Error Management in Data Capture

Objectives (Entry)

This unit will indicate the checklist to check the quality of data. Moreover, the unit will discuss the difference between accuracy and precision of data. Moreover, the unit will explain the establishment of the data quality standard and documenting the data quality. At the end of the unit the student will apply the knowledge of the lesson metadata and quality manage the error in GIS data capture.

Managing Errors in Data capture (Clarification)

1. Checking quality of Data

The documentation of data quality is very important. It is said that undocumented data set is worthless. Documentation of data will be discussed more detail in the section of Metadata. The documentation of some data set is very extensive and is published separately.

The followings should be illustrated and at least to be checked.

1. Currency of data

2. Source of Data

3. Coverage of data

4. Projection System, Coordinates and datum information

5. The original scale of map

6. Positional Accuracy

7. Attribute Accuracy

8. Logical consistency of data

9. Readability of cartographic representation

10. Relevance of data to the GIS project

11. Data format

12. Method of data compilation

13. Reliability of data provider

14. The medium of original document

15. Level of detail of categories

16. Density of observation

It is important to be aware that sometimes the costs of using and converting publicly and commercially available digital files outweight their value.

Although extensive digital data source is available, the in house data conversion might be necessary if the available data does not meet the specific requirements of the project.

It is wise to sample datasets and testing them in project work is critical. If the data does not perform the specific requirement, the data conversion process should be decided because it is time and cost consuming process.

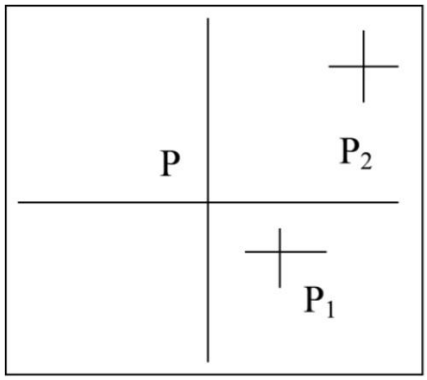
2. Accuracy and Precision

The key issue is accuracy and precision of data.

Accuracy is indeterminable. (Department of Surveying Course Notes, University of Otago)

This is because there are errors associated with measurements that cannot be fully determined. For instance, in the finite time taken to measure a quantity (or position), one cannot make the assumption that the quantity (or environment around the position) has remained stable during the measurement.

Accuracy



Accuracy of the measurement is the nearness of that measurement to truth. P is a true position or more correctly known position. Position P1 and P2 are measurements of position P. P1 is considered to be more accurate than position P2.

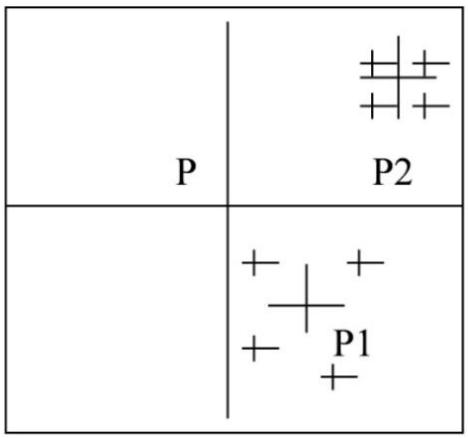
Precision

Precision is defined as the spread (or dispersion) of the measured values of a quantity.

The position P1 (and similarly P2) is the average of four measurements, as denoted by the smaller crosses close to each of the larger crosses.

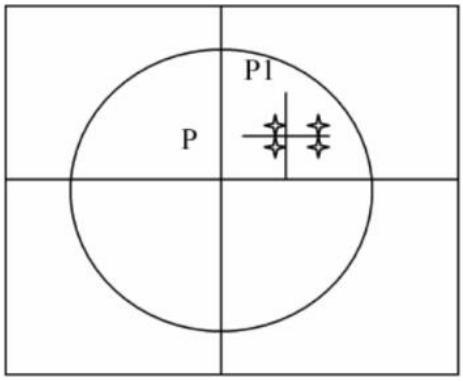
P2 is more precise than P1, as the distance of each point from its corresponding average is less than in P2 than P1. However P2 is less accurate than P1, with respect to known position P.

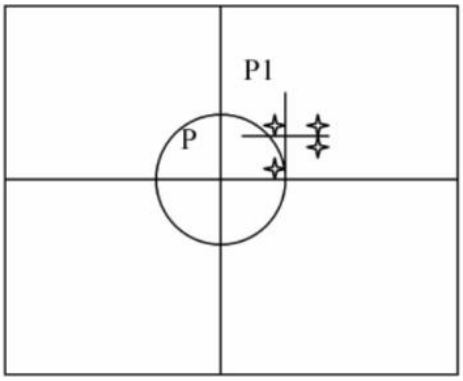
The need to understand that precision provides an indication of the quality of data but does not provide an indication of the accuracy of data.



