Geodata quality

Spatial Data Analysis and Simulation modelling,
2020, Simon Scheider

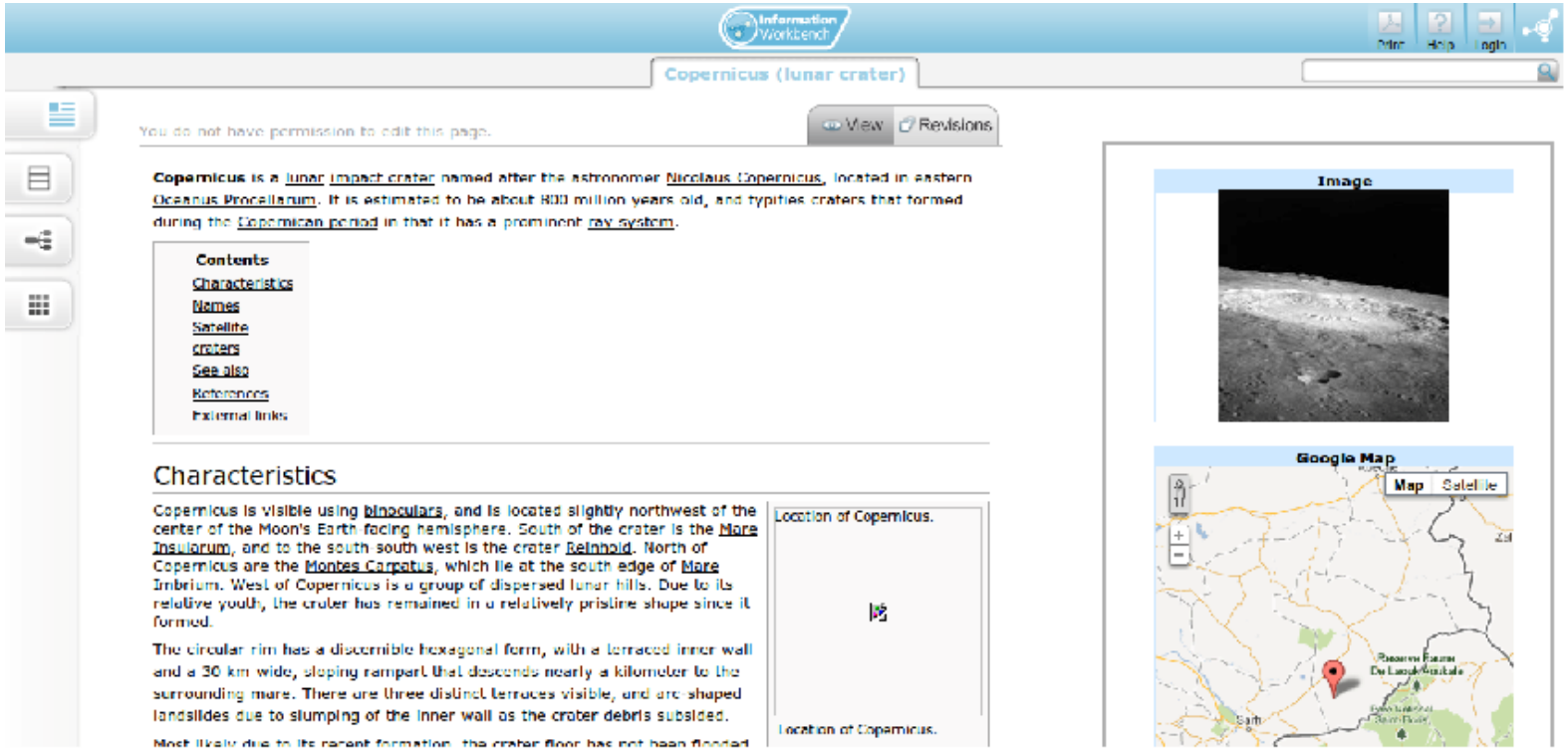
Outline

- (Some examples of) spatial uncertainty
- What is data quality?
- Geodata quality dimensions
- Veregin’s evaluation matrix
- Accuracy
- Resolution (and scale)
- Completeness

