(1993)

GIS BASED DECISION MAKING MUST CONSIDER DATA QUALITY

Heinz Stanek, Andrew U. Frank
Technical University Vienna
Department of Geoinformation
Gusshausstrasse 27-29, A-1040 Vienna, Austria
stanek@geoinfo.tuwien.ac.at
frank@geoinfo.tuwien.ac.at

ABSTRACT

The great benefit of GIS is that it simplifies the process of decision-making. The availability of data quality parameters improves decisions in GIS applications. Besides the economic aspects, questions of liability will have to be considered in the future. In order for data quality to be a useful tool it is not sufficient that the necessary information be supplied to the decision-maker. In a second step, an appraisal of the quality of the information provided must be carried out. This applies to all aspects of spatial decision making, from real estate acquisition to environmental protection measures.

There is, however, the problem of differentiating between the constituent elements of data quality, i.e. positional accuracy, attribute accuracy, temporal update level, or scale. They are often compounded in a lineage description related to a special field of applications. The goal must be to provide an analytical description of data quality, which is practically "independent" of the processes used.

Another problem is assessing the influence of data quality on the decisions made. One solution is, for instance, to find a "weighted" combination of basic components of data quality, which is then applied to assess the influence of data quality on a specific type of decision. Up to now, scale has been used as virtually the only numeric quality parameter.

The paper reviews our current understanding of these problems and lists specific items for future research. Different data quality statements and the availability of data quality in conventional processes, as compared to GIS-based ones, are analyzed. Some examples are included to illustrate specific demands which data quality parameters must meet. They also show the close relationship between data quality and applications. The authors reflect on the possibilities of structuring data quality for processing in GIS, which they think is the key for universal applications.

INTRODUCTION

Quality is a key aspect of all technological and scientific work. No matter what the project, or at which stage it stands, quality criteria will have to be complied with: in the conceptional phase just as well as in the process of tendering, or after its completion, when it has become fully operational. The definition of quality criteria will depend on the needs of the project and can reach from how to observe a particular feature, to how to process these observations, to what objectives have to be met. Important quality criteria are laid down in the form of standards. For GIS, the standards DIN ISO 9000-9004 are applicable (Caspary 1992). Besides defining quality standards for software (ISO 9000 Part 3), also services are taken into account and standardized (ISO 9004 Part 2).

Quality is to be described comprehensively by selecting and defining certain characteristic quality criteria that can be measured and calculated. Furthermore, these criteria should be as far as possible uncorrelated in order to allow comparison of different applications. The additional information about data quality can then help the user in decisions about the further use of the product.

Stanek, H., and A. U. Frank. "Gis Based Decision Making Must Consider Data Quality." Paper presented at the EGIS'93, Genoa 1993.

When geographical information systems are introduced, adapted or worked with in an organization, it is of vital importance to formulate, define and verify data quality criteria. The impetus to do this should be particularly strong when working with GIS-projects because they often have a very long life-span of up to several decades, during which they are continually utilized.

There are different approaches of how to split GIS-project costs between hardware, software and data acquisition. Caspary (1992) assumes, for instance, that the ratios are as follows: 10% spent on software, 25% on hardware and 65% on data. This estimate was provided on the basis of figures for a first-time investment for installing a GIS-system. However, when taking into account the costs for updating the data, the major cost share will probably shift towards data acquisition and updating. The life-span of GIS-projects can be assumed as follows: 4 years for the hardware, 8 years for the software and the duration of the project for the data.

When it is decided to install a GIS it should be borne in mind from the start that the data needs to be updated regularly, and that this updating must also be carried out in compliance with the defined quality criteria.
Due to technological progress, certain data quality parameters, which were initially identical for the entire map, may have to be modified or adapted locally or for the GIS as a whole. It might be efficient to plan several differentiated areas of data quality from the beginning.
One component of data quality is positional accuracy. This term is very well defined and documented through the work of geodesists, but is unable to provide a complete description of data quality in GIS. Other quality criteria also have an effect, such as completeness, recency (of data), ways of handling the data, etc. The attributes used in GIS must also be described by means of suitable parameters (Frank 1987).

