GEOG 204

LECTURE 14

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

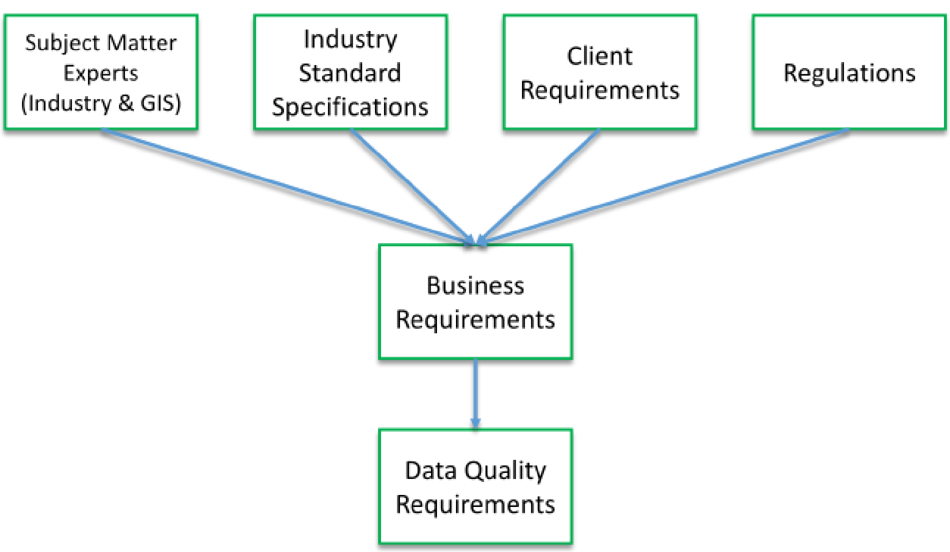
Data Collection

- 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

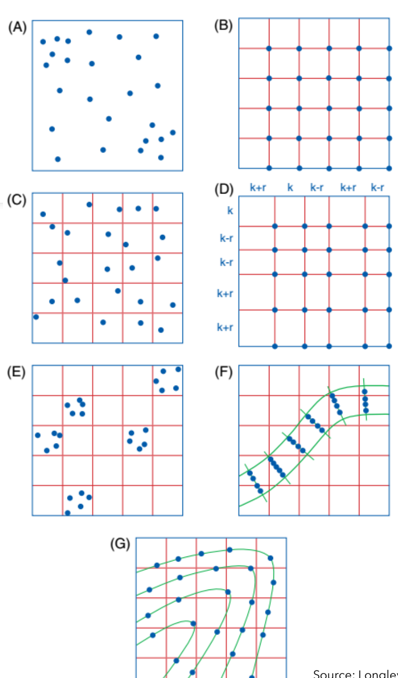
6

Sampling

- Provides knowledge about a whole population
    - i.e. make inference about a population from the sample data
- Larger sample sizes are more accurate representations of the whole
    - Large samples are costly: time, labour
        - Can be wasteful since we can statistically infer from appropriate samples
- A sampling strategy with the minimum bias is the most statistically valid

7

7



Sampling

Spatial sample designs: (A) simple random sampling, (B) systematic sampling, (C) stratified random sampling, (D) stratified sampling with random variation in grid spacing, (E) clustered sampling, (F) transect sampling, and (G) contour sampling.

Source: Longley, Paul A.; Goodchild, Michael F.; Maguire, David J.; Rhind, David W.. Geographic Information Science and Systems, 4th Edition. Wiley.

8

8

Random Sampling

- Random sampling: each member of the population has an equal chance of being selected
  - Advantages:
    - Can be used with large sample populations
    - Avoids bias
  - Disadvantages:
    - Can disproportionately represent some parts of the population at the expense of others

9

9

Systematic Sampling

- Systematic Sampling: Samples are chosen at regular intervals
  - Sample locations are evenly distributed for example every two metres along a transect line
    - systematic sampling implies a regularly spaced grid
      - Advantages:
        - It is more straight-forward than random sampling
        - Provides a good coverage of the study area
      - Disadvantages:
        - It is more biased: not all points have an equal chance of being selected
        - It may lead to over or under representation if there is periodicity in the data (e.g. sampling at the same interval as the location of erosion barriers along a beach. Or a city road grid)

10

10

Stratified sampling

- Stratified sampling: used when the parent population is made up of sub-groups that of interest.
  - Divide the sampling design into strata(classes), and then select a sample from each stratum
  - The strata are defined so that individuals inside each class are similar based on the characteristic believed to influence the phenomena