Moon landing in Algeria

- DBpedia version (2015-04) showed
  that the moon landing (Copernicus crater) happened here on Earth
  (in Algeria)
- (Why?
  Because DBpedia
  disregarded CRS…)

(see Janowicz et al. 2016)



The place of the Earth (on DBpedia)

```
SELECT distinct ?lat ?long ?populationCount
WHERE {
<http://sws.geonames.org/6295630/> geo:lat ?lat ; geo:long ?long ;
 ptop:populationCount ?populationCount.}
```

lat	*long*	*populationCount*
0	0	6814400000

Listing 3.3 A query for the geographic coordinates of the Earth and its population.

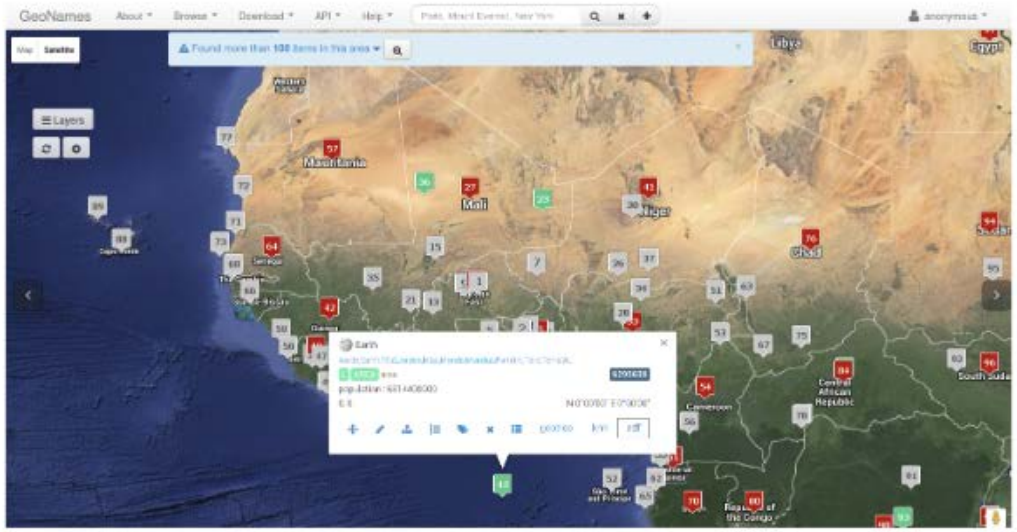


Fig. 4 The point-feature representation of the Earth.

(see Janowicz et al. 2016)

Spatial uncertainty

- Geodata is imperfect and uncertain in many respects
- from conceptualization to generalization, from measurement to analysis, information loss is unavoidable
- For example:
  Positional uncertainty:
  Street networks in Santa Barbara, CA from two different data sources (Li & Goodchild, 2011).

What is data quality?

The totality of data characteristics that bears its ability to satisfy stated and implied needs.

