2006, Angewandte Geoinformatik – AGIT (Stobl, Blaschke & Griesebner, eds.), Wichmann, pp: 494-503

What Does Data Quality Mean? An Ontological Framework

Gerhard NAVRATIL and Andrew U. FRANK

Abstract

There is a vast amount of literature on quality of spatial data. The terms used in the publications include quality, uncertainty, or fitness for use, which all have detailed definitions. The connections between these terms and the reason for their occurrence, however, are not so clear. In this paper we propose a framework that connects all those terms.

Zusammenfassung

Es gibt eine große Anzahl von Publikationen zur Qualität räumlicher Daten. Die darin verwendeten Begriffe sind unter anderem Qualität, Unsicherheit oder Nutzbarkeit, die alle eine genaue Definition haben. Die Beziehungen zwischen diesen Begriffen und die Ursache für ihr Auftreten sind jedoch nicht ganz so klar. In diesem Artikel schlagen wir ein Gerüst vor, das die Forschungsgebiete miteinander verknüpft.

1 Introduction

It is well established that measurements always contain random deviations. Gauß proposed a general method for treating these deviations and Helmert discussed the different types of deviations (HELMERT 1872: 1-27). Main contributors in the discussion of measurements were astronomers, surveyors, and cartographers.

The increasing influence of geographic information systems in decision processes attracted more attention for the topic of data quality. A large number of texts (see for example BURROUGH 1986; CHRISMAN 1983; CHRISMAN 1991; FISHER 1987; GOODCHILD 1991; GUPTILL & MORRISON 1995; VEREGIN 1999; WORBOYS 1998) were published, which dealt with data quality. Conferences like the Symposium on Spatial Accuracy Assessment or the Symposium on Spatial Data Quality concentrate on data quality but it is also an important topic for other conferences like

- Angewandte Geoinformatik (AGIT),
- Conference on Spatial Information Theory (COSIT),
- GIS Planet,
- International Conference on Spatial Analysis and GEOmatics, Research & Development (SAGEO),

- International Symposium of ACM GIS, and
- International Symposium on Spatial Data Handling (SDH).

According to the International Organization for Standardization (ISO) quality is the „*totality of characteristics of a product that bear on its ability to satisfy stated and implied need.*" A brief overview about elements of data quality and definitions of the basic term has been provided by VAN OORT (2006: 12-18).

Discussion of uncertainty started when dealing with risk assessment in decision processes. Examples for applications are seeding of hurricanes (HOWARD, MATHESON & NORTH 1972) or the protection against wildfires (NORTH, OFFENSEND & SMART 1975). Early approaches used decision trees and probabilities to specify possible loss or gain. Later discussions developed a broader view (FISHER 1999).

A general framework for the different aspects of data quality is missing. Fisher stated that the concepts of uncertainty and data quality should be complementary. Fisher complained that „*neither seems to have paid little more than lip service to developments in the other*" (FISHER 2003).

In this paper we describe how data are captured and show the sources for errors. We discuss the sources of these errors in an ontological framework. This allows the separation of different influences and clarifies the different concepts of data quality.

2 5-Tier Ontology

The 5-tier ontology (FRANK 2001) structures our reality into different tiers. Tier 0 is the physical reality. Tier 1 contains the observations of tier 0. The observations are grouped in objects. These objects form tier 2. Tier 3 is the social reality (SEARLE 1995) where physical objects may get a new function. Pieces of paper, for example, may be used as money in a specific social context. Finally, each actor in society has a subjective view on reality. This is tier 4.

Tier 0 describes the physical reality we live in. The underlying assumption is that there is only one physical environment where each point in space and time has determined properties. Knowledge about the properties is gathered by observations. The separation between reality and observation has been introduced by the Greek philosopher Plato.

Observations form tier 1. The observations considered here are observations of properties at points. The observations return a value *v* for a specific property at a specified point in space and time:

$$p(\underline{x},t) = v\,.$$

The observations in tier 1 describe the world. Unfortunately the amount of data is high. Discussions about space do not consider the point observations but group them to objects. Objects have some important properties. Among the most important is that they exist over an extended period in time (AL-TAHA & BARRERA 1994).