Stratified sampling

- Advantages:
  - If the proportions of the subgroups are known, the results are representative of the whole population
  - Correlations and comparisons can be made between subgroups
- Disadvantages:
  - The proportions of the subgroups must be known

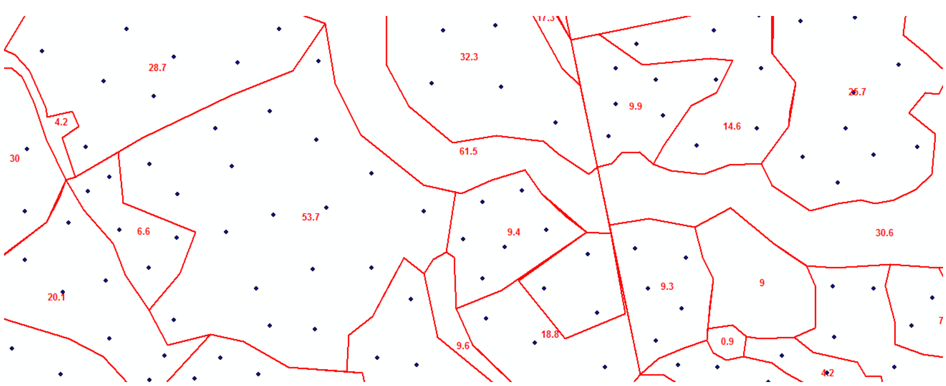
Air Photos for Stratified Sampling

- Looking for distinct, uniform areas
  - Crown size (age), harvest history
  - Hardwoods (gray) and softwoods (green)

13

Stratified Sampling

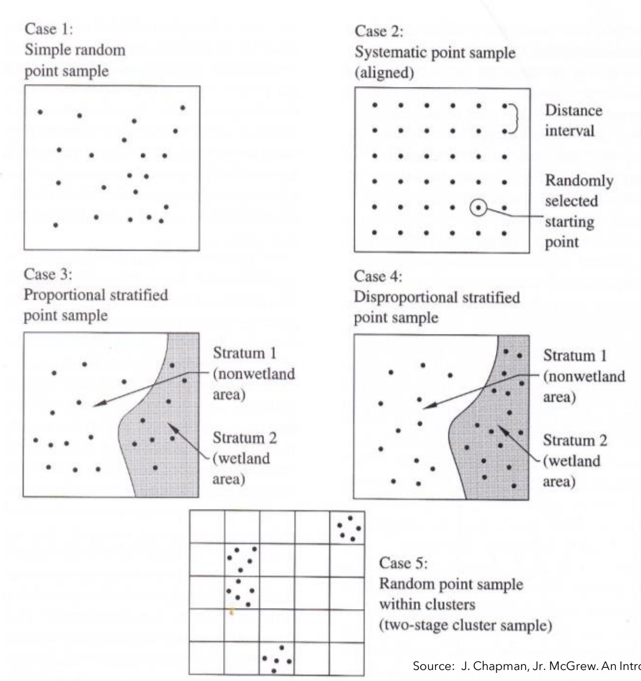
- Generate sample points randomly
  - X points per area, e.g. 1 point every 3 hectares
  - Each point tied to polygon = unique stand



14

Stratified Sampling: Population

15

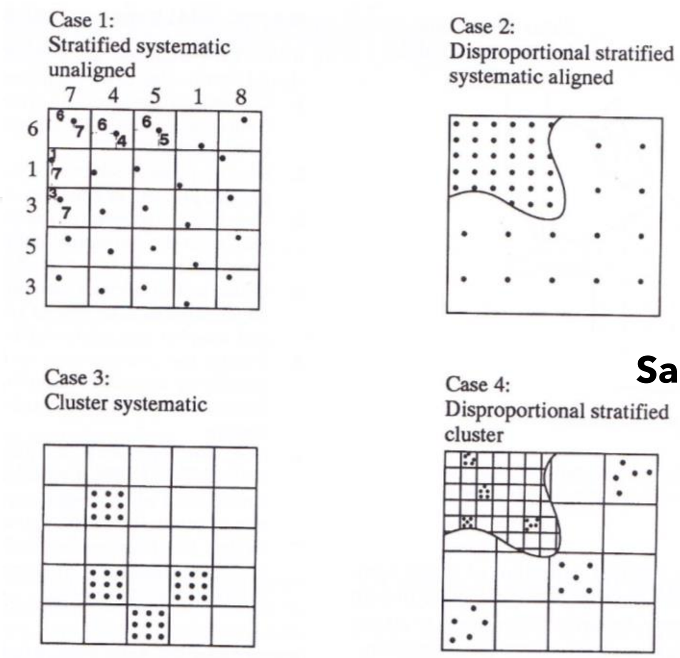


Point Sampling Methods

Source: J. Chapman, Jr. McGrew. An Introduction to Statistical Problem Solving in Geography

