

Uncertainty and Data Quality

SIE 510 GIS Applications

March 26, 2009

Spatial Data Quality

Why should we care about spatial data quality?

Increased data production by the private sector, where there are no required quality standards. Production of data by national mapping agencies (e.g., US Geological Survey, British Ordnance Survey) had to conform to national accuracy standards (i.e., mandated quality control).

Increased use of GIS for decision support, - the implications of using low-quality data become more widespread (including the possibility of litigation if minimum standards of quality are not met).

Any application will have requirements for some level of data quality- need to know whether available data will support the application

Spatial Data Quality Standards

How does one specify the quality of spatial datasets?

ISO and FIPS standards identify five components of spatial data quality:

- positional accuracy
- attribute accuracy
- logical consistency
- completeness
- lineage

Statistical Models for Uncertainty

Methods exist for describing errors in observations and measurements

These can be applied to GIS if we consider GIS databases as collections of measurements.

Two cases:

- Nominal data
- Interval/Ratio data

Accuracy

The closeness of results, computations or estimates to true values (or values accepted to be true).

Since spatial data is usually a generalization of the real world, it is often difficult to identify a true value, and we work instead with values which are accepted to be true

Precision

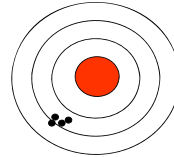
Repeatability or conformity among a set of measurements.

The number of decimal places or significant digits in a measurement

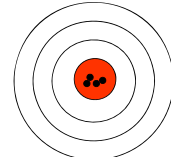
Precision is not the same as accuracy - a large number of significant digits does not necessarily indicate that a measurement is accurate

A GIS works at high precision, generally much higher than the accuracy of the data itself

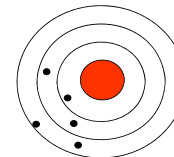
Accuracy and Precision



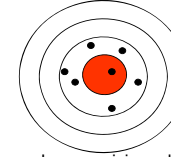
High precision – low accuracy



High accuracy and precision



Low precision – low average accuracy



Low precision – high average accuracy

Reporting Precision

The number of digits used to report a measurement should reflect the measurement's accuracy.

A measure of 10.5 meters suggests accuracy to a tenth of a meter. 100.05 suggests accuracy to a hundredth of a meter.

GIS measurements carry excess digits that should be removed by rounding

Accuracy

All spatial data are inaccurate to some degree

Important questions are:

- how to measure accuracy
- how to track the way errors are propagated through GIS operations
- how to ensure that users don't ascribe greater accuracy to data than it deserves

Positional Accuracy

Accuracy of the spatial component of the database. The metrics used depend on the dimensionality of the entities under consideration.

Accuracy reporting for points

- Accuracy is defined in terms of the distance between the encoded location and the "actual" location.
- Truth considered to be an independent source of higher accuracy
- Error can be defined in various dimensions: x, y, z
- Metrics of error are extensions of classical statistical measures (mean error, RMSE (root mean squared error), inference tests, confidence limits, etc.)

Positional Accuracy

Measurements of x and y are both subject to error.

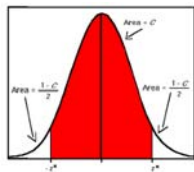
If 100 people were asked to measure the coordinates of a road intersection on a topographic map, the results would show variation in both coordinates.

The variation in both coordinates typically follows a normal distribution, with most measurements clustered, and a few extremes in both positive and negative directions.

The amount of variation is likely to be similar in both coordinate directions. Errors are likely to be uncorrelated, in the sense that the direction and amount of error in one coordinate is independent of the direction and amount of error in the other.

Errors are also likely to be unbiased, in the sense that the average of all 100 measurements will be very close to the true location. If these assumptions are true the errors in both coordinates can be visualized as a three-dimensional bell curve, bivariate Gaussian, or circular normal distribution.