So far, attempts to describe data quality have primarily concentrated on the aspects of establishing and updating databases. The current possibilities of data processing in a GIS open up exciting new fields of application. But it is important to realize that information about data quality must not be omitted when interpreting the GIS-related results. In conventional forms of data modelling a number of additional pieces of information, like special manipulations or manner, that go beyond the processing of geometrical representations are used for the interpretation of results. The interpretation of data quality relating to a specific process is usually subject to technical and legal guidelines.

THE STRUCTURE OF DATA QUALITY

That data quality in GIS can no longer be disregarded is demonstrated by the large number of publications on this specific topic. Its importance is also highlighted by the fact that the National Center for Geographic Information and Analysis has focused its Initiative 1 "Accuracy of Spatial Databases" on this issue. Goodchild (1992) offers a comprehensive and fundamental overview of existing attempts to structure and handle data quality parameters in databases. The work done in National Center for Geographic Information and Analysis Initiative 1 concentrated primarily on the description and structuring of data quality and on its representation in databases.

In addition to modelling data quality in GIS - or as an integral part of a specific database - an efficient visualization of data quality is becoming an ever more important aspect. The National Center for Geographic Information and Analysis, realizing the need for further work in this field, started Initiative 7 "Visualization of Spatial Data Quality"(Beard, et al. 1991), devoted to this topic. Also GIS education is now attaching more importance to data quality problems and the Core Curriculum of the National Center for Geographic Information and Analysis (Goodchild, Kemp 1990) also contains several units on "accuracy of spatial databases" and "managing error". In accordance with this issues in GIS the accuracy of databases can be structured as follows:

- positional accuracy
- attribute accuracy

- logical consistency
- completeness
- lineage

This heterogeneous structure is a set of parameters for describing data quality which are to the highest possible degree "independent", i.e. do not influence each other.
In statistical terms, we can assume an idealized normal distribution for positional accuracy. A general distribution can, however, not be stated for the discrete parts of attribute accuracy. Logical consistency and completeness can either be given or not independent of positional accuracy. Statements about lineage can be available in greater or lesser detail. The above classification does not tell us anything about the data object to which it is applied: positional accuracy can obviously refer to a point, but also to a straight line or even to an area. In a raster-based system it can also refer to a pixel.

The further processing of information in a GIS also raises the question of handling data quality parameters. For cases of normally distributed positional accuracy, error propagation can be applied, on the condition that the modelling and the analysis exist in the form of an analytical function. In this way it would be possible e.g. to compute the accuracy of the distance between two points which have been determined stochastically. This does not determine the overall accuracy of the straight line connecting the two points, but only determines the accuracy of the line length. (Dutton 1991). It has to be noted that correlation, which certainly exists due to observation, processing and modelling, is not considered.

Assessment of attribute accuracies can, for instance, be made by map overlay and statistically evaluated. A problem in this is the overlay of spatial references caused by the raster elements: their positional accuracy has an influence on the results, and thereby violates the requirement of parameter "independence".

DECISION-MAKING BASED ON CONVENTIONAL MAPPING

Traditionally, maps at specific scale have been used for spatial interpretation and practical work. When we assume a generalized accuracy of representation on maps to be in the order of approx. 0.1 mm and multiply by the scale of the map, we obtain how accurate our spatial references are. In the planning process for a new railway line it is quite possible to do the general planning of the tracks with maps at a scale of 1:10.000. It will be immediately possible to exclude those areas, which will under no circumstances be part of the tract of land on which the tracks are to be laid. Those restricted areas, however, which run along the planned railway line and the adjacent areas will be subjected to closer inspection. How broad the tract of land is that will have to be inspected in greater detail depends on the building regulations, as well as on the uncertainty of representation in the map data. Maps at ever larger scales will then be used for the key areas and for highest accuracy even new on-site measurements might be necessary.

Another aspect is completeness. Completeness is frequently achieved by recording the date of updating on the document, e.g. on the most recent version of a map or map section. In various instances, such as for cadastral purposes, the date of the updated version might even be indicated on very small sections of the map, and even for individual parcels.