16

16

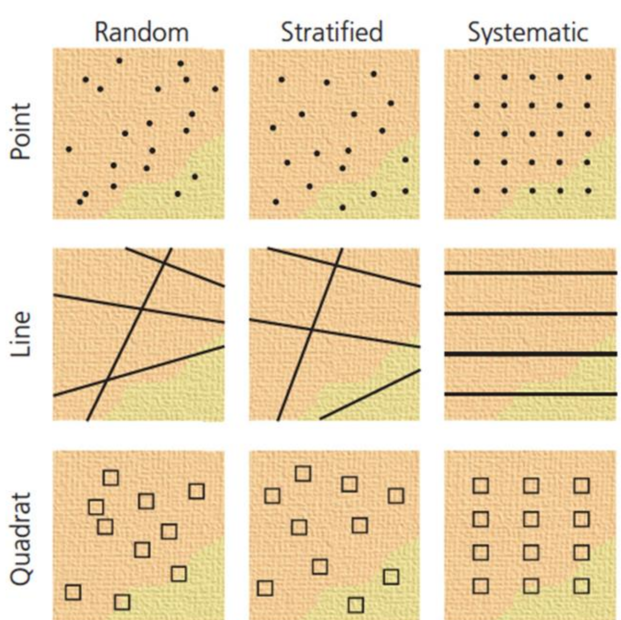


Hybrid Point Sampling Methods

Source: J. Chapman, Jr. McGrew. An Introduction to Statistical Problem Solving in Geography

17

17



Source: https://www.geography-fieldwork.org/a-level/before-starting/methods/sampling/

18

18

Levels or Scales of Measurement

- Nominal
  - Categorical data e.g. land use type, religious affiliation
- Ordinal
  - Ranked data , e.g. main, secondary, minor roads
- Interval:
  - Interval between any two units can be measured on scale. Zero value is assigned arbitrarily e.g. Celsius and Fahrenheit scales (80°F is not twice as hot as 40°F)
- Ratio:
  - interval data with an absolute zero value

19

19

Summary of Levels of Measurement

Level of measurement	Brief description
Nominal	Each value or unit of data is assigned to one of at least two categories or qualitative classes; no assumptions are made about relationships between categories—only that they are "different."
Ordinal	Values themselves are placed in some rank order.
Strongly ordered	Each value or unit of data is given a particular position in a rank-order sequence; that is, each value is assigned its own particular rank.
Weakly ordered	Each value or unit of data is assigned to a category, and the categories are then rank ordered.
Interval	Each value or unit of data is placed on a measurement scale, and the interval between any two units of data on this scale can be measured; origin or zero starting point is assigned arbitrarily (i.e., origin does not have a "natural" or "real" meaning).
Ratio	Each value or unit of data is placed on a measurement scale, and the interval between any two units of data on this scale can be measured; origin or zero starting point is "natural" or non-arbitrary, making it possible to determine the ratio between values.

Source: J. Chapman, Jr. McGrew. An Introduction to Statistical Problem Solving in Geography

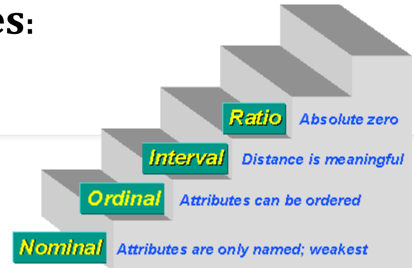
OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

20

20

GIS Attribute field types:

Interval/ Ratio:



Integer (long/short)
Short: - 32,000 to +32,000
Long: - billions to + billions

Real (float/double)
Float: 'precision'- field length; scale - number of decimals
Double: 15 digits

Nominal: **Text, (includes code number)**

Ordinal: **Date, Text, (short integer)**

21

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
    - e.g. Round off errors
  - **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.

22

22

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)

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 = $\leq$ 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

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?

