Errors in GIS

GEOG 300, Lecture 14

Dr. Anthony Jjumba

1

Error and Uncertainty

- Uncertainty
  - A lack of sureness about something… the same as a lack of knowledge
    - lack of knowledge about level of error
      - Indicates the extent of unreliability
- Error
  - degree of doubt in a measurement
  - e.g. 5% error (calculated by difference between known & measured values)

2

Error and Uncertainty

- UNCERTAINTY…
  - To you the scientist:
    - A statement of knowledge
    - Useful information on the limits
- UNCERTAINTY…
  - To the general public and decision makers:
    - Sign of weakness
    - Lack of trust and confidence
    - Confusing

Propagation of Errors

- In GIS applications we combine data from different sources.
  - The data have uncertainties associated with them.
    - How can you state the effect of these uncertainties to the final result?
    - What impact does error in each data layer have on the final result?

4

Propagation of Errors

- Example:
  - Find the most likely locations where human settlement will infringe on the Agricultural Land Reserve.
    - The data for the study area are available in a raster format
    - You have identified about 10 factors likely to influence infringement e.g.
      - agricultural productivity (dollars per hectare)
      - Existing settlements (presence or absence)
      - Parcel size (number of owners per hectare)

5

Propagation of Errors

- Example:
  - the factors can be combined, or cascaded, to a single measure of ***likelihood*** on a scale of 0 through 10
    - the cascading rules could group factors into composites such as "social impact", "agricultural impact" and then weight each group relative to the others
    - the rules used in cascading include weighted addition: likelihood = w1x1 + w2x1
      - as well as simple conditions:
      - likelihood = 0 if settlement = "present"
      - and reclassifications:
        - likelihood = 3 if factor A has value of X

6

Propagation of Errors

- Error Analysis
  - Sources of errors in the above example?
    - Input data
    - relative importance of factors
  - It quickly becomes evident that rigorous analytical methods are needed to ascertain errors in the final outputs.

"the ultimate arbiter of cartographic error is the real world, not a mathematical formulation"

Nicholas R. Chrisman

7

Propagation of Errors

- Errors propagate due to the multiplication of numerical error and initial measurement error.
- Estimating the degree of an error is an important area of research in GIS and computational science.

8

Error Analysis

- Error Analysis:
  - Some Methods
    - Sensitivity Analysis
      - Used to determine how different values of an input variable(s) impact the output.
      - Described as the study of how uncertainty in the output of a model can be attributed to different sources of uncertainty in the inputs
    - Validation
      - Used to determine how close the results match up to data in the real world

9

Error Analysis

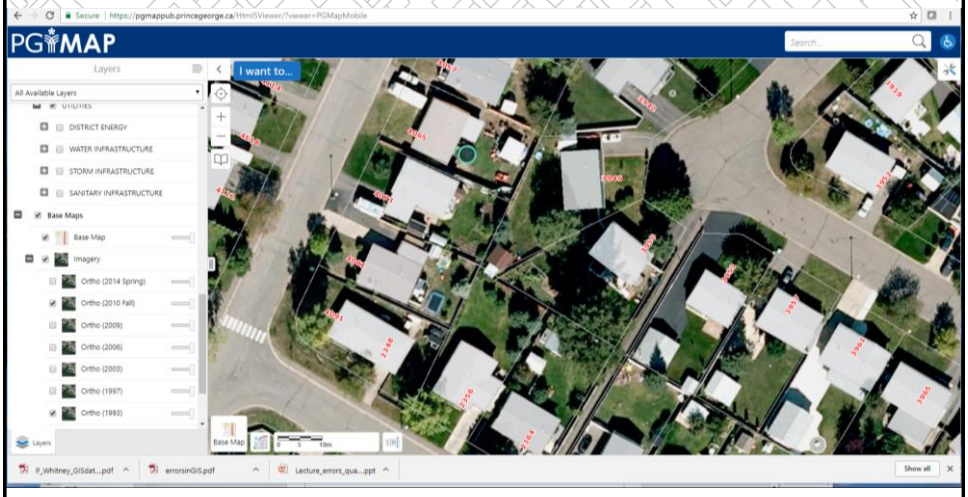
- Error Analysis:
  - Cross-validation
    - remove one data value at a time and predict its associated value using the values at the rest of the locations

