Unit 100 - Data Quality Measurement and Assessment

Written by Howard Veregin, Department of Geography,University of Minnesota, Room 414 267 19th Avenue South, Minneapolis, MN 55455, USA veregin@atlas.socsci.umn.edu

This section was edited by Gary Hunter, Department of Geomatics, University of Melbourne, Australia.

Advanced Organizer

Unit Topics

- This unit covers the following topics:
  - Basic definitions of quality for geospatial data;
  - Differences between quality control and truth-in-labeling paradigms;
  - Descriptions and assessment of data quality components for geospatial databases, including accuracy, resolution, completeness and consistency.

Intended Learning Outcomes

- after reading this unit, you should be able to
  - A basic introduction to the meaning of data quality in the context of geospatial data;
  - A simple model for categorizing quality components based on geographical dimension (space, time, theme) and data quality component (accuracy, resolution, completeness and consistency).
  - Basic techniques for data quality assessment.
  - Pointers to relevant literature covering topics in more detail.

Full Table of Contents

Metadata and Revision History

Data Quality Measurement and Assessment

1. Data Quality

- What is quality?
    - Quality is commonly used to indicate the superiority of a manufactured good or to indicate a high degree of craftsmanship or artistry. We might define it as the degree of excellence in a product, service or performance.
    - In manufacturing, quality is a desirable goal achieved through management and control of the production process (statistical quality control). (Redman, 1992)
    - Many of the same issues apply to the quality of databases, since a database is the result of a production process, and the reliability of the process imparts value and utility to the database.

- Why is there a concern for DQ?
    - Increased data production by the private sector, where there are no required quality standards. In contrast, production of data by national mapping agencies (e.g., US Geological Survey, British Ordnance Survey) has long been required to conform to national accuracy standards (i.e., mandated quality control).
    - Increased use of GIS for decision support, such that the implications of using low-quality data are becoming more widespread (including the possibility of litigation if minimum standards of quality are not attained).
    - Increased reliance on secondary data sources, due to the growth of the Internet, data translators and data transfer standards. Thus, poor-quality data is ever easier to get.

- Who assesses DQ?

  Model 1. Minimum Quality Standards.
    - This is a form of quality control where DQ assessment is the responsibility of the data producer. It is based on compliance testing strategies to identify databases that meet quality thresholds defined a priori.
    - An example is NMAS, the National Map Accuracy Standards adopted by the US Geological Survey in 1946.
    - This approach lacks flexibility; in some cases a particular test may be too lax while in others it may be too restrictive.

  Model 2. Metadata Standards.
    - This model views error as inevitable and does not impose a minimum quality

standard a priori. Instead, it is the consumer who is responsible for assessing fitness-for-use; the producer's responsibility is documentation, i.e., "truth-in-labeling."
    - An example is SDTS, the Spatial Data Transfer Standard.
    - This approach is flexible, but there is still no feedback from the consumer, i.e., there is a one-way information flow that inhibits the producer's ability to correct mistakes.

  Model 3. Market Standards.

    - This model uses a two-way information flow to obtain feedback from users on data quality problems. Consumer feedback is processed and analyzed to identify significant problems and prioritize repairs.
    - An example is Microsoft's Feedback Wizard, a software utility that lets users email reports of map errors.
    - This model is useful in a market context in order to ensure that databases match users' needs and expectations.

- What are the dimensions of geographical DQ?

  The conventional view is that geographical data is "spatial". We often use the terms "geographical data" and "spatial data" interchangeably. However, this is problematic.

    - First, it ignores the inherent coupling of space and time (geographical entities are actually events unfolding over space and time)
    - Second, geography is really about theme, not space. Space (or space-time) is just the framework on which theme is measured. Without theme, we have only geometry.