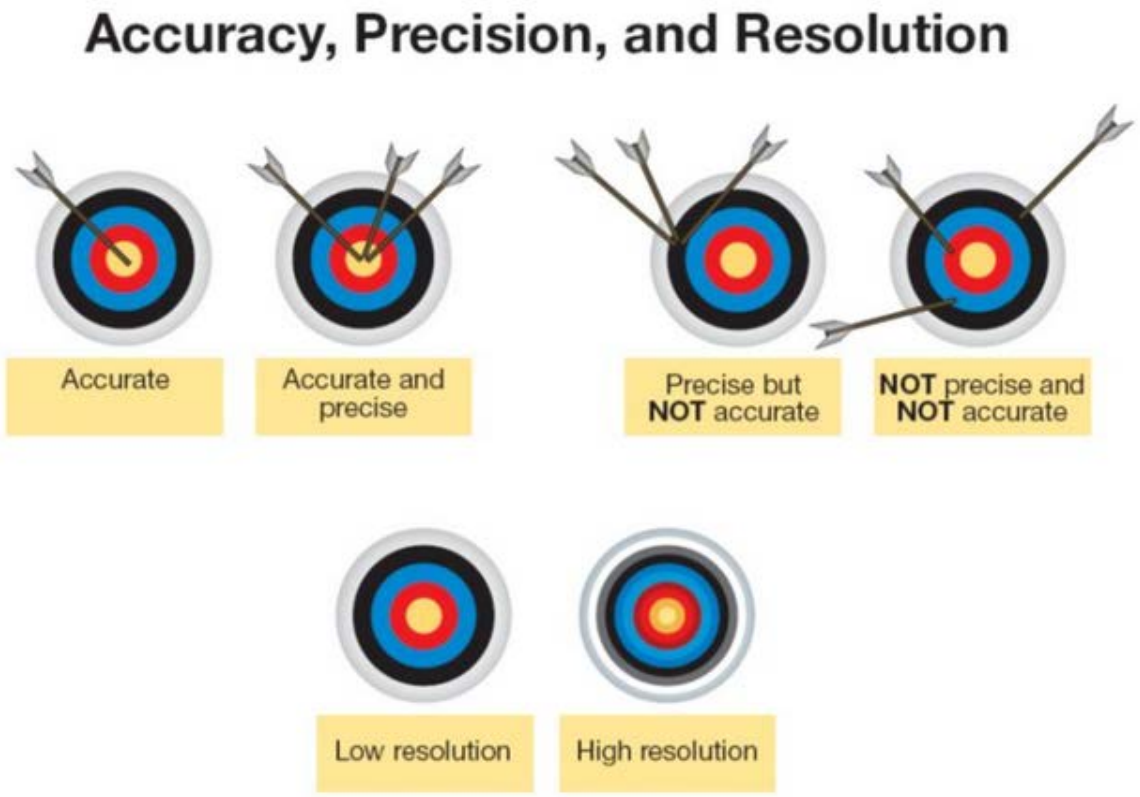
Quality can be assessed with the following approaches:

- following norm specifications (tenders): this is a technical approach, where the product and the production process are the essential elements to assure quality;
- quality as "fitness for use": this is a practical approach directed to the market, where quality is determined by the customer
- quality as "excellence": in this approach it is the point of view of the seller to determine the concept of quality;
- quality as recognized by experts

Geodata quality dimensions

Accuracy, precision and resolution are important quality terms that are often confused in the literature.

- *Accuracy:* the inverse of (systematic) measurement error.
- *Resolution:* the level of detail of data.
- *Precision:* May either mean the resolution of digital number representations,
- … or (as shown on the left) the standard error of measurement.



Veregin's geodata quality evaluation matrix

Geodata quality can be measured along all three *dimensions of geodata* (space, time, theme) for the following *quality dimensions*:

- Accuracy
- Resolution or precision (level of detail)
- Completeness
- Consistency (satisfaction of logical rules)

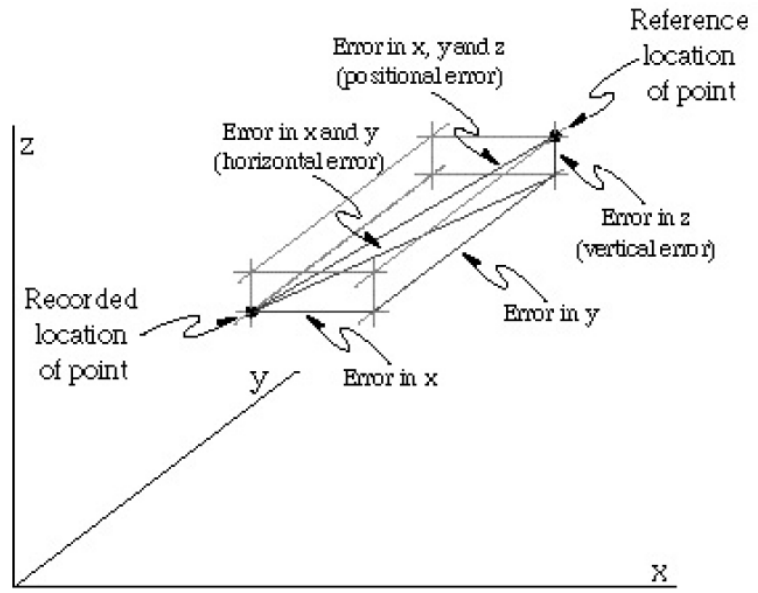
	Space	Time	Theme
Accuracy			
Precision			
Consistency			
Completeness			