In other fields, the requirement of completeness might be fulfilled by other criteria: when planning a new road, the sense of completeness will be strongly influenced by visible elements, such as houses or groups of houses. Individual houses will be considered "more complete" and more accurate than housing estates. The manipulation of maps by erasing, drawing over objects, etc. is considered as a negative quality criterion.
Besides the accuracy of drawings, yet other quality indicators will have to be respected in decision making:

The methods of data acquisition and their quality are generally known. They can vary for creating one and the same map. The front sides of the parcels, where the building line usually runs, are recorded with high geodetic accuracy. The line in the rear of the property (garden side boundary), because there are generally determined with simple methods. Thus the representation of the buildings is governed by different geodetic accuracies. In calculating the area of a parcel this factor needs to be taken into account, even though it does not impair the completeness and correctness of the map itself. It is possible to visualize these different degrees of accuracy by either distinctly marking the boundary points or boundary lines in different ways or by reasonably limiting the number of digits presented on the screen.

The methods of processing and reproducing maps or digitizing parts of maps have an effect on data quality. Besides analytical transformations or the choice of projections, also cartographic methods have to be taken into account. The question is whether the mathematical manipulations constitute ways of processing which can be repeated easily for part or the whole of the area that is represented.

Cartographic processing endeavors to achieve highly esthetically representations in order to make maps better readable and thus more user-friendly. Since the map, depending on its purpose, will always focus on, but also be limited to, the representation of specific objects or specific areas. Let us take the example of sea charts. No responsible skipper would ever use a road map for steering his vessel in the coastal waters. First of all, the map projection of the road map would be entirely unsuitable, and, secondly, it does not contain any information on shallow waters, sand banks, small offshore islands or reefs, etc., which all constitute tremendous hazards for the ship and the crew.

The nautical charts demonstrate further aspects of data quality: they record different navigational aids such as beacon lights, dangerous depths, etc. These recordings must be absolutely accurate and complete, i.e. offer the highest degree of positional accuracy on the one hand, and of attribute accuracy (depth, beacons etc.) on the other. Only authorized agencies may update the charts and record additional features. The date of the update must always be registered on the chart as well. In case of an accident and a resulting claim, the chart will be used as evidence both in dealing with the insurance and in court. The cause of the accident may have been a lacking update of the chart. In this example, the relevant data quality parameters are completeness and lineage. A second aspect is planning the ship's course while taking into account positional accuracy (e.g. the ship has to navigate at a safe distance from the coast).In a large number of processes recording the time and date of the most recent update is required by legal regulations (e.g. maximum admissible age of a map). In certain cases the date of a document may also point to the form of processing used when it was established or updated.

DATA QUALITY AND DECISION-MAKING IN GIS

We are now trying to bring together the various aspects of data quality discussed above and apply them to decision making in GIS.

The introduction of a GIS is frequently motivated by the wish to speed up the processing of existing tasks with the aid of an efficient tool. Professional, legal and economic aspects may be responsible for this decision, but in most cases a GIS is acquired because the users expect that given tasks can be processed more quickly, and that the results obtained by GIS processing are more reliable than conventional ones.

Obviously, some users have been goaded into unrealistic expectations about the quality of the data and the decisions resulting from GIS-modelling by beautiful and colorful high-definition graphics on a screen. They led them to believe that the quality of data and of the decisions resulting from its processing will be equally excellent. However, depending on the GIS, different quality criteria ,as mentioned before, will have to be met, as will be demonstrated by the following examples. In land development, planning procedures require that parcel-relevant information must be complete. The parcel boundaries themselves are not so important for the first planning blueprints. Of critical importance is, however, completeness. For soil evaluations, on the other hand, the boundaries of estates or groups of parcels must be available in high accuracy. Assignment of the soil types is available with less accuracy, due to the discrete determination by means of soil samples. Data on altitude is here less important than data on the inclination of slopes - because this might affect drainage.