The simplification of the object formation is evident when looking at a small example. Let us assume that the mass *m* of a plate is specified by observation of the density *d* of points in tier 1. The mass of the plate can be seen as the set of densities for the points forming the plate or as a single parameter, the mass. The connection between these two views is

$$m(V) = \int_V d(\underline{x}, t) dV \; .$$

Tier 3 describes the socially constructed reality. Physical objects from tier 2 may have a special function in a specific social context. Contracts as physical objects, for example, are sheets of paper with text and signatures on it. Within a legal context a contract is a binding regulation for the signatories. The meaning of the object has changed due to the construction process.

Between tiers 2 and 3 the processes for creating the objects may change. A piece of land in tier 2 may be created by placing a fence along the outline of the land. In the social reality the object „land parcel" is created and changed by clearly defined cadastral processes.

Cognitive agents living in the environment have a subjective view of the reality based on observations the agent made himself or were reported by other agents (compare FRANK 2000). These observations form the knowledge of the agent and influence his view of reality. This subjective view is tier 4.

3 Observation of Reality

It is well established that the results of observation processes are not error-free. The physical reality in tier 0 does not have errors (it just is). Thus there is a predefined value, the „real" value, which should be returned by an observation process. However, the result *v* of observations process deviates from this „real" value *v'* by the value ε:

$$\underline{v'} + \varepsilon = v \; .$$

The standard assumption for the distribution of the deviation ε is normal distribution. Gross errors and systematic errors are eliminated by using control observations and specific observation methods that are insusceptible to systematic effects. Helmert lists the following assumptions for the remaining random deviations (HELMERT 1872: 6):

- The number of positive and negative deviations is equal.

- The frequency of occurrence increases with decreasing absolute value of the deviation.
- The deviation 0 has the maximum frequency of occurrence.

In other words, the average of the random deviations tends to zero as the sample size tends to infinite. Thus the average is an appropriate method to eliminate these deviations if the sample is large enough.

Statistical methods are used to model the deviations. The law of random error propagation provides a measure for the quality of the result of data processing. A typical measure is „accuracy" as defined in the context of data quality (DRUMMOND 1995; SALGÉ 1995).

Decisions based on observations must apply concepts of statistical testing since observations always contain deviations. Statistical tests serve to determine whether or not a basic postulate, a hypothesis, holds. Statistical tests decide about acceptance and rejection of the hypothesis based on the probability that the sample data is observed while assuming the hypothesis to be true. Two kinds of errors are possible in a statistical test. Type I is the error of rejecting a correct hypothesis. The probability of such an error is the significance level α. The probability level of making a correct decision is thus $(1-\alpha)$, the confidence level. Type II is the error of accepting a hypothesis that is actually false. The probability of making such an error is denoted as β. The probability of making a correct decision about a false hypothesis is the power of the test $1-\beta$ (MÜLLER 1991; REISSMANN 1976: 342-343).

A different problem emerges from the fact that observations of properties at points are impossible in a strict way. The observation process covers a small area and returns the average value for this area as a result. Observing the temperature at a specific point, for example, requires a temperature sensor. The sensor has a certain area where the temperature induces a response that can be quantified. Although we treat the observations as a point observation, the resulting value is the average temperature for the area of the sensor. Variations with a smaller spatial extent than the observations will thus be lost. Typical examples are satellite images where the area used to determine the value of a „point observation" may comprise a few hundred square meters.

4 Forming Objects

Forming objects requires two steps: Classification and boundary definition. The observations from tier 1 are used to identify both, the type of objects and the boundary.