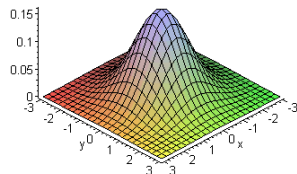
Positional Accuracy



Univariate Case

- $1 * \sigma = 68$ percent
- $1.96 * \sigma = 95$ percent

A 95% confidence interval covers 95% of the normal curve



Bivariate Case

- $1 * \sigma = 39$ percent
- $2.146 * \sigma = 90$ percent
- $2.45 * \sigma = 95$ percent

Positional Accuracy

The circular normal distribution is commonly used to describe positional accuracy.

The Circular Map Accuracy Standard (CMAS) is defined as the 90th percentile of the circular normal distribution, or 2.146 times its standard deviation.

Graphically, it forms a circle about the true location of the point, within which the observed location is expected to lie 90% of the time.

Using the example of the topographic map, it might turn out that 90% of the 100 people determined the road intersection's coordinates to within 0.5 millimeter of their true locations at the scale of the map, leaving 10 people with positional errors of more than 0.5 millimeter.

Testing Positional Accuracy

Use of an independent source of higher accuracy

- A larger scale map
- Global Positioning System (GPS) observations
- Independent survey data

Establish a set of test points

Compare to corresponding data points

Testing Positional Accuracy

ID	X(true)	X(data)	delta X	deltaxsq	Y (true)	Y (data)	Delta y	delta y sq	sum dx+dy
1	12	10	-2	4	288	292	-4	16	20
2	18	22	-4	16	234	228	6	36	52
3	7	12	-5	25	265	266	-1	1	26
4	34	34	0	0	243	240	3	9	9
5	15	19	-4	16	291	287	4	16	32
6	33	24	9	81	211	215	-4	16	97
7	28	29	-1	1	267	271	-4	16	17
8	7	12	-5	25	273	268	5	25	50
9	45	44	1	1	245	244	1	1	2
10	110	99	11	121	221	225	-4	16	137
									442
									44.2
							RMSE		6.6483081

$$RMSE = \sqrt{\sum e^2 / n} \quad \text{Square root of averaged square error}$$

Useful to also display the distribution – expectation is normally distributed with zero mean

Testing Positional Accuracy

Use of internal evidence

unclosed polygons, lines which overshoot or undershoot junctions, are indications of inaccuracy

the sizes of gaps, overshoots and undershoots may be used as a measure of positional accuracy

Testing Positional Accuracy

Compute accuracy from knowledge of the errors introduced by different sources

1 mm in source document

0.5 mm in map registration for digitizing

0.2 mm in digitizing

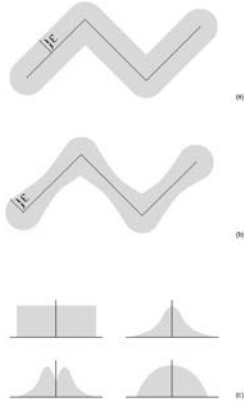
if sources combine independently, we can get an estimate of overall accuracy by summing the squares of each component and taking the square root of the sum

Positional accuracy of lines and areas

More complex - 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 a line, within which the "actual" line exists with some probability.

There is little agreement on the shape of the band, - planimetrically and in cross-section



Positional error and autocorrelation

Spatial autocorrelation is the tendency of values close by to be more similar to each other.

Errors in position measurements may be spatially autocorrelated depending on the measurement process

Spatial autocorrelation effectively reduces the degrees of freedom in spatial data.

Attribute accuracy

The accuracy of the attribute values encoded in a database

The metrics used depend on the level of measurement of the data:

Interval/ratio data (e.g., precipitation) can be treated like a z-coordinate (elevation) and assessed using metrics normally used for vertical error (such as the RMSE).

Nominal data (e.g., land use/land cover) is normally assessed using a cross-tabulation of encoded and "actual" classes at sample of locations.

Attribute accuracy

For categorical attributes such as classified polygons

Are the categories appropriate, sufficiently detailed and defined?