When building regulations are modified, the location of specific boundary points (and thus the boundaries delineating public property e.g. roads, streets) must be determined with extremely high accuracy - since property in these areas are very expensive.

The boundary lines on the side of the property pointing away from the street can be less accurate. This may, however, change when a new road is planned on the back side of the parcel. Then the criterion for high positional accuracy will have to apply there as well.

When switching to a GIS, the user often discovers that he would like to be able to carry out more far-reaching analyses, which he was unable to do before for lack of equipment or methodologies. The design of "radiation-models" in agricultural areas, destined to create optimal conditions for plant growth or, else, determining the optimal route for a drainage canal are examples for the far-reaching possibilities inherent in GIS. It seems that everything can be linked up in GIS. Coordinates and other values can always be assumed with the desired number of digits. Scale seems to exist no longer. All the data has to be imported into the GIS model. Together with all the characteristics provided in the system, the data is now ready for being processed in a set model. Points are described by spatial coordinates, lines as relations between them. Additional information is added as attribute data. Objects of varying complexity can be built from these basic components. The sets of data are stored in databases and kept ready for further processing until accessed next via GIS.

The central issue of this paper is, however, modelling of the components of data quality, which have been described in the above paragraphs. Besides modelling, the visualization of these data is under discussion. The relationships referred to before and the complex heterogeneous structure of data quality - that goes beyond a mere description of accuracy information - pose problems for which there is no complete solution at the present time (Beard 1991). One first attempt has been to include information on accuracy as attribute data. Including accuracy statements for coordinate components actually seems a viable solution. By means of error propagation all subsequent information can be stated as a function of coordinate points.

In visualizing a straight line that connects two points we are confronted with the problem of determining its accuracy (Dutton 1991). National Center for Geographic Information and Analysis initiative 7 investigates the visualization of accuracy as areas whose boundaries are not straight lines. If several such lines form the boundary of a parcel, it is possible, according to Blakemore (1984), to distinguish five "error zones": definitely in, possibly in, ambiguous, possibly out, and definitely out. The probable area given as "error band". This error band can also be used for visualization of positional accuracy of the line.

The fact that GIS claims universal applicability makes the following demands on the description of data quality:

1. Derived quantities must remain comprehensible, i.e. it must be possible to "see through" them and find out of which fundamental/original data they are derived from, or how they were processed before as to data quality. The most uncertain component of a combination determines the quality of the result in more or less dominant form. The combination must be described with regard to all data quality components that are relevant for a specific case.

2. For heterogeneous sub-areas in data quality, the criterion applies to each sub-area. It must be possible to reconstitute the quality information locally. This requirement is particularly evident with regard to positional accuracy. If an area of a map at a scale of 1: 1000 was digitized and other areas were determined by means of geodetic observation with an accuracy of 1 cm, data quality description for each must be maintained separately. Whether it is admissible to combine them can only be determined with a view to the purpose of the task. If the result is to be used in further analysis, statements about data quality should be as far as possible derived from original data. Since data quality parameters occur as original data in model-like interpolation will be necessary in order to determine the parameter required of any particular point. Obviously, statements about lineage or completeness cannot be interpolated. The date of recording will refer to a group of points, relations, or entire objects. The date of an attribute can be a different one.

3. It seems desirable to classify the contributions of data quality parameters in processing. This will allow a structured analysis of the influence which individual data components exert on data quality. Such an approach would allow users of a GIS to process as far as possible with the available data. For example the planning of railway tracks at least partly with the aid of

the existent GIS. In critical areas it would be necessary to systematically feed basic spatial elements with a high degree of positional accuracy into the system.

These three demands describe an ideal concept, whose realization in practical systems can, at the present time, be achieved only in parts and in the form of models. Furthermore, temporal aspects connected with modifications in the sets of data would have to be taken into account. These modifications are the result of updating activities on the one hand, and of specific modifications such as new technologies or changes in the legal basis, on the other. Temporal modelling of data quality is not further pursued in this paper.