  A better definition of geographical data includes the three dimensions of space, time and theme (where-when-what). These three dimensions are the basis for all geographical observation. (Berry, 1964; Sinton, 1978)

  Geographical data quality is likewise defined by space-time-theme. Data quality also contains several components such as accuracy, precision, consistency and completeness. The result is a matrix. [FIGURE 1]

---

2. Accuracy

- Accuracy is the inverse of error. Many people equate accuracy with quality but in fact accuracy is just one component of quality (See Figure 1).

- Definition of accuracy is based on the entity-attribute-value model [FIGURE 2]

    - Entities = real-world phenomena
    - Attribute = relevant property

- Values = Quantitative/qualitative measurements

- An error is a discrepancy between the encoded and actual value of a particular attribute for a given entity. "Actual value" implies the existence of an objective, observable reality. However, reality may be:

  - Unobservable (e.g., historical data)
  - Impractical to observe (e.g., too costly)
  - Perceived rather than real (e.g., subjective entities such as "neighborhoods")

- In fact, it is not necessary to posit an objective reality in order to assess accuracy, since all geographical data are collected with the aid of a model that specifies -- implicitly or explicitly -- the required level of abstraction and generalization.

  - This is the database "specification" and is closely related to the "terrain nominal" concept of perceived reality (Salgé, 1995).
  - 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 (Brassel et al., 1995).
  - Accuracy is always a relative measure, since it is always measured relative to the specification.
  - To judge fitness-for-use, one must judge the data relative to the specification, and also consider the limitations of the specification itself (CEN, 1995).

2.1. Spatial Accuracy

- Spatial accuracy is the accuracy of the spatial component of the database. The metrics used depend on the dimensionality of the entities under consideration.

- For points, accuracy is defined in terms of the distance between the encoded location and "actual" location.

  - Error can be defined in various dimensions: x, y, z, horizontal, vertical, total [FIGURE 3].
  - Metrics of error are extensions of classical statistical measures (mean error, RMSE or root mean squared error, inference tests, confidence limits, etc.) (American Society of Civil Engineers 1983; American Society of Photogrammetry 1985; Goodchild 1991a).

- For lines and areas, the situation is more complex. This is because 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) (Goodchild 1991b).

  - 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 (and little empirical work) on the shape of the band, both planimetrically and in cross-section (Chrisman, 1982; Blakemore, 1983; Honeycutt, 1986; Caspary and Scheuring, 1993). [FIGURE 4]

2.2. Temporal accuracy

- Temporal accuracy is the agreement between the encoded and "actual" temporal coordinates for an entity.
- Temporal coordinates are often only implicit in geographical data, e.g., a time stamp indicating that the entity was valid at some time. Often this is applied to the entire database (e.g., a map dated "1995").
- More realistically, temporal coordinates are the temporal limits within which the entity is valid (e.g., Pothole Q54D-35-021 existed between 2/12/96 and 8/9/96).
- Temporal accuracy 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. A database can be temporal accurate but still out of date; historical applications depend on such data.

2.3. Thematic Accuracy

- Thematic accuracy is the accuracy of the attribute values encoded in a database.
- The metrics used here depend on the measurement scale of the data:
    - Quantitative 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). See section 2.1.
    - Qualitative data (e.g., land use/land cover) is normally assessed using a cross-tabulation of encoded and "actual" classes at sample of locations. This produces a classification error matrix [FIGURE 5].
        - Element in row i, column j of the matrix is the number of sample locations assigned to class I but actually belonging to class j.
        - The sum of the main diagonal divided by the number of samples is a simple measure of overall accuracy.
        - An error of omission means a sample that has been omitted from its actual class. An error of commission means an error that is included in the wrong class. Ever error of omission is also an error of commission.
        - There is a large body of research on this topic (e.g., van Genderen and Lock, 1977; Congalton et al., 1983; Aronoff, 1985; Rosenfield and Fitzpatrick-Lins, 1986).

---

3. 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 (Veregin and Hargitai, 1995).

- Resolution is an aspect of the database specification that 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.