10

Error Mitigation

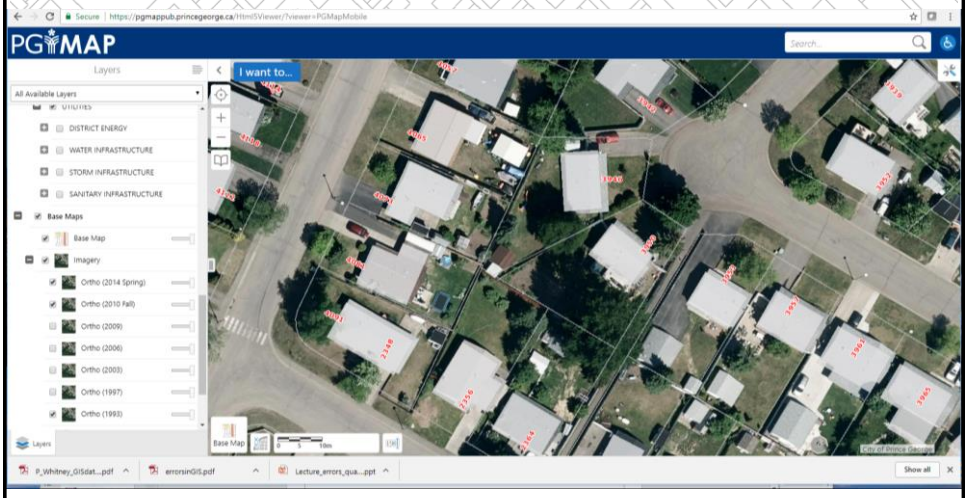
- Techniques to mitigate sources of errors?
  - Quality Control in measurements
  - Develop standards
  - …
- Consider spatial location differences between imagery
  - Are there errors or just expected differences

11



2010 Imagery

12



2014 Imagery

13

RMSE

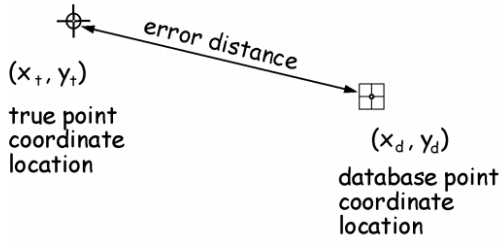
- **Root Mean Square Error (RMSE)**
  - A measure of the differences between predicted values (results) and the values actually observed (real world).
  - The RMSE represents the standard deviation of the differences between predicted values and observed values

14

RMSE

- **Root Mean Square Error (RMSE)**

One needs measured coordinates and true coordinates



The lower the error distance, the more the accuracy

$$\text{error distance} = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2}$$

RMSE

- $RMSE = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 \ldots e_n^2}{n}}$

e = error distance
n = number of observations

- Try it
  - https://www.dropbox.com/s/fb0e7g07an08bgb/rmse_sample_data.xlsx?dl=0

16

Some not good practices

- Bad Analysis Practices
  - Overstating the accuracy of the results
    - Data accuracy/quality
      - Vector
      - Raster
    - Area proportion overlays
    - Mixed scales
    - Attribute accuracy
  - Ignoring topological errors

17

MAUP

- The Modifiable Areal Unit Problem (MAUP)
  - “a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomena resulting in the generation of artificial spatial patterns”
  - Geographical boundaries are imposed.
    - They do not relate in any meaningful sense to the variables of interest.
  - Data are aggregated from source observations to zones or output areas which have no special meaning in terms of the underlying geographical process of phenomenon.

18

MAUP

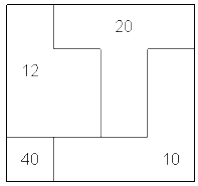
100	200	400	100
200	300	500	100
400	200	100	100
300	100	200	100

Total population in 16 areas

10	20	20	20
5	10	20	10
15	15	20	10
40	10	10	10

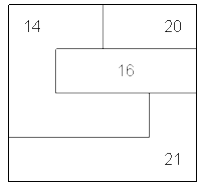
Percentage unemployment

(A)



Percentage unemployment aggregated to four areas

(B)



Percentage unemployment aggregated to four areas

19

http://www.restore.ac.uk/geo-refer/91023cwors00y00000000.php

Ecological Fallacy

- Ecological Fallacy:
  - The interpretation of statistical data where inferences about individuals are deduced from data for the group to which those individuals belong.
  - Relationships observed for groups do not necessarily hold for individuals, and vice versa
  - For instance, assume that you measured the average weight of moose in Jasper, you cannot expect that the next moose you see is of the same weight.

20

Coordinate adjustments	*Generalization*
rubber sheeting/transformations	linear alignment
projection changes	line simplification
datum conversions	addition/deletion of vertices
rescaling	linear displacement
Feature Editing	*Raster/Vector Conversions*
line snapping	raster cells to polygons
extension of lines to intersection	polygons to raster cells
reshaping	assignment of point attributes
moving/copying	to raster cells
elimination of spurious polygons	post-scanner line thinning
Attribute editing	*Data input and Management*
numeric calculation and change	digitizing
text value changes/substitution	scanning
re-definition of attributes	topological construction / spatial indexing
attribute value update	dissolving polygons with same attributes
Boolean Operations	*Surface modelling*
polygon on polygon	contour/lattice generation
polygon on line	TIN formation
polygon on point	Draping of data sets
line on line	Cross-section/profile generation
overlay and erase/update	Slope/aspect determination
Display and Analysis	*Display and Analysis*
cluster analysis	class intervals choice
calculation of surface lengths	areal interpolation
shortest route/path computation	perimeter/area size/volume computation
buffer creation	distance computation
display and query	spatial statistics
adjacency/contiguity	label/text placement

Common Sources of Errors in GIS Analysis

All Analyses?

Hunter, G. J., and Beard, K. Understanding error in spatial databases. The Australian Surveyor 37, 2 (1992), 108–119. 433

21