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

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

27

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

28

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

29

29

Data Quality - Key Issues

- Precision and Accuracy
  - Speak to data quality and the errors in the data.
    - Applies to geographic position, attribute/thematic information, conceptual accuracy (when modeling)
  - **Accuracy:** Closeness with which spatial data marches the values in the real world.
  - **Precision:** Exactness in the measurement or description of the data
    - Precise data may be inaccurate

30

30

Data Quality - Key Issues

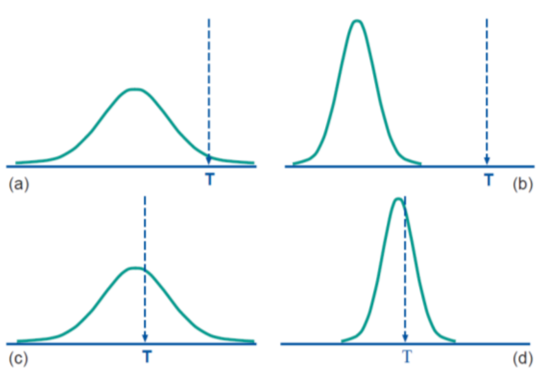
- Precision and Accuracy
  - If there are systematic variations in either the instruments used, or the phenomenon measured, this affects both accuracy and precision.

31

31

Data Quality - Key Issues

- Precision and Accuracy



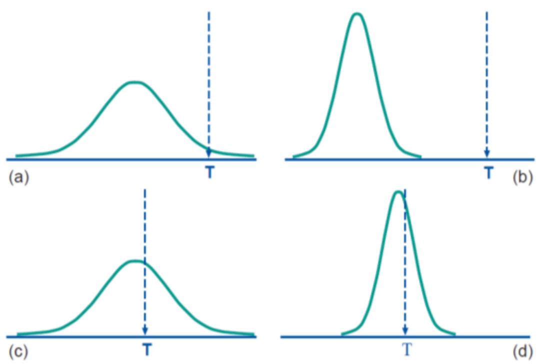
Consider 40 students measuring the length of a line

32

32

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.

Precision and Accuracy

- GIS software uses 'double-precision' – capable of storing 15 digits
  - E.g. decimal places (of meters – UTM) or 10 (latitude/longitude – WGS)
    - 560157.324687 or 52.4974294521

Lat	Lon
51.592443225	-122.242216653
51.590503265	-122.254802119
51.590917946	-122.252207907

- In most cases, this level of precision is <u>not</u> warranted by the data.
  - What is the justification for reporting millimeter precision on trail lengths?

Data Quality - Key Issues

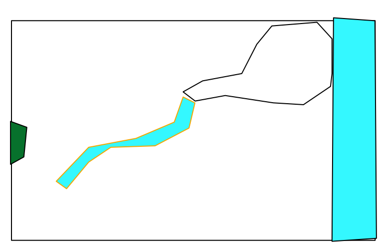
- *"All observations are inexact"*
- Spatial data are inaccurate to some degree therefore
  - accuracy assessment is important
  - tracking how errors are propagated through GIS operations is important
  - Take care not to assign greater accuracy to data than what it has
- Some data are intentionally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

35

35

Data Quality - Key Issues

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



36

36

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)

37

37

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

38

Data Quality

Some Considerations/Illustrations

Know the Standards

Specifications:

File Format:

- E
- (

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:

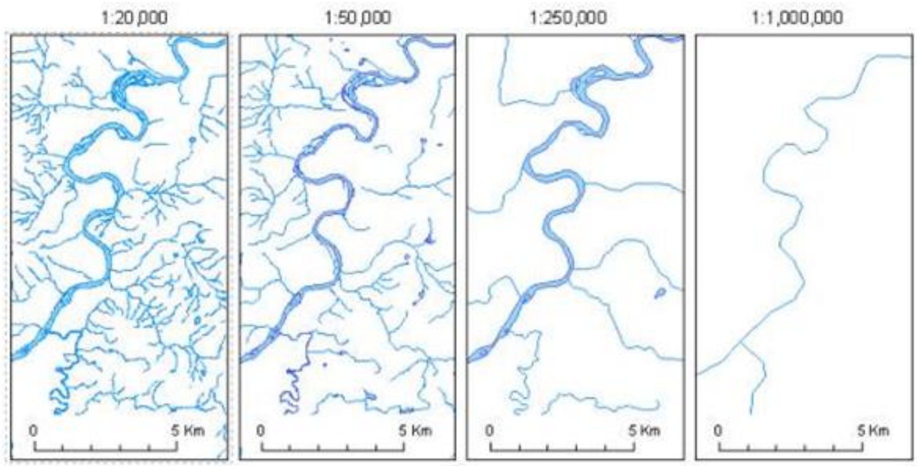
The tile size for digital data capture will be 1:100,000 (1/4 letter blocks - 6x5 1:20,000 per mapsheet) using TRIM base in the NAD 83 datum, however, circumstances may require the use MOF 1:20,000 graphics (.fc1) mapsheet files as the base reference.