4.1 Classification

Classification may use affordances (GIBSON 1979) to specify the necessary properties. Affordances are what objects or things offer people to do with them. A table, for example, must have a horizontal, plane, inflexible surface within a suitable height range to afford people a basis for placing things while standing or sitting. This concept is extremely

flexible since suitable height may range from 40 cm for a traditional Japanese table to 1,3 m for tables at standing buffets. The concept does not specify the size or shape of the surface and still we can recognize furniture fitting the concept. The specification of type uses the observations and checks if the observations fit to one of the known concepts.

Another problem is the classification scheme itself. The terms „mountain" and „hill" stand for two different classes. Some elevations like „Matterhorn", „K2", or „Großglockner" definitely belong to the class mountain, whereas in other cases elevations clearly belong to the class hill. Where is the border between a hill and a mountain? These questions lead to fuzzy sets and the resulting fuzzy logic (ZADEH 1974; ZADEH 1965).

4.2 Boundary Definition

We deal with objects in everyday life. The rule generally used to determine the boundary of the objects is simple: Everything that is physically connected to the object belongs to the object. This strategy becomes visible when dealing with stacks of objects, e.g., books. Removing the top book requires determination of its boundary, i.e., determination of the bottom area of the book. Shifting the book shows the extent of the book and allows taking this book. This strategy creates problems if the top books stick together and therefore move in the same way. Then a change of strategy is necessary.

The simple test for the extent of an object thus requires moving the object. This is possible for objects in figural space and some objects from vista space (MONTELLO 1993). Objects in environmental or geographical space, however, must be treated differently. Let us assume the following example: Agricultural subsidies are paid for production in mountainous regions. The problem is thus the definition of mountainous regions. Figure 1 shows a simple approach. The boundary is defined by providing a limiting value for the height above sea level, in this case 1.000 m. Every point above 1.000 m is a part of the mountain and areas within that region will receive subsidies. The boundary of the mountain is the contour line with the specified height. Theoretically this is an unambiguous definition of the boundary.

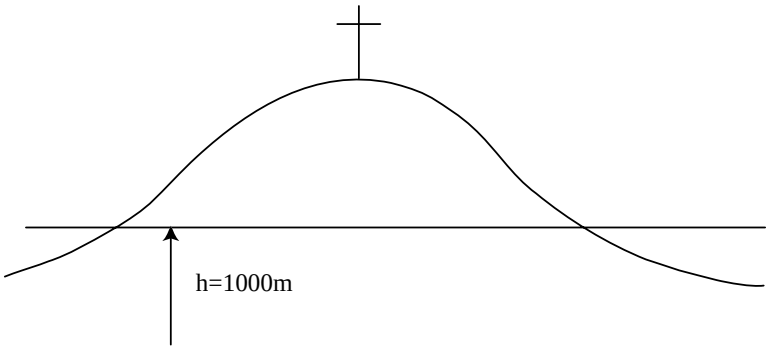


Fig. 1: Definition of the boundary for a mountain

However, the definition contains two elements that may be subject to error: The representation of the terrain is not error-free and the definition of the limiting value may contain errors. These influences will be discussed separately.

4.2.1 Influence of errors in the terrain model

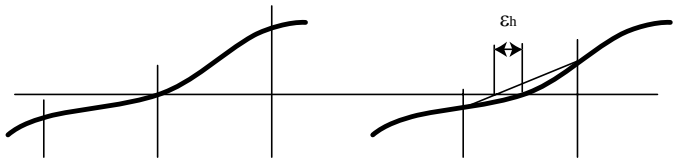
Height data of each point of the mountain comes from observations and thus the height is not error free. The observations are usually structured as a raster. Improvements by adding break lines are possible (KRAUS, ASSMUS et al. 1982). A simple formula for deriving the accuracy of contours derived from such a digital elevation model is (ACKERMANN 1978)

$$\sigma_z = a + b \cdot Z' .$$

The parameter $Z' = \tan\alpha$ represents the maximum slope, *b* expresses the influence of the slope on the height accuracy, and the parameter *a* denotes the accuracy that is independent of the slope. As discussed by Kraus these parameters are not independent from the observation process and Kraus derived the following formula for the horizontal accuracy of the contours (KRAUS 1994):

$$\sigma_{zp} = \frac{a}{Z'} + b .$$

In addition the observations are done in a specific raster and the position of the raster points as well as the resolution may have an impact on the result. Since in general the points will have a height different from the specified height of the boundary, interpolation will be necessary. Figure 2 shows the result of the interpolation for the 1-dimensional case. The left image shows the surface and the original raster. In the left image the raster was shifted and the position of one of the original raster points was interpolated based on the height of that point. The position is moved by approximately a quarter of the raster width.