- Resolution is linked with accuracy, since the level of resolution affects the database specification against which accuracy is assessed. Two databases with the same overall accuracy levels but different levels of resolution do not have the same quality; the database with the lower resolution has less demanding accuracy requirements. (For example, thematic accuracy will tend to be higher for general land use/land cover classes like "urban" than for specific classes like "residential".)

3.1. Spatial Resolution

- Spatial resolution is well-defined in the context of raster data were it refers to the linear dimension of a cell. [FIGURE 6]

- For vector data resolution might be defined as the minimum mapping unit size. Sometimes mean polygon size is used instead, but this is erroneous since smaller polygons may be observable but just not present on the map.

- Resolution is distinct from the spatial sampling rate, although the two are often confused with each other.

    - Sampling rate refers to the distance between samples, while resolution refers to the size of the sample units.
    - Often resolution and sampling are closely matched, but they do not necessarily need to be. When the sampling rate is higher than the resolution, sample units overlap; when the sampling rate is lower than the resolution, there are gaps between sample units.

3.2. Temporal Resolution

- Temporal resolution is length (temporal duration) of the sampling interval.

    - For example, the shorter the shutter speed of a camera, the higher the temporal resolution (other factors being equal).
    - Temporal resolution affects the minimum duration of an event that is discernible. If the duration is less than the resolution, the event is invisible or at best leaves a smudge (like carriages on nineteenth-century daguerreotypes). [FIGURE 6]

- Temporal resolution is distinct from temporal sampling rate.

    - Resolution is the length of the sampling interval, while sampling rate is the

frequency of sampling over time (e.g., once a day, once a week, etc.).
    - For example, a motion picture camera might have a temporal resolution of 1/1000 second (i.e., the shutter speed to capture a single frame), and sampling rate of 24 frames per second.

3.3. 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). [FIGURE 6]

---

4. Consistency

- Consistency refers to the absence of apparent contradictions in a database. Consistency is a measure of the internal validity of a database, and is assessed using information that is contained within the database.
- Consistency can be defined with reference to the three dimensions of geographical data.
    - Spatial consistency includes topological consistency, or conformance to topological rules, e.g., all one-dimensional objects must intersect at a zero-dimensional object (Kainz, 1995).
    - 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 (Langran, 1992).
    - 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.
- Attribute redundancy is one way in which consistency can be assessed. For example, an entity might have the value "Delaware" for the attribute "State" but the value "Lincoln" for the attribute "County". This is inconsistent since there is no Lincoln county in Delaware [FIGURE 7].
    - In this example redundancy is partial, since the state Delaware eliminates the possibility of the county Lincoln, but the county Lincoln does not necessarily imply the state Maine, since Maine is one of twenty-four states containing a Lincoln County.
    - The identification of an inconsistency does not necessarily imply that it can be corrected.
    - The absence of inconsistencies does not necessarily imply that the data are accurate (Redman 1992).

5. Completeness

- 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 (Brassel et al., 1995)

    - "Data completeness" is a measurable error of omission observed between the database and the specification. Even highly generalized databases can be "data complete" if they contain all of the objects described in the specification.
    - "Model completeness" refers to 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.

- Incompleteness can be measured in space, time or theme . 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: The list contains only buildings in Hennepin County (one county in Minnesota, rather than all of Minnesota).
    - Temporal incompleteness: The list contains only buildings placed on the Register by June 30, 1995.
    - Thematic incompleteness: The list contains only residential buildings.

- Errors of commission can also be assessed. These errors can lead to "over-completeness".

    - Errors of commission in space, time and theme for the previous example: The list also contains buildings in Wisconsin; the list contains buildings added to the list in 1996; the list contains historic districts as well as buildings.

6. Summary of Important Points

- Data quality is the degree of excellence in a database. Quality is assessed relative to the database specification, which defines the desired level of generalization and abstraction. The quality of this specification, and its appropriateness for particular applications, can also be assessed.

- Quality assessment and reporting is based on minimum quality standards (compliance testing or quality control), metadata standards (truth-in-labeling and fitness-for-use), or market standards (feedback from users).