DATA QUALITY MODELLING

The requirements of accuracy description lead to the formulation of a hierarchy for error management (Veregin 1989), which comprises error reduction management, error propagation, error detection and measurement and error source identification. Decision-making in GIS requires modelling and therefore only error information that is contained in the GIS model can be processed.

As mentioned above, accuracies of quantities and their functions can be expressed in mathematical terms. In (Kraus 1992) the effects of accuracy information on slopes, which were derived from DEM are investigated in the example of an analytically formalized surface. The description of accuracy characteristics has a long tradition. On the basis of point determination by means of geodetic measuring and evaluation procedures, adjustment computation has come a long way. This procedure consists of assumptions on the distribution of errors and of statistical methods of determining the most probable values and accuracies. For other procedures of spatial data acquisition, such as digitizing and subsequent transformation, adjustment was adapted and modified accordingly. The central task hereby is to express the adjustment model mathematically.

The coordinates and also the points are frequently classified depending on the coordinate accuracy that has to be met. This classification is prescribed by law. The resulting standardized accuracy is then to be used for all further processing. Thus the standardized value is substituted for the original one, or, in other terms, the original, real accuracy model is replaced by a standardized one. It should be noted that in the process of standardization also the correlation between point coordinate values is suppressed. The advantage of this procedure is that it is not related to a particular location in the map. As to attribute accuracy the need for classification is even greater (Veregin 1989.)Sensitivity analysis makes it possible to determine attribute error. The effects of attribute errors can be quantified by means of a suitability analysis as described in (Lodwick et al. 1990).

For the purpose of analyzing attribute accuracy by means of random samples, verification is carried out. Accuracy statements are then derived by means of statistical methods. These statements are, however, influenced by spatial references obtained through the random samples. (Dunn 1990) analyzes positional accuracy in digitizing land use. Also Monte Carlo simulations have been variously used for this purpose (Fisher 1991).
The influence of systematic effects can, however, lead to distortions. For modelling in GIS, data quality should be structured as follows:

1. A formal set of DATA QUALITY parameters is to describe an area of reality by means of statistical and logical procedures in a form that can be verified by the computer. This group is to comprise accuracies, sensitivity analysis and logical consistency.

2. The second set can be stored in, but cannot be verified by, the computer. This group contains lineage and in part also completeness.

3. The third set should contain the global stable parameters such as precision, raster width.

The desired combination of several or all data quality parameters for a particular application in order to facilitate processing in databases and also in GIS will largely depend on their belonging to one of the above-mentioned groups. The combination works best if all relevant parameters are contained in group 1. In this case, the scale in which the blueprint or

draft was made se
traditional descripti
Decisions r
Frequently the for
the attribute data c
the requirement o
typical decision pr
characteristic appli
from formalizing
statements relatin
conventional proce

SUMMARY

The invest
primarily aimed at
been positional acc
are to be taken in
envisaged applicati
The analysis of co
are included, sho
requirements of d
reference. The ulti
to all the elements
quantities that are
making.

REFERENCES

Beard, Kate, M., I
Visualization of S
1001, Castine, Ma
Paper 91-26.

Beard, Kate." po
Buttenfield, Barba
Meeting June 9-1

Caspary, Willi. "
(1992): 360-367.

Dunn, R., Harriso
Digital Databases
Systems 5 (2 199

Dutton, Geoffrey
Buttenfield, Barba
Meeting June 9-1

Fisher, Peter. "Mo
Journal of Geogr

draft was made serves as a global data quality parameter. Due to its analytical character, this traditional description also supports the sense of spatial accuracy.

Decisions relating to land estate property will draw on this error model to a high degree. Frequently the form of processing, such as photogrammetry or digitizing, is added. Quality of the attribute data can also frequently be derived from the form of processing applied, without the requirement of direct spatial references. The aim of further investigations is to classify typical decision processes in GIS according to their data quality structure. For this purpose, characteristic applications are being analyzed in greater detail in the form of case studies. Apart from formalizing these specific data quality parameters it is being tried to define quality statements relating to decisions made in GIS and to compare them with those assumed for conventional processing.