Accuracy

- Accuracy is the inverse of measurement error.
- Needs to be measured relative to some “ground truth”, which is given by some *specification* of what is considered to be true
- The specification serves as the standard against which accuracy is assessed. Thus the "actual" value is the value we would expect based on the specification.
- For example, we may specify a standard way of measuring terrestrial positions in geodesy, used as ground truth for measuring the error of GPS coordinates.

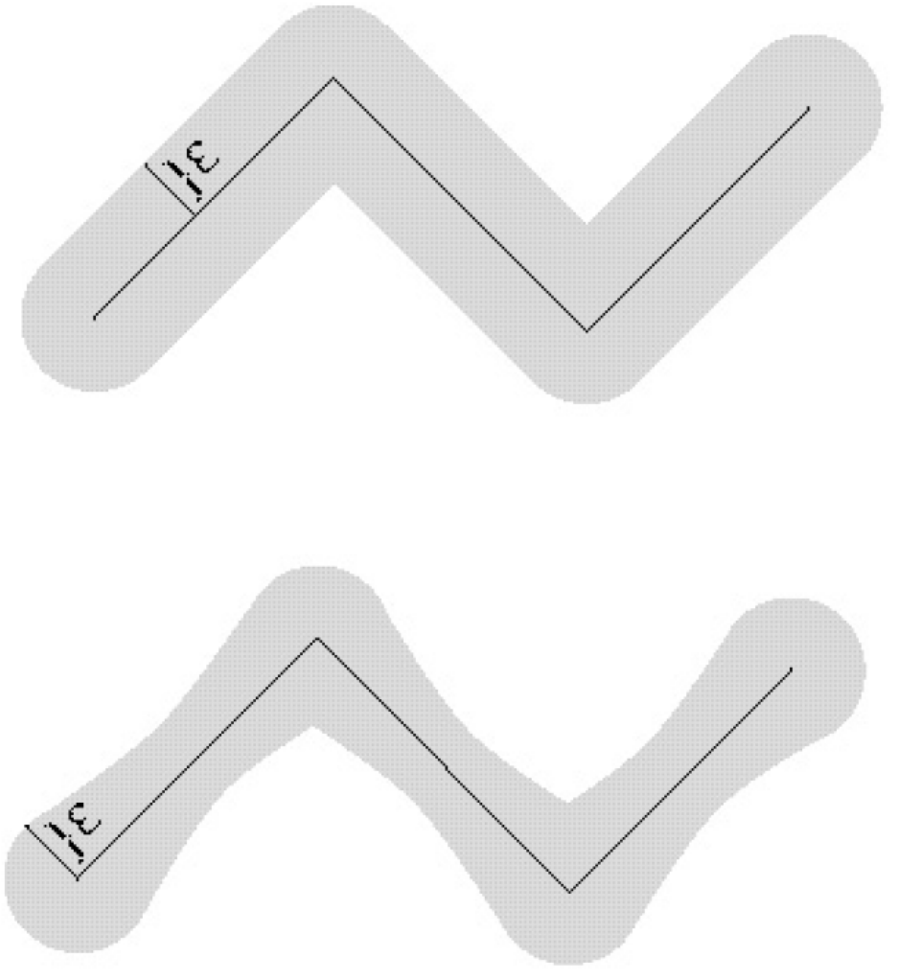
Spatial accuracy

- Spatial accuracy is the accuracy of the spatial component of the database. The metrics used depend on the dimensionality of the geometry
- For points:
  - Error can be defined in various dimensions: x, y, z, horizontal, vertical
  - Metrics of error are extensions of classical statistical measures (mean error, RMSE or root mean squared error, inference tests, confidence limits, etc.)