- Data quality is contains several components, including accuracy, precision, consistency

and completeness. Each component can be assessed in space, time and theme (the three basic dimensions of geographical data).

- Various assessment methods can be used for each component/dimension combination. Some methods are well-developed and others are not.

---

7. References and Bibliography

- American Society of Civil Engineers (Committee on Cartographic Surveying, Surveying and Mapping Division) 1983 Map uses, scales and accuracies for engineering and associated purposes. New York: American Society of Civil Engineers.

- American Society of Photogrammetry (Committee for Specifications and Standards, Professional Practice Division) 1985 Accuracy specification for large-scale line maps. Photogrammetric Engineering and Remote Sensing 51: 195-199.

- Aronoff S 1985 The minimum accuracy value as an index of classification accuracy. Photogrammetric Engineering and Remote Sensing 51: 99-111.

- Beard M K 1989 Use error: The neglected error component. Proceedings, Auto Carto 9; 808-817.

- Berry B 1964 Approaches to regional analysis: A synthesis. Annals, Association of American Geographers 54: 2-11.

- Blakemore M 1983 Generalisation and error in spatial data bases. Cartographica 21: 131-139.

- Brassel K, Bucher F, Stephan E-M and Vckovski A 1995 Completeness. In Guptill S C and Morrison J L (eds) Elements of spatial data quality. Oxford, Elsevier: 81-108.

- Burrough P A 1986 Principles of geographical information systems for land resources assessment. Oxford, Clarendon.

- Campbell W G and Mortenson D C 1989 Ensuring the quality of geographic information system data. Photogrammetric Engineering and Remote Sensing 55: 1613-1618.

- Caspary W and Scheuring R 1993 Positional accuracy in spatial databases. Computers, Environment and Urban Systems 17: 103-110.

- Chrisman N R 1982 A theory of cartographic error and its measurement in digital data bases. Proceedings, Auto Carto 5: 159-168.

- Chrisman N R 1991 The error component in spatial data. In Maguire D J, Goodchild M F and Rhind D W (eds) Geographical information systems. New York, Wiley: 165-174.

Comité Européen de Normalisation (CEN) 1995 Geographic Information - Data Description - Quality (Draft). Brussels: CEN Central Secretariat.

- Congalton R G, Oderwald R G and Mead R A 1983 Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing 49: 1671-1678.

- Duecker G T and Platt J T 1990 The role of automated data checks in the quality assurance of GIS data bases. GIS/LIS '90: 264-271.

- Federal Geographic Data Committee (FGDC) 1994 Content Standards for Digital Geospatial Metadata (June 8). Washington DC: Federal Geographic Data Committee.

- Fegeas R G, Cascio J L and Lazar R A 1992 An overview of FIPS 173, The Spatial Data Transfer Standard. Cartography and Geographic Information Systems 19: 278-93.

- Goodchild M F 1988a Stepping over the line: Technological constraints and the new cartography. The American Cartographer 15: 311-319.

- Goodchild M F 1988b The issue of accuracy in global databases. In Mounsey H (ed) Building Databases for Global Science. London, Taylor and Francis: 31-48.

- Goodchild M F 1991a Issues of quality and uncertainty In Muller J C (ed) Advances in cartography. London, Elsevier: 113-139.

- Goodchild M F 1991b Keynote address. Proceedings, Symposium on Spatial Database Accuracy: 1-16.

- Goodchild M F 1995 Sharing imperfect data. In Onsrud H J and Rushton G (eds) Sharing geographic information. New Brunswick NJ, Center for Urban Policy Research: 413-425.

- Guptill S C 1993 Describing spatial data quality. Proceedings, 16th International Cartographic Conference: 552-560.

- Honeycutt D M 1986 Epsilon, generalization and probability in spatial data bases. Unpublished manuscript.