Acceptable and unacceptable known position

In order to determine the accuracy of the position, the measured position must be compared with a reliable known position. The reliability of the known position must be an equal or better level of reliability than that of the measured position; otherwise the comparison is not valid.

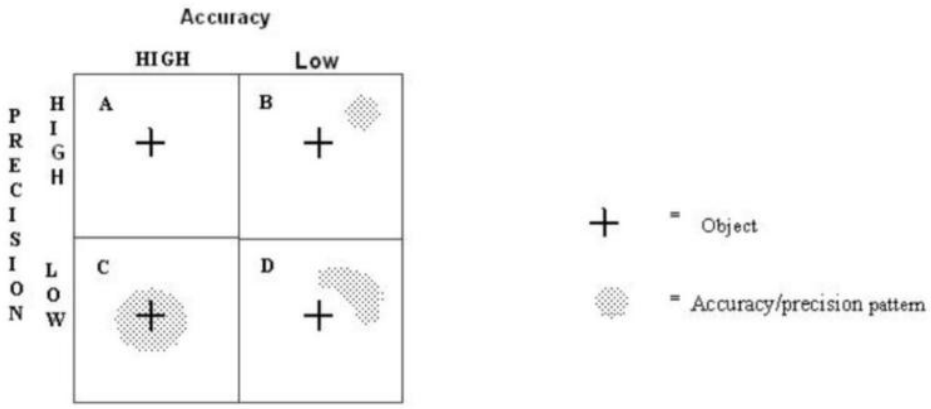




P is known and has a level of reliability as illustrated by the circle, while P1 is measured and has a level of reliability that is determined by the position precision. Position P1's level of reliability is denoted by the spread of the smaller crosses close to P1.

Assuming that the precision of P1 is an appropriate indicator of the position's reliability, the known position in first figure is unacceptable for the determination of the accuracy of P1. This is because P1 could be moved anywhere within the area defined by the level of reliability for P, which could in fact, result in an increase of the distance between the known position and the measured position, thus decreasing its accuracy. The known position in second figure, however, has a level of reliability such that if P1 where to be moved to be within the area that

defines the known position then the distance between P and P1 would be decreased, or the accuracy of P1 would be improved.



Precision and accuracy together define the reliability of the database coordinates. The graphic illustrates the various relationships we can have.

- A is the best possible combination: highly accurate data, very precisely measured.

- B represents low accuracy, but very precise measurement

- C represents low precision, but very accurate data

- D is the worst case: low accuracy and low precision

Accuracy is synonymous with absolute accuracy, while precision is synonymous with relative accuracy.

Absolute accuracy requires measured data to be referenced in terms of known or existing data

Relative accuracy does not need to be referenced to other data in order for measurements to be correct. This is usually the case where data has been collected as a completely self-contained survey.

The spatial concept of precision and accuracy can be applied to attribute. The attribute accuracy will be the degree to which attribute information on a map or in a digital database matches true or accepted value. The attribute precision refers to the level of measurement and exactness of descriptions in a GIS database. It is important to realize that precise attribute data, no matter how carefully

measured may be inaccurate. Surveyors may make mistakes or data may be entered into the database incorrectly.

High precision does not indicate high accuracy. High accuracy data does not indicate high precision. High accuracy and high precision are both costly.

3. Establishing data quality standard

Standard for the procedures to create both spatial and non-spatial data and, standard for the final products should be established from the start.

The standards should be resolved the accuracy and precision of spatial and non-spatial data, conventions for naming geographic features, criteria for classifying data etc.

The standards should reflect the demands of accuracy, precision and completeness of spatial and non-spatial data of project in the light of ultimate project goal. Standards such as USGS Glossarial Data Standards, Spatial Data Transfer Standard and USGS Map Accuracy Standards etc. already existed for wide range of mapping and GIS applications. These standards should be considered as a guideline and modified regarding specific project requirements.

The staff who will be compiling and entering data must aware the standards. For example, the digitizing staff must aware and strictly follow the allowable RMS of the project in order to create the spatial database spatially accurate enough in order to meet the project demand. Quantitative method to calculate the allowable RMS will be illustrated in GIS Data Quality section.

Throughout the project, all data should be regularly spot checked and tested in order to make sure that the standards are followed and to point out and correct the error at the early stage.

4. Documenting the Data Quality

Document the database dictionary

The following information about the data set in the spatial database should be documented.

- Feature name

- Feature Source

- Source Organization

- Source organization types

- Source format

- Date of the data

- Source medium

- GIS input method

- GIS format

- Description

Document the Positional Accuracy

Document the positional accuracy based on national map accuracy standard, the distance error of each tested point, average distance error, standard deviation (S.D) of distance error, root mean square error (RMS) and accuracy at different confident level would be documented.