Spatial accuracy

- For lines or areas:
  - error is a mixture of *positional error* (error in locating well-defined points along the line) and *generalization error* (error in the points selected to represent the line)
  - The *epsilon band* is usually used to define a zone of uncertainty around the encoded line, within which "actual" line exists with some probability. However, there is little agreement on the shape of the band, both planimetrically and in cross-section.



Temporal accuracy

- Temporal accuracy is the agreement between the encoded and *valid time* (time interval *valid for an entity)*:
  For example, pothole Q54D-35-021 existed between 2/12/96 and 8/9/96
- Valid time is often implicit in geodata
- Valid time is not the same as "database time", which is the time the information was entered into the database.
- Temporal accuracy is not the same as "currentness" (or up-to-dateness) which is actually an assessment of how well the database specification meets the needs of a particular application.

Thematic accuracy

Thematic accuracy is the accuracy of the attribute values encoded in a database.

- Quantitative data (e.g., precipitation) can be treated like a z-coordinate (elevation) and assessed with usual metrics (such as the RMSE)
- Qualitative data (e.g., land use/land cover) is assessed using a crosstabulation Percent Correctly Classified (PCC)

Table: Error/ Confusion Matrix

Sample Data	Reference Data						Total
	Exposed soil	Cropland	Range	Sparse woodland	Forest	Water	
Exposed soil	1	2	0	0	0	0	3
Cropland	0	5	0	2	3	0	10
Range	0	3	5	1	0	0	9
Sparse woodland	0	0	4	4	0	0	8
Forest	0	0	0	0	4	0	4
Water	0	0	0	0	0	1	1
Total	1	10	9	7	7	1	35