- Kainz W 1995 Logical consistency. In Guptill S C and Morrison J L (eds) Elements of spatial data quality. Oxford, Elsevier: 109-137.

- Langran G 1992 Time in geographic information systems. London: Taylor and Francis.

- Lanter D 1991 Design of a lineage-based meta-database for GIS. Cartography and Geographic Information Systems 18(4): 255-261.

- Lanter D and Veregin H 1992 A research paradigm for propagating error in layer-based GIS. Photogrammetric Engineering and Remote Sensing 58: 526-533.

Moellering H (ed) 1991 Spatial database transfer standards: Current international status. London: Elsevier.

- Parkes D N and Thrift N J 1980 Times, spaces, and places: A chronogeographic perspective. New York: Wiley.
- Redman T C 1992 Data quality. New York: Bantam.
- Rosenfield G H and Fitzpatrick-Lins K 1986 A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing 52: 223-227.
- Salgé F 1995 Semantic accuracy. In Guptill S C and Morrison J L (eds) Elements of spatial data quality. Oxford, Elsevier: 139-151.
- SDTS 1992 The Spatial Data Transfer Standard (FIPS-173).
- Sinton D 1978 The inherent structure of information as a constraint in analysis. In Dutton G (ed) Harvard papers on geographic information systems. Reading MA, Addison-Wesley.
- Stearns F 1968 A method for estimating the quantitative reliability of isoline maps. Annals, Association of American Geographers 58: 590-600.
- Thapa K and Bossler J 1992 Accuracy of spatial data used in geographic information systems. Photogrammetric Engineering and Remote Sensing 58(6): 835-841.
- Tychon G G and Johnson M R 1990 GIS data exchange: Standards and formats. In Heit M and Shortreid A (eds) GIS applications in natural resources. Boulder CO, GIS World Inc: 155-161.
- van Genderen J L and Lock B F 1977 Testing land-use map accuracy. Photogrammetric Engineering and Remote Sensing 43: 1135-1137.
- 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.

Citation

To reference this material use the appropriate variation of the following format:

Howard Veregin,(1998) Data Quality Measurement and Assessment, *NCGIA Core Curriculum in GIScience*, http://www.ncgia.ucsb.edu/giscc/units/u100/u100.html, posted March 23, 1998.

Created March 23, 1998.

Unit 100 - Data Quality Measurement and Assessment

Table of Contents

Back to the Unit

Unit 100 - Data Quality Measurement and Assessment

Metadata and Revision History

1. About the main contributors

- Written by Howard Veregin, Department of Geography,

  University of Minnesota, Room 414,
  267 19th Avenue South, Minneapolis, MN 55455, USA

2. Details about the file

- unit title
  - Data Quality Measurement and Assessment
- unit key number
  - 100

3. Key words

4. Index words

5. Prerequisite units

6. Subsequent units

7. Other contributors to this unit

8. Revision history

- Created: March 23, 1998

Back to the Unit.

	Space	Time	Theme
Accuracy			
Precision			
Consistency			
Completeness			

FIGURE 1. Matrix showing geographical dimensions (columns) and data quality components (rows).

Name	Width (ft)	Cover	Speed (kph)
Belmont Rd.	36	asphalt	60
Latrobe St.	22	concrete	50
etc...			

FIGURE 2. Example of E-A-V model showing entities (e.g., Belmont Rd.), attributes (e.g., Width) and values (e.g., 36 ft.).

