PERIODICA POLYTECHNICA SER. CIVIL ENG. VOL. 39, NO. 2, PP. 77-84 (1995)

DATA QUALITY IN GIS SYSTEMS

Ákos DETREKŐI

Department of Photogrammetry
Technical University of Budapest
H-1521 Budapest, Hungary
Tel: 463-1187

Received: 15 May, 1995

Abstract

The quality of data depends on various factors: e. g. accuracy, precision, consistency and lineage. Beside these factors, the definition of the data reliability will be discussed in detail. In the data quality management it is necessary to estimate the values of the quality components before the realization of the data base (a priori estimation) in the period of data bases design. It is required to have redundance in the data for the a posteriori estimation of the quality components in the control period. In the paper some mathematical-statistical methods for the a priori and a posteriori estimation of data quality are presented. Along with the theoretical foundation some examples and numerical values will be given, too.

Keywords: data quality, GIS, data bases design, precision, reliability.

1. Introduction

Digital systems are capable of processing data more precisely than analogue systems, but their overall quality still depends on the quality of their source data, which remains analogue in most of the cases. In GIS the situation is similar, the final quality depends on source quality (BERNHARDSEN, 1992). The GIS data quality is compromise between needs and costs. In practice the choice is often a question of shopping what is currently available or can be acquired in a reasonable amount of time.

Four aspects of data acquisition comprise the criteria for selecting data quality (BERNHARDSEN, 1992):

- need,
- costs.
- accessibility,
- time frame.

The most important factors of data quality are (NCGIA, 1990; BERNHARDSEN, 1992):

- lineage.
- georeferencing,
- attribute data,
- consistency of links between geometries and attributes,
- geometry link consistency,
- data completeness,
- data currentness.

In the paper the quality of the georeferencing and the quality of the attribute data will be discussed.

In the quality management it is necessary to estimate the values of the quality components before the realization of the data base (a priori estimation), and it is required to have redundance in the data for an a posteriori estimation of the quality components. The a priori estimation is a tool of the design of GIS. The a posteriori estimation has an important role in the control of data bases, it is very often useful by making of data quality report.

2. Measurement of Data Quality

There are various values to measure the data quality. The most important are the following:

- accuracy,
- precision,
- reliability,
- degree of the misclassification.

The definitions of these measures are in various books often different. In the paper will be given first a possible definition of each measure, then the estimation of these measures in the design and in the control processes will be discussed separately.

The *accuracy* is defined as the closeness of results to the true values. Normally the accuracy is characterised by the standard deviations or root-mean-squares . If the results are characterised by the random variable x, then the standard deviation is the following:

$$s = \sqrt{E[(x - E(x))^2]}\,, \tag{1}$$

where $E[\,]$ is the symbol of the expectation.

The *precision* is defined as the number of decimal places or significant digits in the measurements. The precision is not the same as accuracy. A large number of significant digits doesn't necessarily indicate that the

measurements are accurate (NCGIA, 1990). The precision expresses the repeatability of the measurement.

The *reliability* gives information about the controllability and about the blunders (gross errors) of the data. The value of the smallest detectable blunder is characterised by the reliability. In the case of normal distributed measurement the value of reliability can be determined (BAARDA, 1968).

The *degree of misclassification* is a measure of the quality of some attribute data. The degree of misclassification can be determined as a function of the misclassification matrix. Often useful functions are the error rate (the number of the errors/hundred units) and the Cohen Kappa index. Other measures are given by GOODCHILD and GOPAL (1989).

3. The Data Quality Estimation in the Design Period of GIS

Technically correct and economical data acquisition is preconditioned by design, consisting essentially in selecting the data sources, the method of the acquisition and the quality of the data.

In the design period the data quality estimation is based upon:

- the experience of the specialists,
- the literature,
- the standards.

The data quality could be characterized by the accuracy, precision, reliability and the degree of misclassification. The a priori estimation of the various measures needs various methods. It will be summarised some methods and values based upon the literature.

The *georeferencing accuracy* depends on the method of data acquisition. Some informing values are the following (BAHR – VOGTLE, 1991; BILL – FRITSCH, 1991):

Method of data acquisition	Accuracy (circa)
Surveying	cm – dm,
Photogrammetry	
– stereoplotting	0.00001 * photo scale figure,
– DTM (elevation)	0.0001 * object distances,
Remote sensing (resolution on the terrain, m*m)	
LANDSAT Thematic Mapper (TM)	30*30,
Système Probatoire d'Observation de la Terre (SPOT)	20*20,
Russian space photograph (KATE)	5*5,
Digitising of maps	0.00025* map scale figure.

Digital map data generated directly from aerial photos or entered during surveying entail fewer steps and consequently are less subject to error than data from digitised maps. From standpoint of accuracy, original sources are always preferable to maps (BERNHARDSEN, 1992).

