generalized any form of data processing, are measures of
utility products of a geographical information system, or more
accurately, the subsequent use of digital map data. The most
error in the subsequent use of digital map data. The most
This paper is concerned with the effects of generalization

the steps between reality and encoded data.

understandable that information will be lost in one or more of
graphical, or by a digital operator following a line. It is
bound by the resolution of a particular working environment.

relevant, as for example in the smoothing of contours and
by a finite number of pixels. In others the process is sup-

of a continuous spatial variation of electromagnetic radiation
some cases it is subject to precise rules, as in the encoding
transfer of data from the real world to a digital image. In

+ INTERRODUCTION

potency problem. In the vector case the discussion concerns the spatio-
temporal data that form the data storage center. For both raster and vector
systems, the estimation of data errors is best to the system's design. The paper reviews approaches to the
evaluation of data quality, and is essential to effect-
etion. In understanding of the effects of generalization.

men can represent the result of a number of forms of
data will differ from the real world. It

london, canada
university of western ontario
department of geography

Michael P. Goodchild

IN GEOGRAPHICAL DATA ENCODING
THE EFFECTS OF GENERALIZATION
There are two significant estimation problems in the raster context.

I. RASTER ACUACY

estimates with the accuracy of the data. Clearly, accuracy can be estimated from the data, but this is not a very strong argument for the accuracy of the data. Data could be checked directly against the data.

II. TIE AND MOST REALISTIC ASSUMPTIONS.

The accuracy of data depends on the density of sample points. The accuracy of a sample set with a randomly selected set of points is not as good as the accuracy of another set of sample points. The closer the density of sample points, the higher the accuracy of the data. Data could be checked directly against the data.

The paper is divided into two sections. The first deals with system design, and the second deals with a generalization of a raster process. Two major steps in the process of rasterization are needed: the production of a raster image, and the transformation of the raster image into a vector image. The raster process is an essential part of the transformation.
The point estimation problem is to determine the proportion of the patch

\[ \frac{K}{N} \]

that a point in the patch will be measured. The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]

The proportion of the patch area that a point will be measured is

\[ \frac{K}{N} \]

The proportion of the patch area that a point will not be measured is

\[ 1 - \frac{K}{N} \]


\[ \text{Length of a smooth curve} \]

\[ L = \int_{a}^{b} \sqrt{1 + (y')^2} \, dx \]

\[ \text{Area of a smooth curve} \]

\[ A = \int_{a}^{b} y \, dx \]

\[ \text{Volume of a smooth curve} \]

\[ V = \int_{a}^{b} \pi y^2 \, dx \]

\[ \text{Surface area of a smooth curve} \]

\[ S = \int_{a}^{b} 2\pi y \sqrt{1 + (y')^2} \, dx \]

\[ \text{Volume of a smooth surface} \]

\[ V = \int_{a}^{b} \int_{c}^{d} z \, dy \, dx \]

\[ \text{Surface area of a smooth surface} \]

\[ S = \int_{a}^{b} \int_{c}^{d} \sqrt{1 + (z')^2 + (z')^2} \, dy \, dx \]

\[ \text{Length of a smooth curve} \]

\[ L = \int_{a}^{b} \sqrt{1 + (y')^2} \, dx \]

\[ \text{Area of a smooth curve} \]

\[ A = \int_{a}^{b} y \, dx \]

\[ \text{Volume of a smooth curve} \]

\[ V = \int_{a}^{b} \pi y^2 \, dx \]

\[ \text{Surface area of a smooth curve} \]

\[ S = \int_{a}^{b} 2\pi y \sqrt{1 + (y')^2} \, dx \]

\[ \text{Volume of a smooth surface} \]

\[ V = \int_{a}^{b} \int_{c}^{d} z \, dy \, dx \]

\[ \text{Surface area of a smooth surface} \]

\[ S = \int_{a}^{b} \int_{c}^{d} \sqrt{1 + (z')^2 + (z')^2} \, dy \, dx \]
Effects of Generalization in Conceptual Data Encoding

...
### Table 1: Overlays of Census Data from Two Time Periods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential</td>
<td>10.0</td>
<td>12.0</td>
<td>12.5</td>
<td>11.5</td>
<td>10.5</td>
<td>11.5</td>
</tr>
<tr>
<td>Commercial</td>
<td>5.0</td>
<td>6.0</td>
<td>6.5</td>
<td>7.0</td>
<td>7.5</td>
<td>6.5</td>
</tr>
<tr>
<td>Agriculture</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

### Notes
- 1 acre = 0.405 hectare
- 926C = 926C + 926C + 926C
- 927C = 927C + 927C
- 928C = Soils Capability for Agriculture
- 929C = Land use, 1976
- 930C = Land use, 1982
- 939C = Land use, 1992
- 940C = Land use, 1996
- Total = Total

### Figure 1: Same Real Geographic Line

The same real geographic line serves as both an independent interpretation of boundaries. In some cases, the boundaries on the two maps are roughly coincident over long periods of time. In other cases, the boundaries have been altered. Figure 1 shows an example in which census data for two different time periods are overlaid. The resulting overlay shows areas of agreement and overlap in the vector data. Additionally, some areas do not overlap at all, indicating potential problems with the data. Between these two datasets, they do not overlap in the raster.
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 \(2n_1n_2/(n_1 + n_2) - 3\) expected when sample points are randomly located along each digitized line. Empirical evidence tends to support this.

Effects of Generalization in Geographical Data Encoding

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
REFERENCES

Accuracy and error in both design and operation are essential factors for greater emphasis be placed on evaluation research and demonstration mode to real utility. It is this is geographical data processing as to advance from a

Michael F. Goodchild

Effects of Generalization in Geographical Data Encoding

Goodchild, M. F., "Representative Form of Contour and

9. Mandelbrod, B. B., "Fractals: Form, Chance and


7. "Theory of the Estimation of Land Parcel, the Map as a

6. Goodchild, M. F., "Fractals and the Accuracy of Geo-

5. Belthoue, D. R., "Error Estimation by Point Counting

4. Goodchild, M. F., "Estimation of Land Parcel, the Map as a


1. Todd, T. W., "The Accuracy of Categorical Maps."