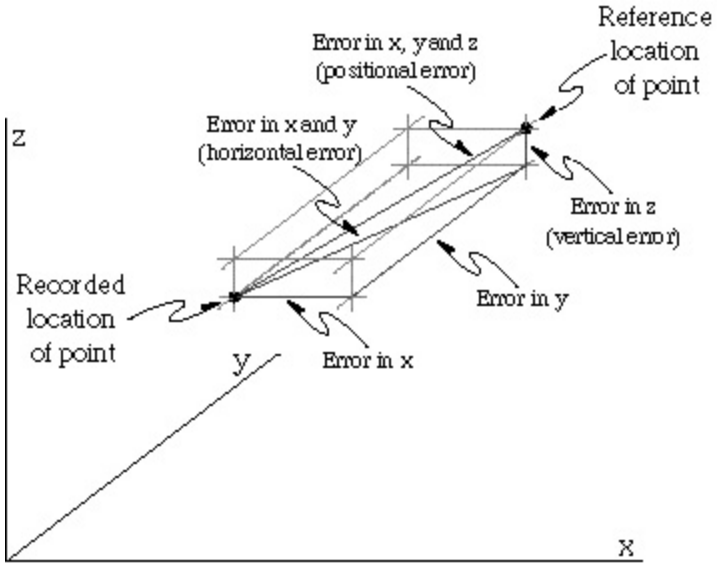
Reference
location
of point
Error in x, y and z
(positional error)
Error in x and y
(horizontal error)
z
Error in z
(vertical error)
Recorded
location
of point
Error in y
Error in x
y
x

Figure 3.