In the large scale domain a special problem is given by the uncertainty of definition of natural points or lines. By KRAUS (1993) some values of the uncertainty of natural point definition were published.

Type of point	Planimetry	Height
House and fence corners	7 – 12 cm	8 – 15 cm
Manhole cover	4 – 6 cm	1 – 3 cm
Field corners	20 – 100 cm	10 – 20 cm
Bushes, trees	20 – 100 cm	20 – 100 cm

In the small scale domain the cartographic adaption may introduce errors, too (BERNHARDSEN, 1992).

The *accuracy of attribute data* depends on the data acquisition method, too. In the literature we can find sometimes information about data quality. That can happen quite easily by direct measurement (e. g. seismic data) or by other kinds of sampling (e. g. data of public opinion test).

The a priori estimated value of data accuracy is very useful, if the required accuracy of data is given (e. g. in a standard). We can use a data acquisition method only if the following inequality is valid:

$$a \leq A, \tag{2}$$

where a is the a priori estimated value of accuracy,

A is the the required value of accuracy.

The inequality (2) could be useful in the case of georeferencing and in the case of attribute data. The a and A values normally are standard deviations or tolerances.

The determination of the A required value of accuracy is an important problem of the various standards. The required value of the accuracy of georeferencing can be determined as the standard deviation of the point determination (it means as the standard deviation of the coordinates) or as the standard deviation of the distances. The tolerance is usually the triple of the standard deviation.

The numerical value of the required accuracy is a function of the data aggregation level. In the analog maps the aggregation level is correlated with the map scale. The required standard deviation of the point determination in the large scale domain is in the order of cm – dm. The

topographical domain can be characterised by the m – 10 m order of the required standard deviation.

The *precision* of data is defined as the number of decimal places or significant digits. For normal measurement of interest in GIS, precision is limited by the instruments and methods used. The value of precision of a data is usually smaller than the value of the accuracy of the same data.

The *reliability* gives information about controllability and about the blunders (gross errors) of the data. For the determination of the reliability we can use the methods of adjustement of survey and photogrammetric measurements (BAARDA, 1968; MIKHAIL, 1981; FÖRSTNER, 1980). The preliminary condition of the determination of the numerical value of the reliability is the existence of redundancy. In the design period of data acquisition we can estimate the redundancy of data. The redundancy of a data set is given by the following equation:

$$h = N - n, \tag{3}$$

where N is the number of all data, n is the number of necessary data for the unique solution.

The ratio $t = h/n$ is a possible measure of the controllability of the data set. If

$0 \leq t \leq 0.01$	the data set isn't controllable,
$0.01 \leq t \leq 0.1$	the data set is poorly controllable,
$0.1 \leq t \leq 0.3$	the data set is sufficiently controllable,
$0.3 \leq t$	the data set is well controllable.

In the standards often is given the number or the precentage of the data to be tested. Let be this number H. It is necessary the validity of the following inequality:

$$h \leq H. \tag{4}$$

The other possibility is the determination of the required reliability. This is e. g the fivefold of the required standard deviation.

The *degree of misclassification* is estimated very difficultly in the design period. Only the literature and the experience of operating GIS give some information for this purpose.

The required accuracy of the attribute data is a function of the object character. E. g in Finland (JAKOBSSON, 1994) for the topographic domain the following values are given:

one error/hundred unit	road numbers, railroads class, administrative boundary class, voltage class;
four error/hundred unit	road class, field class, use of building;
fifteen error/ hundred unit	path class.

4. Data Quality Estimation after the Data Acquisition

The estimation of the accuracy, precision, reliability and the degree of misclassification is possible using various methods.

The *accuracy* of data is characterised normally by the standard deviation. The real value of the standard deviation of a data can be determined in the following ways:

- using the least squares or other estimation,
- using more accurate data,
- analysis of attribute data.

If we use the *least squares estimations* (e. g. in the case of determination of georeferencing data) the standard deviation values can be calculated using well-known equations. For example the covariance matrix of coordinates using the observation equation is the following (MIKHAIL – ACKERMAN, 1976):

$$\mathbf{M} = \frac{1}{h}\mathbf{v} * \mathbf{W}\mathbf{v}\,(\mathbf{A} * \mathbf{W}\mathbf{A})^{\wedge}\,, \tag{5}$$

where h is the redundancy,
$\mathbf{v}$ is the vector of the residuals,
$\mathbf{A}$ is the design matrix of the observation equations,
$\mathbf{W}$ is the weight matrix of the observations.

The standard deviations of the coordinates are the squares of the diagonal elements of the covariance matrix.

