Geographic Information Systems (GIS) in Planning and Resource Management

ENVS 6189 3.0 – Session IV

John Sorrell
York University.
sorrell@yorku.ca
Session Purpose:

To discuss the various descriptions of data quality, some assessment techniques and effective management of error when engaged in GIS work.
Data Quality:

All geographic data inherently has error.

- Cost and Consequence
- Accountability
- Proactive Approach

Applications can be enhanced or corrected when this knowledge is available and considered
Precision and Accuracy:

Important distinctions:

1. **Accuracy** is the degree to which information on a map or in a digital database matches “true” or how close its values are to “accepted” values. Accuracy is an issue pertaining to the quality of data and the number of errors contained in a dataset or map. Includes: Positional, Attribute and Logical accuracy

2. **Precision** refers to the level of measurement and exactness of description in a GIS database. Precise locational data may measure position to a fraction of a unit. Precise attribute information may specify the characteristics of features in great detail.

   Precision and Accuracy are mutually independent
Error and Uncertainty:

... more important distinctions:

1. **Error** is used to describe the deviation between measured values and true values. Error usually includes an indication of the variation or discrepancy in quantitative terms which is why data quality is usually expressed in terms of error. There are three types of error: 1) Gross; 2) Systematic; and, 3) Random.

2. **Uncertainty** refers to the lack of confidence in the data that is due to incomplete knowledge of the data. Uncertainty is a measure of what we don’t know.

There is a cause and effect relationship between error and uncertainty.
Components of Data Quality:

Characteristics that affect data usefulness can be divided into three (3) categories or levels with nine (9) components.

MICRO - MACRO - USAGE

Micro level components are data quality factors that pertain to the individual data element. There are four (4) components:

- **Positional Accuracy** – planimetric and elevation;
- **Attribute Accuracy**;
- **Logical Consistency**;
- **Spatial Resolution**.

Usually expressed in quantitative units and evaluated by statistical testing of the data against an independent source of known and higher quality.
Components of Data Quality: MICRO - MACRO - USAGE

Assessment of Positional Accuracy:

Individual Point Accuracy –

- Bounding Rectangle (min, max for X,Y)
- Mean Centre or Centroid - Simply the mean of X and Y
- Standard Deviation for X,Y (distance)

\[
X = \frac{\sum_{i=1}^{N} X_i}{N} \quad Y = \frac{\sum_{i=1}^{N} Y_i}{N}
\]

\[
S_x = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N-1}}
\]

\[
S_y = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}}
\]
Components of Data Quality:

MICRO - MACRO - USAGE

Assessment of Positional Accuracy:

Overall Positional Accuracy
Root Mean-Square Error (RMSE)

\[
\begin{align*}
\text{rms}_x &= \sqrt{\frac{X^2}{n}} \\
\text{rms}_y &= \sqrt{\frac{Y^2}{n}} \\
\text{rms}_p &= \sqrt{\left(\text{rms}_x\right)^2 + \left(\text{rms}_y\right)^2}
\end{align*}
\]

Where,

\[X^2 = dx_1^2 + dx_2^2 + \ldots + dx_n^2\]
\[dx = \text{- discrepancies in X coordinate} = X_{\text{reference}} - X_{\text{sample}}\]
\[y_x = \text{- discrepancies in Y coordinate} = Y_{\text{reference}} - Y_{\text{sample}}\]
\[E^2 = \sum e_i^2 + e_2^2 + \ldots + e_n^2\]
\[n = \text{total number of points checked}\]

*Spatial variation – see overhead*
Components of Data Quality:  
Evaluation of attribute accuracy is obtained by comparing values of sample data units with reference data obtained from field checks or data with higher degrees of accuracy. Measurement scale of attribute data will determine method of assessment used.

Interval and Ratio:  
- mean deviation between actual and observed is used to generate an error index similar to RMSE.

\[ PCC = \left( \frac{\sum d}{n} \right) \times 100 \]

Where,  
- \( \sum d \) = sum of values along diagonal  
- \( n \) = total number of sample locations

Nominal and Ordinal:  
- an error matrix is constructed to show the frequency of discrepancies between encoded values and actual values.
- Classification Error Matrix and Percent Correctly Classified (PCC) are used to determine overall accuracy.
Components of Data Quality:

MICRO - MACRO - USAGE

Macro level components pertain to the data set as a whole. There three (3) components:

- Completeness – coverage, classification and verification
- Temporal Resolution – world and database time.
- Lineage

GIS in Environmental Studies
Components of Data Quality:

Level 1                Level 2                   Level 3
Agriculture            Grains                   Broccoli
                      Truck Crops               Carrots
                      Livestock                 Tomatoes
                      Cattle                   Other
                      Hogs                     Sheep
Forest                  Coniferous              Pine
                      Spruce                   Deciduous
                      Fir                      Mixed wood
Urban, Water

Sample Classification to Illustrate Concepts of Completeness (nominal measurement scale)

Exhaustive and Mutually Exclusive…

*GIS in Environmental Studies*
Components of Data Quality:

Usage components of data quality are specific to the resources of the organization. They include:

- Accessibility
- Direct and Indirect Costs
Compounding Error – Propagation:

GIS work usually employs multiple set of spatial and attribute data. It is unlikely all data are equal in quality components.

Propagation occurs when one error leads to another.
Compounding Error – Cascading:

Cascading occurs when erroneous, imprecise or inaccurate information skews a GIS solution.

Cascading occurs in an additive or multiplicative fashion.

GIS in Environmental Studies
Sources of Error:

<table>
<thead>
<tr>
<th>Category</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection</td>
<td>errors in field data collection</td>
</tr>
<tr>
<td></td>
<td>errors in existing maps used as source data</td>
</tr>
<tr>
<td></td>
<td>errors in the analysis of remotely sensed data</td>
</tr>
<tr>
<td>Data Input</td>
<td>inaccuracies in digitizing caused by operator and equipment</td>
</tr>
<tr>
<td></td>
<td>inaccuracies inherent in the geographic feature (e.g., edges, such</td>
</tr>
<tr>
<td></td>
<td>as forest edges, that do not occur as sharp boundaries)</td>
</tr>
<tr>
<td>Data Storage</td>
<td>insufficient numerical precision</td>
</tr>
<tr>
<td></td>
<td>insufficient spatial precision</td>
</tr>
<tr>
<td>Data Manipulation</td>
<td>inappropriate class intervals boundary errors</td>
</tr>
<tr>
<td></td>
<td>error propagation as multiple overlays are combined</td>
</tr>
<tr>
<td></td>
<td>slivers caused by problems in polygon overlay procedures</td>
</tr>
<tr>
<td>Data Output</td>
<td>scaling inaccuracies</td>
</tr>
<tr>
<td></td>
<td>error caused by inaccuracy of the output device</td>
</tr>
<tr>
<td></td>
<td>error caused by instability of the medium</td>
</tr>
<tr>
<td>Use of Results</td>
<td>the information may be incorrectly understood</td>
</tr>
<tr>
<td></td>
<td>the information may be inappropriately used</td>
</tr>
</tbody>
</table>
Summary:

- Error is associated with all representations of geographic information;
- Error encompasses horizontal, vertical, elevation and attribute data;
- Implicit assumptions about data quality must be made explicit so they are addressed for application development or decision making;
- Each data source and processing method introduces a level of error into the final products;
- Propagation and Cascading are a result of data layers (inputs) not having equal levels of accuracy being combined to generate a new data layer;
- The objective in dealing with error should not be to eliminate it, but to manage the implications.

End.