PCC= (1+5+5+4+4+1)*100/35= 57.1%

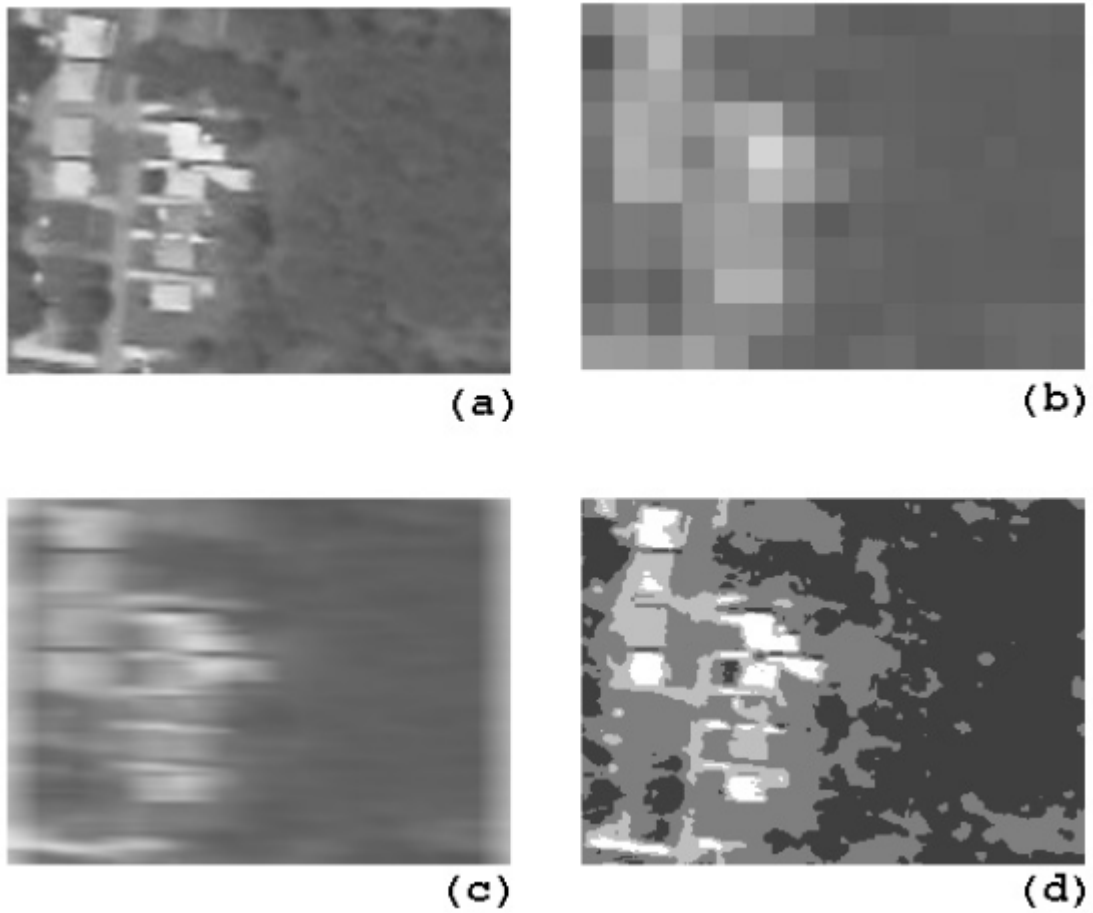
Resolution

Resolution (or precision) refers to the amount of detail that can be discerned in space, time or theme.

- Resolution is always finite because no measurement system is infinitely precise, and because databases are intentionally generalized to reduce detail
- Resolution determines how useful a given database may be for a particular application.
- High resolution is not always better; low resolution may be desirable when one wishes to formulate general models.