(a) original raster

(b) shifted raster with interpolation of point with specified height

Fig. 2: Dependency between raster points and interpolation result

The problem becomes even more complicated in the 2D-case. The four corner points of a raster cell will, in general, not lie in a plane. Therefore, a method must be selected to interpolate the height between the raster points. One of the simplest methods is dividing the squares in two triangles. Unfortunately, this can be done in two directions and the results vary. Fisher analyzed the effects of different interpolation implementations using the

example of viewshed analysis and showed a dependency between the type of interpolation and the result of the function (FISHER 1993).

Also the width of the raster influences the quality of the result. As discussed in section 0 the observations of properties of points are the result of an averaging process. The collection process determines the size of the area that is used to determine the value. Therefore the collection process introduces a scale dependency in the data set. This scale dependency is directly visible with analogue representations like printed maps but also exists with digital data sets.

4.2.2 Influences of errors in the limiting value

Where do bounding values for the definition of objects come from? There are two different cases:

- The value may be defined by law.
- The value may be derived from other observations.

The first case is the simpler one. Legal rules usually do not consider tolerances (TWAROCH 2005). The result is like the picture in Figure 1. Problems arise if the comparison of the defined value with measurements is not conclusive. Is a distance of 2,99 m with a standard deviation of 5 cm smaller than 3 m? This question becomes increasingly important with more complex calculations like areas (KAY 2006, NAVRATIL & ACHATSCHITZ 2004) or volumes (KRAUS 2000).

Values derived from other observations, however, must allow variations. Classification of land cover derived from satellite images must consider deviations in the image caused by objects smaller than the area covered by the picture element. Figure 3 shows the effect of such an imprecise value for the definition of the boundary of a mountain. In contrast to the result of Figure 1 the boundary in this case is imprecise due to the variation of the bounding condition.

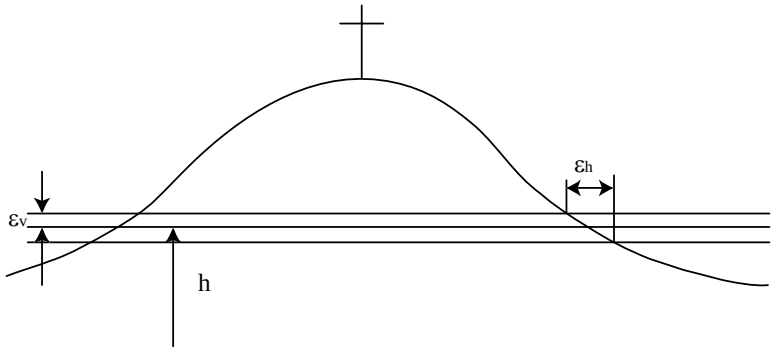


Fig. 3: Definition of boundary with imprecise bounding value

5 Socially Constructed Objects

In tier 3 the social context is added and the objects in tier 3 are socially constructed objects. Generally a description of quality of data for socially constructed objects becomes pointless if we are not considering the use. The definition of forest according to the Austrian law is (FORSTGESETZ 1975 §1a)

- Areas with trees of specified kinds with a size of at least 1.000 m^2 and an average width of at least 10 m,
- Areas according to above definition where the trees are temporary removed or reduced,
- Areas without trees, which are directly necessary for forestry.

According to this definition areas, which are not used for forestry, are not a forest. This definition does not involve the existence of trees in the areas defined as forests. The definition is purely based on forestry. Since forestry requires free areas for transport and storage of trees these areas are included in the definition. It is possible to provide the quality of a data set using the definition for the forest law because there is a rule for inclusion and exclusion of objects (areas).

