GEOG 204

LECTURE 7

1

Data Quality

- The power of GIS analysis is based on the assembly and manipulation of layers of data, but errors may rapidly propagate during analysis
- "Garbage in, garbage out"
  - Poor data quality leads to the poor decisions based on resulting from the analysis.
- …. High quality data are expensive

2

2

Data Quality

- Geographic Information Systems
  - The context
    - Widely used for decision support applications
    - Reliance on data sourced from a myriad providers
      - Citizen Scientists, Open Data Portals, Government,
    - Low-quality data in decision making can have severe consequences
    - Inappropriate use of GIS functions can introduce errors
      - geometric and other transformations to the spatial data

3

3

Data Collection

- Data Collection:
  - Traditionally, most spatial data were collected and held by individual, specialized organizations
    - national mapping agencies
    - energy supply companies,
    - local government departments
  - Increasingly, many users, agencies are collecting their own data.
    - Low cost of data capture equipment
    - Quality control is as much the responsibility of the producer as it is for the user

4

4

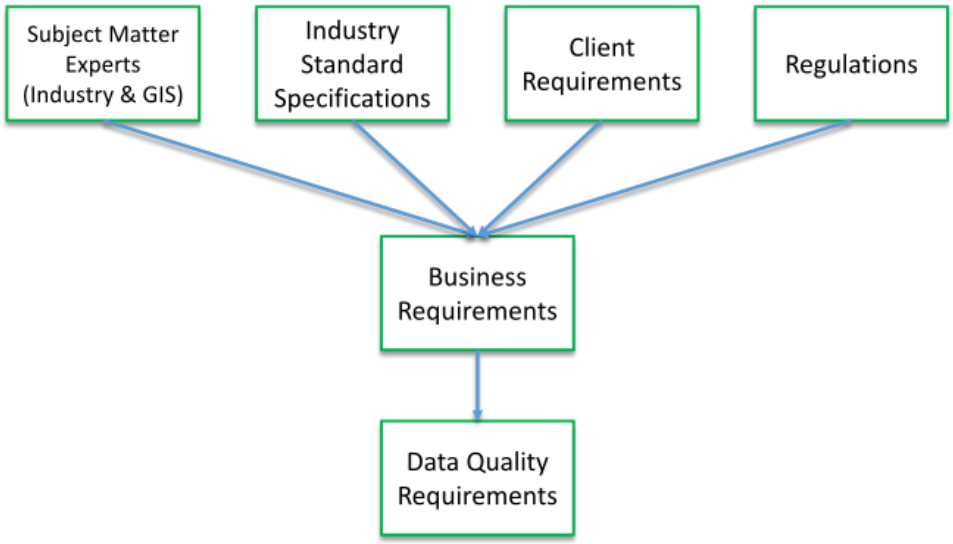
Data Collection

- Data Collection:
  - If data are to be shared, considerations include
    - what data exists
    - where
    - format
    - quality requirements/specifications
    - metadata: the 'data about data'

5

5

To collect data



Source: ESRI

6

6

Errors in Data

- Sources of Errors:
  - **Human errors** include mistakes, such as reading an instrument incorrectly, and faulty judgments (e.g. ambiguous boundaries such as high water mark).
  - **Environmental characteristics**, such as variations in temperature can result in measurement errors
  - **Instrument errors** Measurements are as precise as the instrument's capabilities.
    - The smallest measurement that can be made is the instrument's resolution.

7

7

Elements of Data Quality

- Data quality elements:
  - Elements or components used to describe the quality of the data
  - They provide information on the suitability for data usage by describing
    - Why (purpose) data were collected
    - when (age) data were collected
    - How the data are created (method)
    - and how accurate the data are (limits)

8

8

Elements of Data Quality

- Accuracy
  - Positional accuracy
    - closeness of locational information (usually coordinates) to the true position
    - Generally, paper maps are accurate to roughly one line width or 0.5 mm
      - On a 1:10,000 scale, 0.5mm is equivalent to?
      - NTS/NTDB 1:50,000 = < 25 metres
      - BC TRIM: 1:20,000 = 10 metres
      - BC/Federal: 1:250,000 = 125 m
  - Thematic/attribute accuracy
    - the closeness of attribute values to their true value

9

9

Elements of Data Quality

- Lineage
  - a record of the data sources and of the operations which created the database
    - how were they digitized, from what documents?
    - when were the data collected? By who?
  - is often a useful indicator of accuracy
- Logical consistency
  - refers to the consistency of the data model (particularly the topological consistency)
    - is the database consistent with its definitions?
    - is there exactly one label for each polygon?
    - are there nodes wherever arcs cross, or do arcs sometimes cross without forming nodes?