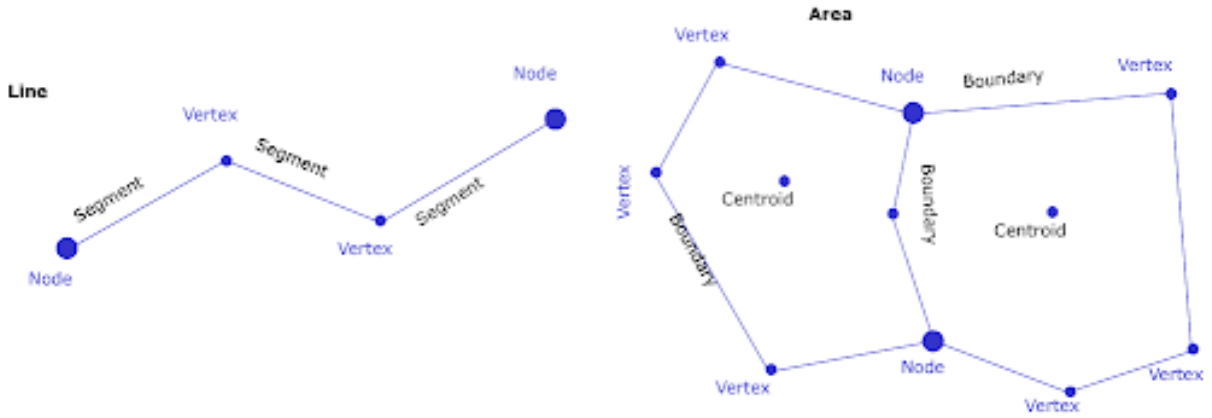
Spatial resolution

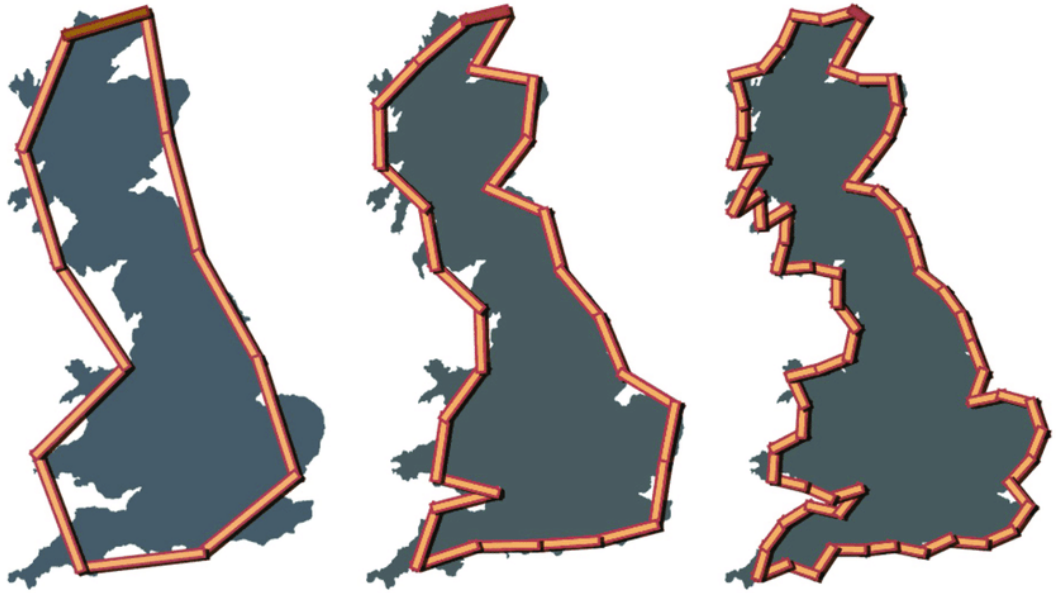
- For raster data: refers to the **linear size of a cell**
- In the maps on the right, resolution increases from map b, over c, d to a



Spatial Resolution

- For vector data, resolution can be defined as the **minimum mapping unit size**. Note: Mean polygon size is erroneous since smaller polygons may be observable but just not present on the map.
- The coastline paradox: Great Britain in different resolutions, leaving open the question of its precise circumference



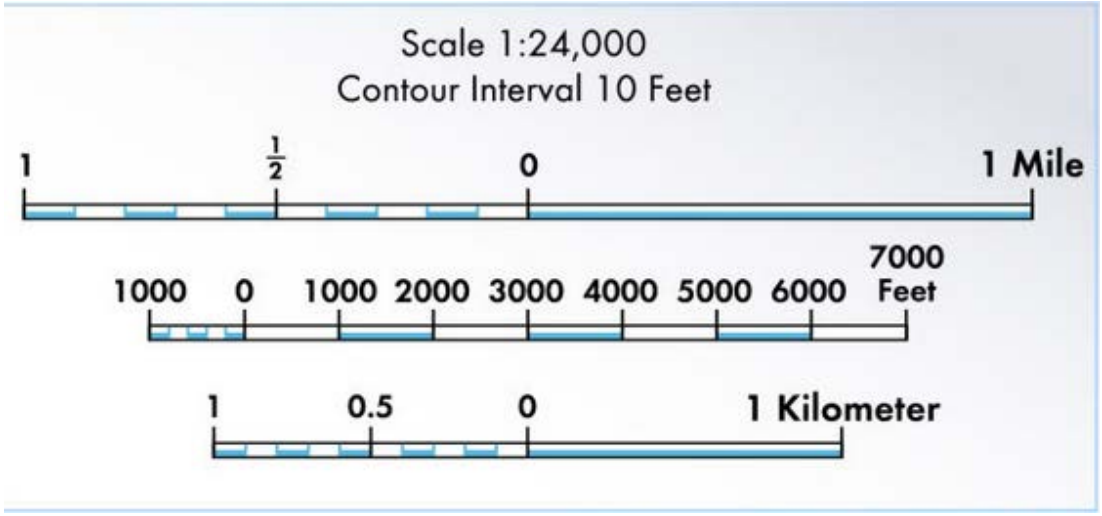


Spatial resolution and map scale

Notice how coastline, roads, labels, etc. become more detailed as the map is zoomed in (https://www.axismaps.com/guide/)

Spatial resolution and map scale

- Resolution and map scale of a map display ("zoom level" of a map) (= the ratio of geometric lengths displayed / lengths on the Earth surface) are closely related.
- Also, geodata of different resolution should not be combined.