The result of the quality definition changes automatically, if the definition of forest changes. Another definition of forest would be „areas covered by crowns of trees". This definition includes single trees and would allow estimating the amount of timber available or the amount of oxygen production by trees. Answering these kinds of questions with a data set collected using the first definition will provide incorrect results.

For any data set it would be possible to invent some strange definition to make it the image of reality. The problem is finding „useful definitions" for the data set. The „usefulness" depends on the use of the data and thus the quality of data sets must be discussed in the context of use.

6 Conclusions

We have seen that data quality emerges from observations processes. The limited quality of the observation process restricts the achievable quality of the final data. Statistical methods are suited to deal with data quality. Uncertainty, on the other hand, is influenced by observations and definitions. Uncertainty arises when observations shall be used to form objects according to defined classes.

When discussing quality of geographic data we have to separate two cases. Data quality measures are sufficient for the quality description if we deal with observation values. Derived objects, however, require also a description of the uncertainty measures. The example discussed in section 4 uses definitions derived from laws or from technical

considerations. The major difference between these two cases is that values specified in laws are error free and values from technical considerations are not.

It became also evident that quality is only meaningful when the use is specified. The definition of classes is only possible if the use of the resulting data set is known. The definition of forest in the Austrian forest law, for example, deals with forestry and not with biomass. This influences the definition itself. This must be visible for the user to avoid misuse of data. One of the future challenges will be making the connection between class specifications and data sets visible for the user.

5 References

(1975). *Forstgesetz (Forestry law)*. BGBl.Nr. 440/1975.

Ackermann, F. (1978). *Experimental Investigation into the Accuracy of Contouring from DTM*. In: Photogrammetric Engineering 44 (12): 1537-1548.

Al-Taha, K. & R. Barrera (1994). *Identities through Time*. In: International Workshop on Requirements for Integrated Geographic Information Systems, New Orleans, Louisiana.

Burrough, P. A. (1986). *Data Quality, Errors and Natural Variation*. In: Principles of GIS for Land Resources Assessment. Oxford, UK, Clarendon Press**:** 103 - 135.

Chrisman, N. R. (1983). *The Role of Quality Information in the Long-Term Functioning of a Geographic Information System*. In: Sixth International Symposium on Automated Cartography (Auto Carto Six), Ottawa, Ontario, Canada, The Steering Committee for the Sixth International Symposium on Automated Cartography.

Chrisman, N. R. (1991). *The error component in spatial data*. In: Geographical Information Systems: principles and applications. D. J. Maguire, M. F. Goodchild & D. W. Rhind. Essex, Longman Scientific & Technical. 1**:** 165-174.

Drummond, J. (1995). *Positional Accuracy*. In: Elements of Spatial Data Quality. S. C. Guptill & J. L. Morrison. Oxford, Elsevier**:** 31-58.

Fisher, P. F. (1987). *The Nature of Soil Data in GIS - Error or Uncertainty*. In: International Geographic Information Systems (IGIS) Symposium (IGIS'87), Arlington, Virginia.

Fisher, P. F. (1993). *Algorithm and implementation uncertainty in viewshed analysis*. In: International Journal of Geographical Information Systems 7 (4): 331-347.

Fisher, P. F. (1999). *Models of Uncertainty in Spatial Data*. In: Geographical Information Systems - Principles and technical Issues. P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind. New York, Wiley & Sons, Inc. 1**:** 191-205.

Fisher, P. F. (2003). *Data Quality and Uncertainty: Ships Passing in the Night!* In: International Symposium on Spatial Data Quality, Hong Kong, Hong Kong University Press.

Frank, A. U. (2000). *Communication with maps: A formalized model*. In: Spatial Cognition II (Int. Workshop on Maps and Diagrammatical Representations of the Environment, Hamburg, August 1999). C. Freksa, W. Brauer, C. Habel & K. F. Wender. Berlin Heidelberg, Springer-Verlag. 1849: 80-99.