Sometimes the standard deviation will be determined using *more accurate data* (e. g. in the case of the control of digitised elevation from topographical maps using levelling). In this case we can use the following equation:

$$s = \sqrt{\frac{1}{n}\,(\mathbf{d} * \mathbf{d})}\,, \tag{6}$$

where n is the number of the compared values,
$\mathbf{d}$ is the vector of the differences of values.

Attribute accuracy must be analyzed in different ways depending on the nature of data. Quantitative data may be defined in non-numerical terms (ordinal data), as discrete variables divided into classes (interval data) or continuous variables without numerical limits (ratio data), (BERNHARDSEN, 1992).

The *precision* is determined by the used instruments.

The *reliability* of each data can be determined using the methods of adjustment. In the case of normal distributed data the following equation

may be useful (BAARDA, 1968):

$$r = gsm\,, \tag{7}$$

where s is the a priori standard deviation of the measurement,
m is the square of diagonal element of the following matrix

$$\mathbf{Q} - \left[\mathbf{A}\,(\mathbf{A} * \mathbf{WA})^{\wedge}\,\mathbf{A}\right]\mathbf{W}\,,$$

where $\mathbf{Q}$ is the cofactor matrix of the measurement,
$\mathbf{A}$ is the design matrix of the observation equations,
$\mathbf{W}$ is the diagonal weight matrix of the measurement,

$$g = g(\alpha, \beta)\,,$$

where α and β are significant levels.

This method of the calculation of the reliability can be used in the case of the least squares adjustment of the measurement.

The *degree of misclassification* is characterised using the misclassification matrix. The determination of the misclassification matrix and the calculation of the function of this matrix (e. g. Cohen Kappa index) was published in the literature (e. g. NCGIA, 1990). Various possibilities of the determination of other measures were given by GOODCHILD and GOPAL (1989).

The results of the a posteriori data quality estimation are given usually in the Data Quality Report. The contents of the Report is determined by standards. There are international standards about data quality generally. The most important standards in this field are the ISO 9001-9004 standards. In various countries national standards of GIS data exist (e. g. BREGENZER, 1992, JAKOBSSON 1995).

5. Conclusion

In the data quality mangement it is necessary to estimate the values of the quality components before the realization of the data bases (a priori estimation), and it is required to have redundance in the data for the a posteriori estimation of the quality components. The a priori estimation is a tool of the quality design of GIS. The a posteriori estimation has an important rule in the control of realized data bases. The result of the a posteriori estimation is useful very often by the draft of quality report. The methods of the adjustment of survey and photogrammetric measurements are applicable for the a priori and a posteriori estimation of the data quality.

References

BAARDA, W. (1968): A Testing Procedure for Use in Geodetic Networks. *Netherlands Geodetic Commision.* Vol. 2. No. 5, Delft.

BAHR, H.-P. – VOGTLE, T. (1991): Digitale Bildverarbeitung. Wichmann Verlag, Karlsruhe.

BERNHARDSEN, T. (1992): Geographical Information System. VIAK IT and Norwegian Mapping Authority.

BILL, F. – FRITSCH, D. (1991): Grundlage der Geo-Informationssysteme 1. Wichmann Verlag, Karlsruhe.

BREGENZER, W. (1992): Die Reform der amtlichen Vermessung RAV in der Schweiz, *Zeitschrift für Vermessungswesen,* Vol. 7/8, pp. 444–449.

BURROUGH, P. A. (1986): Principles of Geographical Information Systems for Land Resources Assessment. Clarendon Press, Oxford.

CAROSIO, A. (1991): Plannumerisierung, ETH Institut für Geodesie und Photogrammetrie, Mitteilung 47.

DETREKŐI, Á. (1994): Data Quality Management in GIS Systems, *Computers, Enviroment and Urban System,* Vol. 18, N. 2, Pergamon Press, New York, Oxford, pp. 81–86.

FÖRSTNER, W. (1980): The Theoretical Reliability of Photogrammetric Coordinates. *XIV. Congress of International Society for Photogrammetry,* Commission III (pp. 223–245)

GOODCHILD, M. – GOPAL, S. ed. (1989): Accuracy of Spatial Databases. Taylor-Francis, London, Bristol.

JAKOBSSON, A. (1994): Quality in making the Topographic Database, *FIG XX. Congress,* 343/1-6., Melbourne.

KRAUS, K. (1993): Photogrammetry, Volume 1. Dümmler Verlag, Bonn.

MIKHAIL, E. M. – ACKERMANN, F. (1976): Observations and Least Squares. IEP Dun-Donelley Publisher, New York.

MIKHAIL, E. M. – GRACIE, G. (1981): Analysis and Adjustment of Survey Measurements. Van Nostrand Reinhold, New York.

National Center for Geographic Information and Analysis NCGIA (1990): Core Curriculum. Introduction.