10

10

Elements of Data Quality

- Completeness
  - degree to which the data exhausts all the possible items
    - are all possible objects included within the database?
  - affected by rules of selection, generalization and scale

11

Elements of Data Quality

- Temporal quality
  - The quality of temporal attributes and temporal relationship of features.
- Data usability
  - Suitability to an application and its related functional requirement

12

Data Quality - Key Issues

- Key Concepts
  - Accuracy, Precision and Uncertainty

- Accuracy:
  - closeness of the measurements, computations to the true values (or values accepted to be true)
    - spatial data are a generalization of the real world, the "true value" is thus an estimate of the real world
  - ~ absence of errors

13

13

Data Quality - Key Issues

- Precision:
  - the number of decimal places or significant digits in a measurement
    - precision is not the same as accuracy - a large number of significant digits doesn't necessarily indicate that the measurement is accurate
  - a GIS works at high precision, mostly much higher than the accuracy of the data itself

14

14

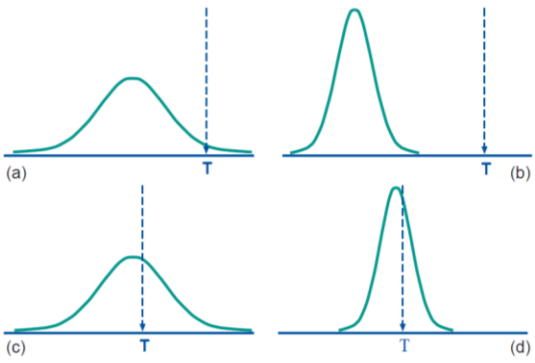
Data Quality - Key Issues

- Precision and Accuracy
  - If there are systematic variations in either the instruments used, or the phenomenon measured, this affects both accuracy and precision.

15

Data Quality - Key Issues

- Precision and Accuracy

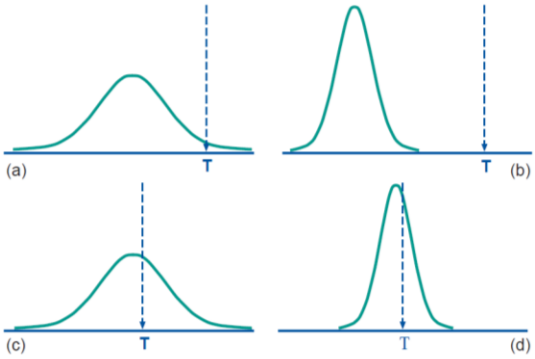


Consider 40 students measuring the length of a line

16

Data Quality - Key Issues

- Precision and Accuracy



Consider 40 students measuring the length of a line

Figure 5.2: A measurement probability function and the underlying true value T: (a) bad accuracy and precision, (b) bad accuracy/good precision, (c) good accuracy/bad precision, and (d) good accuracy and precision.

17

17

Data Quality - Key Issues

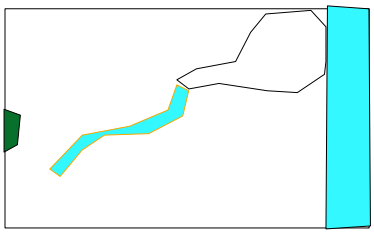
- *"All observations are inexact"*
- Spatial data are inaccurate to some degree therefore
  - accuracy assessment is important
  - tracking how errors are propagated through GIS operations is important
  - Take care not to assign greater accuracy to data than what it has
- Some data are intentionally imprecise
  - It is important to know the limitations

18

18

Data Quality - Key Issues

- Uncertainty: our imperfect and inexact knowledge of the world
  - Positional uncertainty
  - Attribute uncertainty
  - Definitional uncertainty
  - Measurement uncertainty



19

Data Quality

Some Considerations/Illustrations

20

Know the Standards

Specifications:

File Format:

- All spatial data provided to the Ministry of Forests must be in Intergraph Design File (IGDS) format Version 8.0 or later. Spatial data in the IGDS (.DGN) files will be in 2D format with the
- Universal Transverse Mercator (UTM) projection (Clarke Spheroid) with no data linkages attached.

Input Scale:

The tile size for digital data capture will be 1:100,000 (1/4 letter blocks - 6x5 1:20,000 per mapsheet) using TRIM base in the NAD 83 datum, however, circumstances may require the use MOF 1:20,000 graphics (.fc1) mapsheet files as the base reference.

Positional Accuracy:

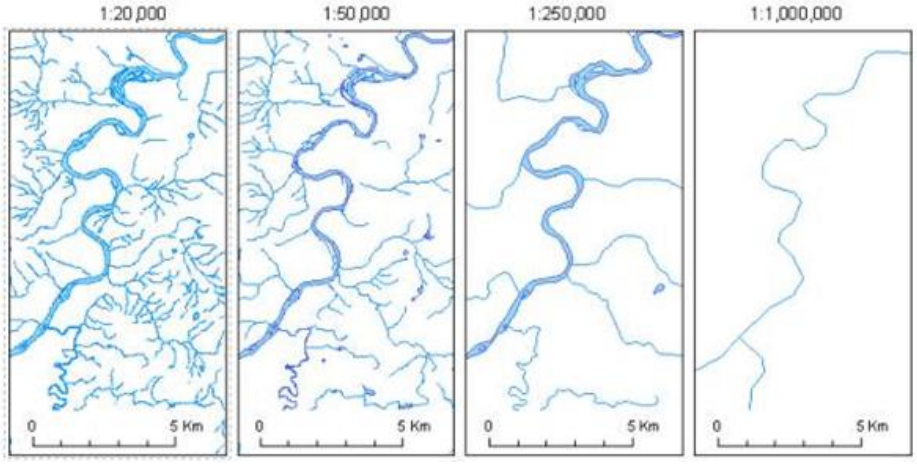
Captured spatial data must have a positional accuracy of 1mm at map scale (for 1:100,000 this is 100 m on the ground) when compared to the original paper map 90 percent of the time.

21

21

SCALE and PRECISION (not accuracy)

Data from a smaller scale has lower resolution (precision)
Details, number of features decrease with smaller scale
[both spatial location details and attributes]

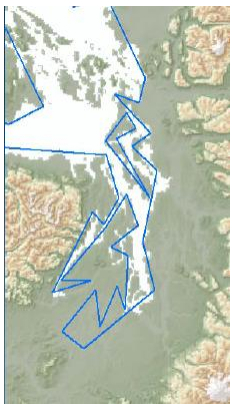


22

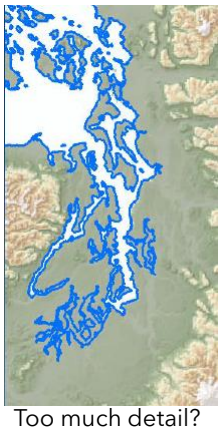
22

Data precision and display Scale

Scale – higher resolution shouldn't be used at smaller scales (too much data) and vice versa (too little).



Too little detail



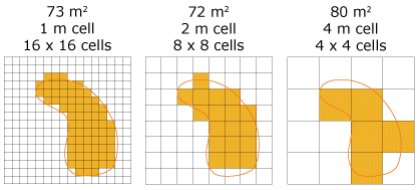
Too much detail?

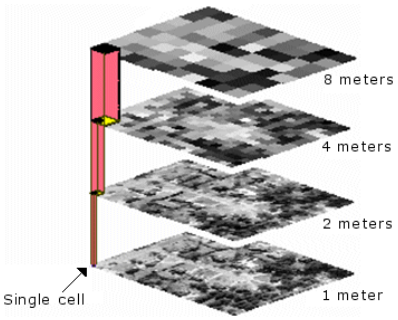
Coastline and lake boundaries: location uncertainty related to tides and fluctuating water levels

Image source: Esri

Raster precision = pixel size resolution

e.g. Landsat 30m, (Google) GeoEye 50cm



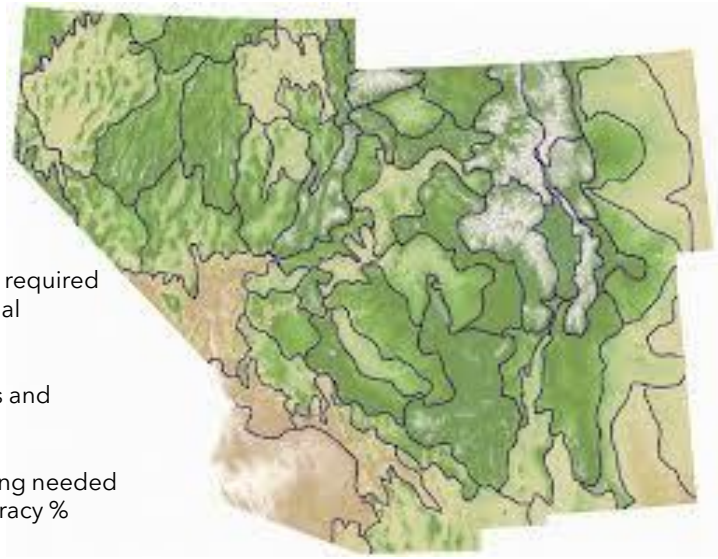


Scale 1:20,000
Cell size: 15 m

Scale 1:20,000
Cell size: 5m

Uncertainty –in natural resources and gradual boundaries

Subjective: 10 people might digitize 10 different sets of lines – polygons and attributes



Consistency required e.g. provincial guidelines

And for soils and geology

Field checking needed to give accuracy %

25

25

Data Quality in Natural Resources

- Some factors causing loss in data quality
  - Scale – spatial data and attributes
  - Density of observations and processing methods
  - Area cover – gaps due to accessibility
  - Age of data – precision and changes

26

26

Summary

- Know the limitations of your data
  - When was it created
  - What level of precision was expected
  - What level of error was accepted
- Don't shoot the messenger if you're the boss
  - Input quality is a limiting factor
- Don't inadvertently lie to the client
- Be careful with simplified/smoothed data
- DISCLAIMERS? use them

27