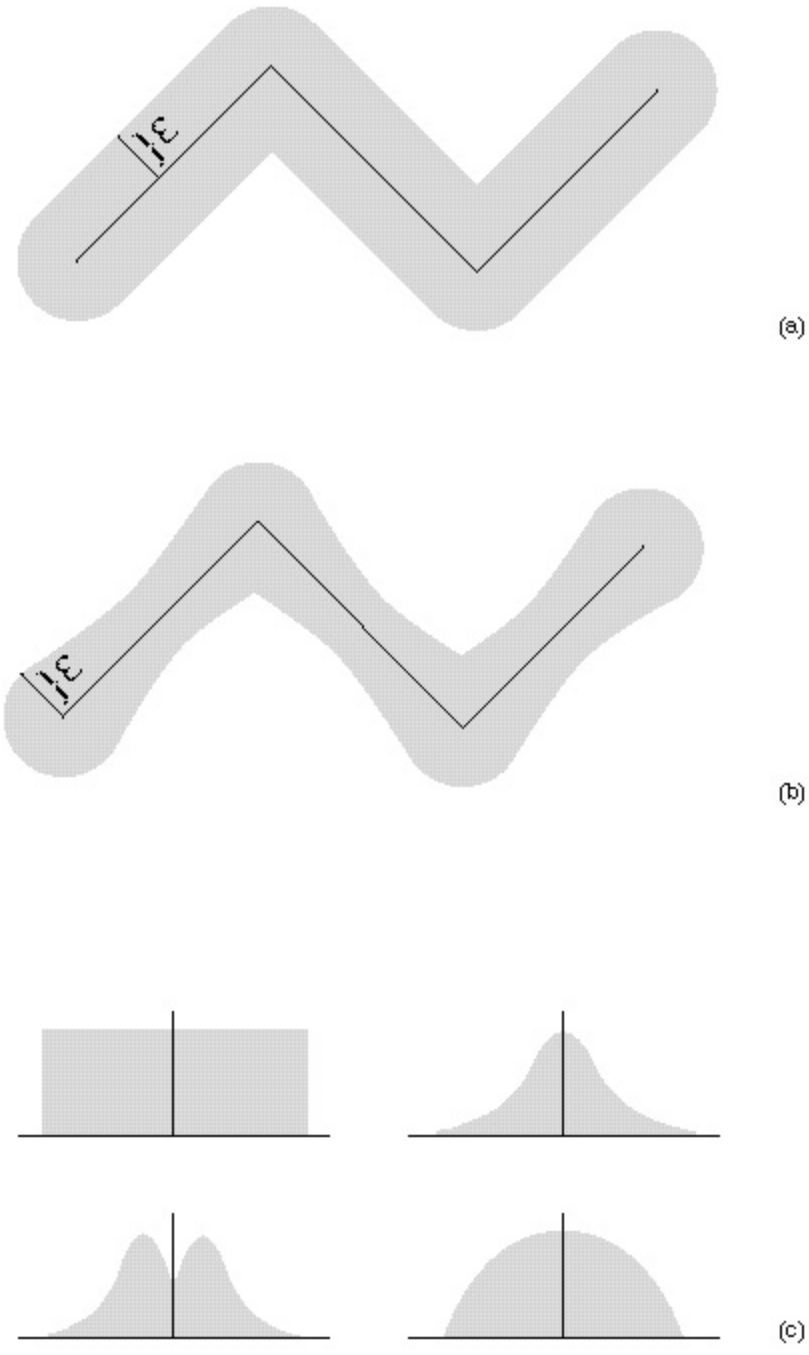

Figure 4.

	Water	Soil	Veg	TOTAL
Water	25	2	3	30
Soil	0	38	2	40
Veg	1	4	25	30
TOTAL	26	44	30	100

FIGURE 5. Example of classification error matrix.

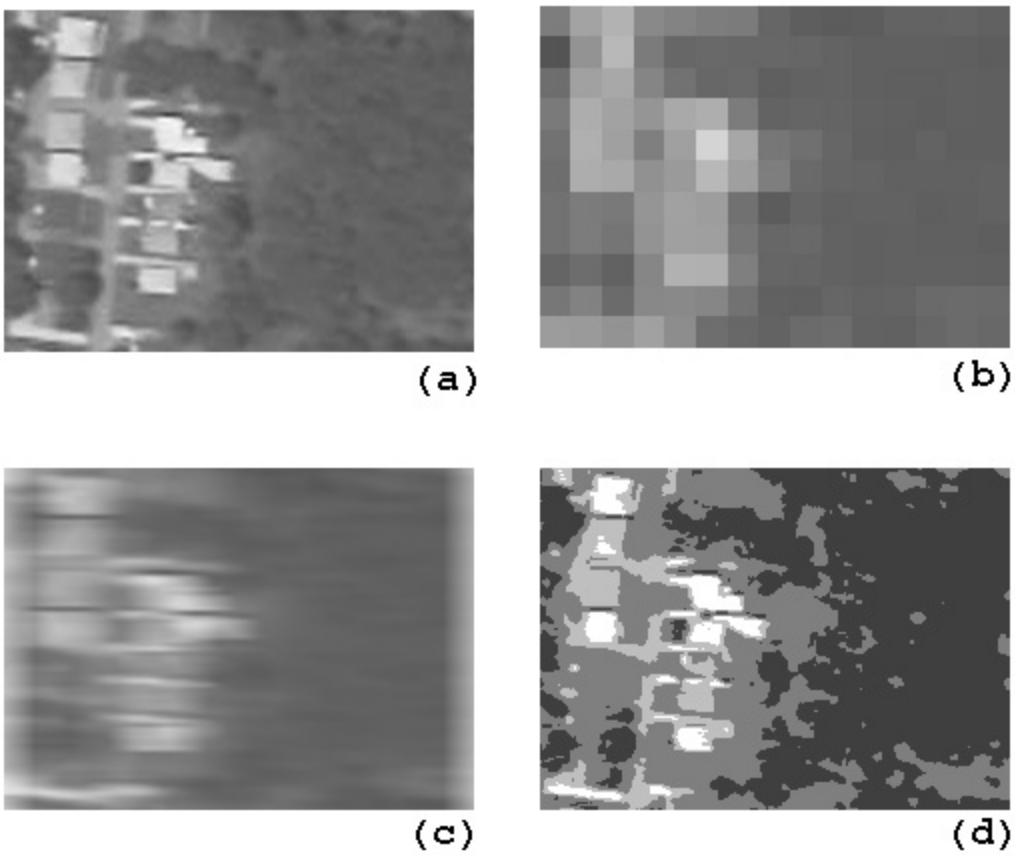

Figure 6.

Entity ID	County	State
1	Lincoln	Delaware
...	...	...

FIGURE 7. Redundancy in attributes.