SUMMARY

The investigations of data quality in GIS that have been carried out so far were primarily aimed at the description and analysis of accuracies. The main object of analysis has been positional accuracy and, in specific cases also attribute accuracy. Which other parameters are to be taken into account, i.e. consistency, completeness, or lineage will depend on the envisaged application.
The analysis of conventional processing sequences, in which also their data quality parameters are included, should improve modelling. Special GIS applications can, as far as their requirements of data quality are concerned, be grouped together by means of a scale of reference. The ultimate aim is to be able to describe data quality comprehensively with regard to all the elements of a GIS. In special cases also a weighted combination of all the required quantities that are to be contained in a parameter seems possible for a specific form of decision making.

REFERENCES

Beard, Kate, M., Buttenfield, Barbara and Clapham Sarah B.,"NCGIA Research Initiative 7: Visualization of Spatial Data Quality Scientific Report for the Specialist Meeting 8-12 June 1001, Castine, Maine" National Center for Geographic Information and Analysis, Technical Paper 91-26.

Beard, Kate." position Statement on Visualization of Data Quality" in Beard, Kate and Buttenfield, Barbara *Initiative Seven - "Visualization of the Quality of Spatial Data" Specialist Meeting June 9-12 1991 Castine*

Caspary, Willi. "Qualitätsmerkmale von Geo-Daten." *Zeitschrift für Vermessungswesen* 7 (1992): 360-367.

Dunn, R., Harrison, A.R., and White, J.C. "Positional Accuracy and Measurement Error in Digital Databases of Land Use: An Empirical Study." *International Journal of Geographic al Systems* 5 (2 1991): 193-208.

Dutton, Geoffrey. "Probability Filtering for Fuzzy Features" 1991 in Beard, Kate and Buttenfield, Barbara *Initiative Seven - "Visualization of the Quality of Spatial Data" Specialist Meeting June 9-12 1991 Castine*

Fisher, Peter. "Modelling Soil Map-Unit Inclusions by Monte Carlo Simulation." *International Journal of Geographic al Systems* 4 (4 1990): 385-398.

Frank, A. U. (1987). Overlay Processing in Spatial Information Systems. AUTO-CARTO 8, Eighth International Symposium on Computer-Assisted Cartography, Baltimore, MD, ASPRS & ACSM.

Goodchild, Michael F., Kemp, Karen, K.(ed.)" Technical Issues in GIS - NCGIA Core Curriculum" National Center for Geographic Information and Analysis, 1990 units 45, 46.

Goodchild, Michael F. *Accuracy of Spatial Databases*. National Center for Geographic Information and Analysis, 1992. Final Report

Kraus, Karl. " Analysis of Geographical Data and Visualization of Their Quality" in *International Society for Photogrammetrie and Remote Sensing XVIIth Congress in Washington, D.C.*, edited by Lawrence W. Fritz and James R. Lucas 741-748, 1992.

Lodwick, Weldon A., Monson, William, and Svoboda , Larry. " Attribute Error and Sensitivity Analysis of Map Operations in Geographical Information Systems: Suitability Analysis. *International Journal of Geographic al Systems* 4 (4 1990): 413-428.

Veregin, Howard. "Error modelling for the map overlay operation." In *Accuracy of Spatial Databases* , ed. Michael Goodchild and Sucharita Gopal.
3-18. Taylor & Francis, 1989

THE C

ABSTRACT

In this paper, a
the Republic o
a digital form,
efficiently solve
and to set up a
software, devel
digital land cad
we do not pro
corporate datal

1. INTRODUC

The Republic o
with approxima
great emphasis
For these reas
pressing tasks
are becoming a
our society. Th
initial difficulti
company has l
beginning.

The first, totall
meet the needs
and 1991. Nov
These informa
connect the wo
both a qualita
towns in our co
the port town
Heureka Igea
system. SIGIS
***) layers in ve**
- administrativ
- geodetic poin
- land cadastre

EGIS '93

Fourth European Conference and Exhibition on Geographical Information Systems

Genoa, Italy, March 29 - April 1, 1993