Frank, A. U. (2001). *Tiers of ontology and consistency constraints in geographic information systems*. In: International Journal of Geographical Information Science 75 (5: Special Issue on Ontology of Geographic Information): 667-678.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Hillsdale, NJ, Erlbaum.

Goodchild, M. F. (1991). *Issues of Quality and Uncertainty*. In: Advances in Cartography. J. C. Mueller**:** 113-139.

Guptill, S. C. & J. L. Morrison, Eds. (1995). *Elements of Spatial Data Quality*. Elsevier Science, on behalf of the International Cartographic Association.

Helmert, F. R. (1872). *Die Ausgleichungsrechnung nach der Methode der kleinsten Quadrate*. Leipzig, Germany, B. G. Teubner-Verlag.

Howard, R. A., J. E. Matheson & D. W. North (1972). *The Decision to Seed Hurricanes*. In: Science Vol. 176 (February 23): 1191-1202.

Kay, S. (2006). *Field Area Checks Using GPS (2)*. In: GIM International 20 (January): 43-45.

Kraus, K. (1994). *Visualization of the Quality of Surfaces and Their Derivatives*. In: Photogrammetric Engineering & Remote Sensing 60 (4): 457-462.

Kraus, K. (2000). *Zur Genauigkeit der Volumenbestimmung*. In: Zeitschrift für Vermessungswesen 125 (12): 398-401.

Kraus, K., E. Aßmus, A. Köstli, L. Molnar & E. Wild (1982). *Digital Elevation Models: Users' Aspects*. In: 38. Photogrammetrische Woche 1981, Stuttgart, Institute for Photogrammetry, University Stuttgart.

Montello, D. R. (1993). *Scale and Multiple Psychologies of Space*. In: *Spatial Information Theory: A Theoretical Basis for GIS*. A. U. Frank & I. Campari. Heidelberg-Berlin, Springer Verlag. 716: 312-321.

Müller, P. H. E. (1991). *Lexikon der Stochastik*. Berlin, Akademie-Verlag GmbH.

Navratil, G. & Achatschitz, C. (2004). *Influence of Correlation on the Quality of Area Computation*. In: International Symposium on Spatial Data Quality, Bruck a.d. Leitha, Austria, Department for Geoinformation and Cartography.

North, D. W., F. L. Offensend & C. N. Smart (1975). *Planning Wildfire Protection for the Santa Monica Mountains: An Economic Analysis of Alternatives*. In: Fire Journal (January): 69-78.

Reißmann, G. (1976). *Die Ausgleichungsrechnung*. Berlin, VEB Verlag für Bauwesen.

Salgé, F. (1995). *Semantic Accuracy*. In: Elements of Spatial Data Quality. S. C. Guptill & J. L. Morrison. Oxford, Elsevier**:** 139-151.

Searle, J. R. (1995). *The Construction of Social Reality*. New York, The Free Press.

Twaroch, C. (2005). *Richter kennen keine Toleranz*. In: Internationale Geodätische Woche, Obergurgl, Wichmann.

van Oort, P. (2006). *Spatial Data Quality: From Description to Application*. In: Centre for Geo-Information. Wageningen, The Netherlands, University Wageningen**:** 125.

Veregin, H. (1999). *Data Quality Parameters*. In: Geographical Information Systems. P. A. Longley, M. F. Goodchild, D. J. Maguire & D. W. Rhind, John Wiley & Sons, Inc. 1: 177-189.

Worboys, M. F. (1998). *Imprecision in Finite Resolution Spatial Data*. In: GeoInformatica 2 (3): 257-279.

Zadeh, L. A. (1965). *Fuzzy Sets*. In: Information and Control 8: 338-353.

Zadeh, L. A. (1974). *Fuzzy Logic and Its Application to Approximate Reasoning*. In: Information Processing 74, Proceedings of the IFIP Congress 1974 (3), pp. 591-594.