The aforementioned positional accuracy indicator can be derived by using an independent source of higher accuracy such as larger scale map, differentially corrected GPS data and raw survey data. Moreover, use the graphic display and editing facilities of GIS to check and correct the unclosed polygons, undershoot and overshoot dangle errors. Moreover compute the accuracy based on knowledge of the errors introduced by different sources, e.g.

- 1 mm in source document

- 0.25 mm in map registration for digitizing

- 0.2 mm in digitizing

If sources combined independently, overall accuracy can be illustrated as follow.

$(1^2 + 0.25^2 + 0.2^2)^{1/2} = 1.05$ mm

Documenting the Attribute Accuracy

It is important to not that while location does not change with time, attributes often do. Attribute accuracy must be analyzed in different ways depending on the nature of the data.

The attribute accuracy of continuous data such as Digital Elevation Model, is expressed as measurement error such as elevation accuracy to 1 m. Residual analyses which will be explained in Intermediate level, can be applied to measure the accuracy of continuous data.

The attribute accuracy of categorical data can be expressed as

- appropriateness of categories

- are the categories sufficiently detailed or defined

- classification errors such as pasture is classified as dry land paddy field

- heterogeneity of polygon attributes such as mixed agriculture where the are may be 70% of wetland paddy and 30% of dry land agriculture (such attribute may not appropriate to analyze the national food sufficiency analyses based on wetland paddy)

- categorical classes may not be well defined such as soil classes are typically fuzzy.

Attribute accuracy can be test and documented by using the error matrix and kappa statistics. The overall accuracy, producer's accuracy and user's accuracy, kappa statistics and error matrix would be documented.

Documenting other information

The key issues involving data sources (except positional accuracy and attribute accuracy), which are discussed in Quality and coverage unit, would be documented. These reflects the data quality apply to the database as a whole, rather than to the objects, attributes or coordinate within it.

5. Managing Errors

Error Propagation

In GIS application, data from different sources with different level of accuracy are combined. It is necessary to manage the error in each data layer on the final result. The effect of error from each data layer to the final result will be complex.

When two layers with similar level of accuracy are overlaid, the accuracy is little better than the original layers. However, many layers are overlaid, the accuracy of resulting composite can be very poor.

In the overlay analyses, if AND operation and Reclassification are extensively used, the accuracy of final analyses result is determined by the accuracy of the least accurate layer. In other cases such as using OR operation, the accuracy of the result is significantly better than the accuracy of least accurate layer.

Sensitivity Analyses

Sensitivity analyses is the response of result on how much the result change when the data input change and how much the result change when the weight given to a factor changes. It was illustrated well in the Geography Craft Project web page.

Calibrating a Dataset to Ascertain How Error Influences Solutions by Kenneth E. Foote and Donald J. Huebner, Department of Geography, University of Texas at Austin.

http://www.colorado.edu/geography/gcraft/notes/manerror/manerror.html

See the example of Sensitivity Analyses by Kenneth E. Foote and Donald J. Huebner, Department of Geography, University of Texas at Austin.

http://www.colorado.edu/geography/gcraft/notes/manerror/html/sensitiv.html

Managing Digitizing Errors

The snapping tolerance, snapping type, dangle length tolerance, fuzzy tolerance must be established in order to correct the gap in closing polygons, overshoot and undershoots.

If necessary enlarge the map photographically to increase the scale of map while holding tolerance constant to digitize detail of geometry. However, it is difficult or impossible to get error-free enlargement cheaply and easily.

It is important to check the topological consistency of geometry and attributes.

Allowable RMS (root mean square) of project must be set in order to maintain the positional accuracy of features.

Refer to the digitizing unit for more detail information to manage and minimize the digitizing errors.

Managing overlay errors

Sliver polygons will have to be removed unless share primitives are allowed. It is important to select the appropriate threshold value, options to keep the edge and appropriate rule to remove the arcs and to merge to the other polygon.

Data Quality Report (Look)

See the example of data quality report of Digital Line Graph (DLG) Dataset.

Data Quality Report of DLG (Digital Line Graph) (http://edcwww.cr.usgs.gov/glis/hyper/guide/2mil)

Exercises (Self assessment)

1. What should be considered to check the quality of data?

2. Accuracy of the measurement is the nearness of that measurement to truth. Yes/No

3. Precision is defined as the spread (or dispersion) of the measured values of a quantity. Yes/No

4. Precision provides an indication of the quality of data but does not provide an indication of the accuracy of data. Yes/No

5. High precision does not indicate high accuracy. High accuracy data does not indicate high precision. Yes/No

6. Absolute accuracy requires measured data to be referenced in terms of known or existing data. Yes/No

7. Relative accuracy does not need to be referenced to other data in order for measurements to be correct. This is usually the case where data has been collected as a completely self-contained survey. Yes/No

8. The digitizing staff must be well informed the allowable RMS of the project in order to maintain the positional accuracy requirement of the project. Yes/No

These questions will be modified based on WebCT quizzing tools.