Gross errors, such as a polygon classified as A when it should have been B, are simple but unlikely

e.g. land use is shopping center instead of golf course

more likely the polygon will be heterogeneous:

e.g. vegetation zones where the area may be 70% A and 30% B

A and B may not be well-defined, may not be able to identify the class clearly as A or B

e.g. soils classifications are typically fuzzy

Reporting Attribute Accuracy

Categorical attribute accuracy typically reported through a confusion matrix

Based on a set of test points or regions

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	9	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

Rows correspond to classes recorded in the database

Columns correspond to classes identified in the field or source of higher accuracy

Reporting Attribute Accuracy

Summarizing the classification error matrix

The proportion of cases in the diagonal represents the proportion correctly classified.

The rows indicate for a supposed class – what the actual classes are.

The columns report the actual classes as they have been assigned in the database.

For the table as a whole, the proportion of entries in the diagonal are the percent correctly classified (PCC). $209/304 = 68.8$ percent

Reporting Attribute Accuracy

Since chance alone can produce some correct classifications need an adjustment – kappa index

The sum of rows times sum of columns divided by grand total for each diagonal cell

$$\kappa = \frac{\sum_{i=1}^n c_{ii} - \sum_{i=1}^n c_i c_i / c_{..}}{c_{..} - \sum_{i=1}^n c_i c_i / c_{..}}$$

kappa is 1 for perfectly accurate data (all N cases on the diagonal), zero for accuracy no better than chance

We expect attribute accuracy to vary over the map, so it would be useful to have an indication of the spatial variation in misclassification probability, not just a summary statistic

Reporting Attribute Accuracy

Per-class accuracies useful, but need to be differentiated into producers' and users' accuracy (also referred to as errors of omission and errors of commission).

User accuracy: assumes the user wants to know how often a category is misclassified – computed as diagonal over the row total

Producer accuracy: assumes the produce knows the class and is interested in how often the correct class is assigned in the database. computed as diagonal over the column total

	A	B	C	D	E	Total
A	80	4	0	15	7	106
B	2	17	0	9	2	30
C	12	5	9	4	8	38
D	7	8	9	65	0	80
E	3	2	1	6	38	50
Total	104	36	10	99	55	304

75%

81%

77%

66%

Reporting Attribute Accuracy

Interval/Ratio case

Magnitude of errors described by RMSE (root mean square error)

Consistency

Refers to the absence of apparent contradictions in a database. measures the internal validity of a database, and is assessed using information that is contained within the database.

Spatial consistency includes topological consistency, or conformance to topological rules, e.g., all one-dimensional objects must intersect at a zero-dimensional object.

Temporal consistency is related to temporal topology, e.g., the constraint that only one event can occur at a given location at a given time.

Thematic consistency refers to a lack of contradictions in redundant thematic attributes. For example, attribute values for population, area, and population density must agree for all entities.

Completeness

Refers to a lack of errors of omission in a database. It is assessed relative to the database specification, which defines the desired degree of generalization and abstraction (selective omission).

There are two kinds of completeness

Data completeness - a measurable error of omission observed between the database and the specification. A generalized database can be data complete if it contains all of the objects described in the specification.

Model completeness - the agreement between the database specification and the abstract universe that is required for a particular database application. A database is model complete if its specification is appropriate for a given application.

Resolution

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.

Spatial resolution is well-defined in the context of raster data - it refers to the linear dimension of a cell.

For vector data resolution might be defined as the minimum mapping unit size.

Thematic Resolution

Thematic resolution refers to the precision of the measurements or categories for a particular theme.

For categorical data, resolution is the fineness of category definitions (e.g., urban vs. residential and commercial).

For quantitative data, thematic resolution is analogous to spatial resolution in the z-dimension (i.e., the degree to which small differences in the quantitative attribute can be discerned).

Lineage

Record of the data sources and of the operations which created the database

how was it digitized, from what documents?

when was the data collected?

what agency collected the data?

what steps were used to process the data?

What is the precision of computational results?

Often a useful indicator of accuracy