; the area covered is the Ottawa-Hull National Capital Region. An enormous number of spurious polygons has been produced principally because of the persistence of boundary lines in the three land use coverages. Even though they occupy a small proportion of total map area, large numbers of spurious polygons and associated arcs create serious problems for topological data structures and should be suppressed, provided that an effective means can be found for discriminating between spurious polygons and real ones. Paradoxically, the more accurately the digitization and the greater the density of sample points, the greater the number of spurious polygons produced. Generalization of boundary lines tends to reduce the spurious polygon problem rather than increase it.

Goodchild (13) has argued that all forms of vector data capture tend to produce close to the maximum possible number of spurious polygons. For two polygons of n_1 and n_2 vertices the maximum is $2 \min(n_1,n_2) - 4$, which is substantially greater than the $2n_1n_2/(n_1 + n_2) - 3$ expected when sample points are randomly located along each digitized line. Empirical evidence tends to support this.

Size is the simplest readily available criterion for suppressing spurious polygons, but other criteria are available which might be used to improve the accuracy of deletion. All spurious polygons have only two arcs. In addition the presence of one spurious polygon makes neighbouring polygons created by the same pair of original arcs more likely to be spurious. Finally, spurious polygons tend to have a distinctive shape. These possibilities are being explored in current research.

VI. CONCLUSIONS

Accuracy has received almost no attention as a design criterion in geographical data processing, although accuracy is always lost at the capture stage, and may also be affected later during processing. There are two reasons for this. First, design has been dominated until recently by technological issues, and accuracy assumed to be outside the designer's control. Recently technical issues have tended to become less important, and at the same time applications have begun to reach the level where the demonstration of capability is no longer sufficient: users demand accuracy and performance as well. Second, there is little literature on map generalization, which has always been treated implicitly, whereas the use of computer processing requires that such issues be addressed as explicitly as possible.

The accuracy of many systems is poor, particularly in the natural resource field, to the extent that system products often have little utility. Accuracy also tends to deteriorate through time, and few systems have been designed with effective methods for updating, since use levels can rarely justify this. If geographical data processing is to advance from a research and demonstration mode to real utility, it is essential that far greater emphasis be placed on evaluating accuracy and error in both design and operation.

REFERENCES

1. Tobler, W. R., "The Accuracy of Categorical Maps." Department of Geography, University of Michigan. Cartographic Laboratory Report Number 4 (1974).
2. Frolov, Y. S. and D. H. Maling, *Cart. J. 6*, 21 (1969).
3. Lloyd, P. R., *Cart. J. 13*, 22 (1976).
4. Goodchild, M. F. and W. S. Moy, *in* "Proceedings of the Commission on Geographical Data Sensing and Processing, Moscow, 1976" (R. F. Tomlinson, ed.), International Geographical Union, Ottawa (1977).
5. Bellhouse, D. R., "Area Estimation by Point Counting Techniques." Statistics and Actuarial Science Group, University of Western Ontario (1979).
6. Goodchild, M. F., "Fractals and the Accuracy of Geographical Measures." Department of Geography, University of Western Ontario (1979).
7. Moy, W. S., "Estimation from Grid Data: the Map as a Stochastic Process." Unpublished M.A. Dissertation, Department of Geography, University of Western Ontario (1977).
8. Richardson, L. F., *Gen. Syst. Yearbook 6*, 139 (1961).
9. Mandelbrot, B. B., "Fractals: Form, Chance and Dimension." Freeman, San Francisco (1977).

10. Switzer, P., *in* "Display and Analysis of Spatial Data" (J. C. Davis and M. J. McCullagh, eds.), Wiley, London (1975).
11. Muller, J.-C., *Can. Cart. 14*, 152 (1978).
12. Mandelbrot, B. B., *J. Fluid Mech. 72*, 401 (1975).
13. Goodchild, M. F., *Harvard Papers in Geographic Information Systems 6* (1978).