Map scale and effective resolution

Scale	Effective Resolution (m)
1:2500	1.25
1:10000	5
1:24000	12
1:50000	25
1:100000	50
1:250000	125
1:500000	250
1:1000000	500
1:10000000	5000

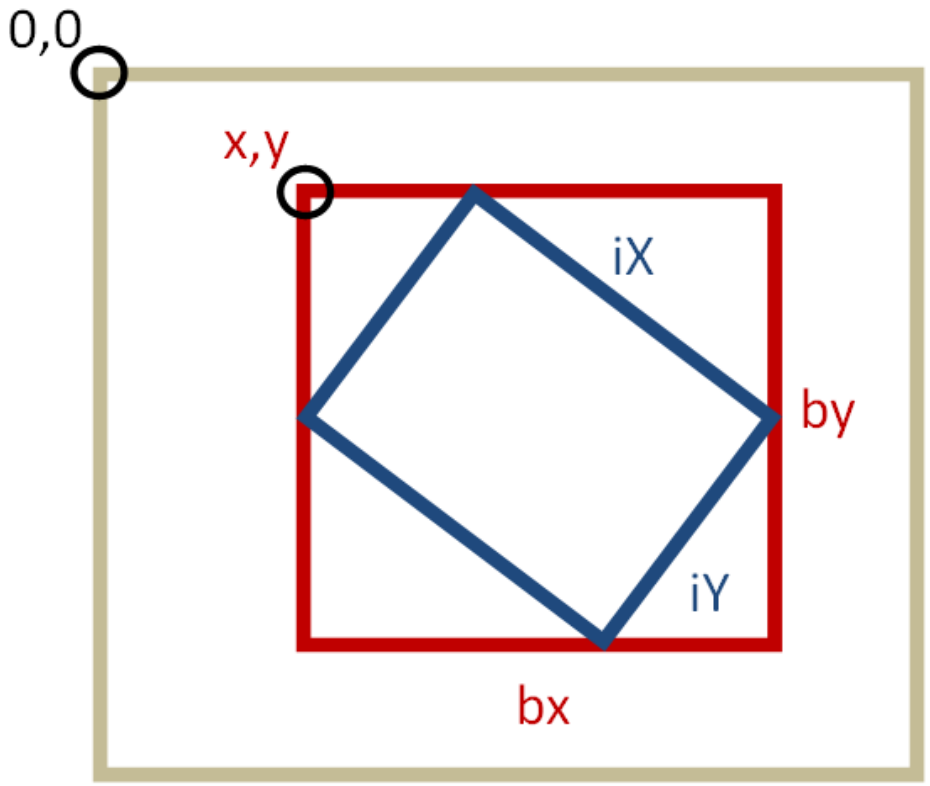
Completeness

Consider a database of buildings in Minnesota that have been placed on the National Register of Historic Places as of the end of 1995.

- Spatial incompleteness (“only buildings in Hennepin County”)
  - Spatial extent of layer
- Temporal incompleteness (“only buildings of June 30, 1995”)
  - Temporal extent of a layer
- Thematic completeness (“only residential buildings”)
  - Coverage (are all objects that exist within extents are covered?)
  - Are all existing qualities covered? (For example, all attributes of a building?)

Completeness: spatial extent

Minimum bounding rectangle (MBR) is often used to specify the extent of objects as well as map layers



Questions?
(Q&A session)

References

- Veregin H and Hargitai P (1995) An evaluation matrix for geographical data quality. In Guptill S C and Morrison J L (eds) Elements of spatial data quality. Oxford: Elsevier 167-188.
- Janowicz et al (2016): Moon Landing or Safari? A Study of Systematic Errors and their Causes in Geographic Linked Data. Proceeedings of GISCience 2016
- Li, L., & Goodchild, M. F. (2011). An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, *2*(4), 309-328
- Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120