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.

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]

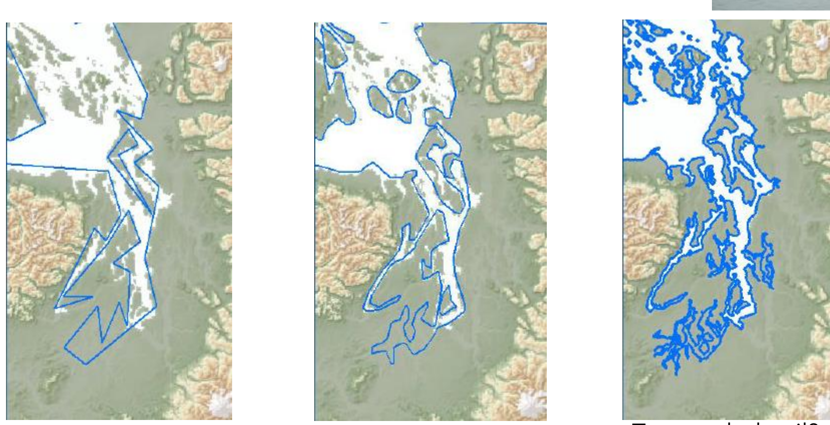


41

41

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

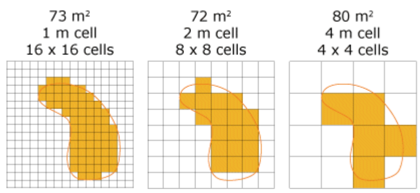
42

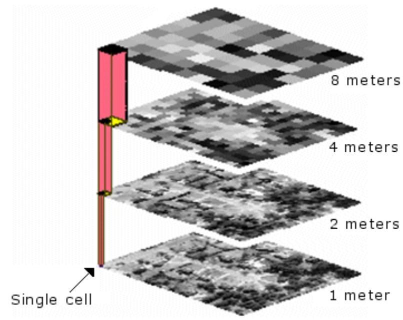
Image source: Esri

42

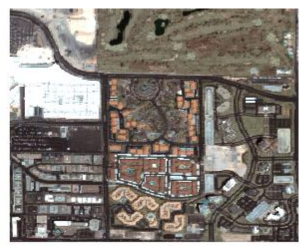
Raster precision = pixel size resolution

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





Scale 1:20,000
Cell size: 15 m



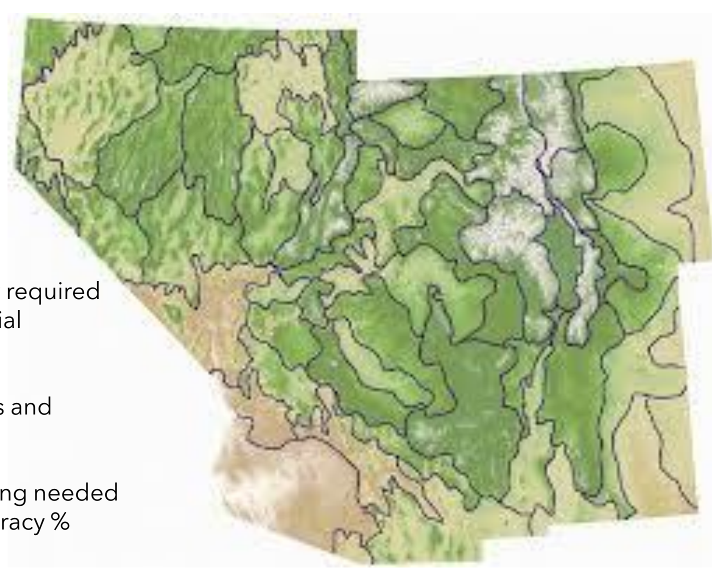
Scale 1:20,000
Cell size: 5m

43

43

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 %

44

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

45

45

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

46

46