THE ROLE OF QUALITY INFORMATION IN THE LONG-TERM FUNCTIONING OF A GEOGRAPHIC INFORMATION SYSTEM

NICHOLAS R. CHRISMAN, *University of Wisconsin, Madison, Wisconsin, USA*

BACKGROUND: TWO OUTMODED MODELS OF MAPS

The development of automation in cartography has finally progressed beyond the stage of marvelling that a computer can make a map. Maps produced by the computer should no longer seem novel, even to the layman. Yet, the digital age has come with a crab-like stride. Computers get faster and storage gets bigger. Resolution and accuracy of many devices improve, but our ideas do not keep up with material progress. There are two attitudes about maps which deserve particular attention because each, in a different way, hinders full exploitation of automated cartography.

Model 1: The Map as Graphic Artifact

Maps have a tangible reality as graphic images. The images consist of symbols used to represent spatial information, both position and attributes. As an automated drafting machine, a computer can plot back a stored map that mimics the traditional product. This achievement may be useful in a limited way, but a pantograph does not deal with the information portrayed by the map – the reason for making a map in the first place.

Concentration on the graphic product alone has trapped cartographers for years. Just as any group hates to admit ignorance, cartographers in the past abhorred blank spots. The heraldic beast may have vanished, but conjecture and surmise are still packaged into a slick graphic presentation that obscures the variations in our knowledge. We have developed expectations, such as smooth contour lines, which are not always supported by adequate evidence.

Model 2: Data Structures based on Spatial Logic

Pleas to examine information content as the basis for digital data structures are not new. At the first *Auto-Carto*, I gave a paper showing the impact of different data structures (Chrisman, 1974). The topological model I advocated has received full theoretical treatment by now (Corbett, 1979). While the theoretical work may have convinced a few, the model has been adopted mostly to solve practical problems.

I am still convinced that the topological approach to map information is necessary; I am no longer convinced that it is sufficient. The topological abstraction is linked to the graphic elements of the traditional map. The model links points, lines and areas according to their tangible connections. The topological relationships have an undeniable role in the internal consistency of the map information, but all other relationships are considered 'attributes' for thematic mapping or record keeping. This formulation does not have the flexibility to handle certain relationships which are crucial to the long-range functioning of a geographic information system.

LONG-RANGE FUNCTIONING OF A GEOGRAPHIC INFORMATION SYSTEM

The term 'geographic information system' (hereafter GIS) is almost dangerously vague. Software is sold as a GIS which may only amount to a computer-assisted drafting station. I would like to reserve the term for a complex type of software which can handle the whole life cycle of spatial information.

Duecker (1978) has identified an important distinction between routine and non-routine systems. Non-routine covers the single purpose, one-shot data base effort, while routine implies an established mechanism to maintain the data for the foreseeable future. In the early years of automation the non-routine had to be dominant due to the experimental nature of the technology. Much of the current GIS software reflects its origins in these non-routine projects; after a massive input phase, the data are considered to be static. Virtually all software with academic and government origins follows this pattern. GIRAS (Mitchell and others, 1977) is an example of a government project with ambitions of Retrieval and Analysis built into its acronym, but the realities of data base production gobble up most resources. The GIRAS data, like many similar projects, consist of a snap-shot of land use. The data structure has no need to record how each line is determined because the same process applies to all.

As a further example, the ODYSSEY system is finally being marketed by Harvard as 'Harvard's GIS'. While the nomenclature may be necessary for marketing, I tried to make a distinction while it was being developed (Chrisman, 1979). The software was designed as a collection of processors to manipulate geographic information. These processors still represent the state of the art for their special functions, but ODYSSEY does not perform all of the data base management functions implied in the broader term GIS.

More provocatively, I would assert that no available commercial software provides a full GIS. While ODYSSEY and GIRAS and other non-commercial developers at least adopted the clarity of the topological model, many commercial groups considered it too complicated (these statements will not be found in corporate literature, but come from personal communication). The commercial groups are responding to the profession's attachment to the map as a graphic product. Yet, in the real meaning of an information system, the computer must have more structure to its data base than merely replotting cartographic spaghetti.

My definition of GIS is strict and my conclusion is that no real GIS has yet been

implemented. Hundreds of systems have been installed, increasingly for routine processing. At first the task is similar to a one-shot project; the backlog of parcel maps (or whatever) must be digitized (Hanigan, 1979). Eventually, these operations plan to switch to routine maintenance. In a refreshingly frank paper, a group working for the City of Milwaukee has discussed the process of getting past the input phase (Huxold and others, 1982) and they specifically mention the underestimation of the maintentance aspect. The current tool is the graphic editing station which assumes that maintenance will mirror the old cartographic process. In the rest of this paper I will try to demonstrate how this concept of routine functioning is inadequate.

QUALITY INFORMATION: A MISSING COMPONENT

A full-fledged GIS can not simply record spatial data, it must also store and understand how these facts are known. This component can best be described as the data quality dimension of a data base. Quality information provides the basis to assess the fitness of the spatial data to a given purpose, and it also provides the handle for long-term maintenance.

The quality of cartographic information seems an obvious concern. An 'accurate map' is part of the popular mythology of cartography, but the profession spends little time on this problem. Few map users notice (or would even care about) the lack of a National Map Accuracy statement at the bottom of a topographic map.

As in many other situations, the development of automation has forced a reevaluation of received opinions and accepted practices. Perhaps, the graphic nature of traditional maps precluded some abuses. Numbers in a data base create an illusion of accuracy and the computer opens new ways of potential abuse. The quality of digital data is an integral part of the information content of the data base. New data structures will have to evolve to encode the quality component, particularly for long-term, routinely maintained projects.

Quality information is not a synonym for positional accuracy measures, although some groups see little else that affects quality (Canadian Council on Surveying and Mapping, 1982). In a standards effort for the USA, the American Congress on Surveying and Mapping's National Committee for Digital Cartographic Data Standards (NCDCDS) (Moellering, 1982) has established a Working Group on Data Quality, as one of four working groups. The next few paragraphs summarize the deliberations of this group (Chrisman, 1983), but they are interpreted in a framework of personal opinion which does not necessarily reflect the views of the working group.

In the opinion of the working group, the foundation of data quality is to communicate information from the producer to a user so that the user can make an informed judgment on the fitness of the data for a particular use. Within this goal, the first responsibility of a producer is to document the lineage of the data. A lineage report traces the producer's work from source material through intermediate processes to the product. In many cases, cartographic agencies have procedure manuals and other documents which contain the relevant information, but this information is not usually considered of great public interest. For

example, the description of computer processes and data structures for GIRAS appeared in the widely-disseminated Geological Survey Professional Papers (Mitchell and others, 1977), while the description of the compilation procedures for the project was placed in the Open File Report series (Loelkes, 1977). In this case, at least the lineage can be constructed from public records. In the case of smaller mapping agencies (at the county or municipal level that accounts for a large proportion of the annual cartography budget, [see Larson and others, 1978]), lineage information may be in the memory of one person, and retirement wipes the slate clean.

Beyond a narrative of lineage, a quality report should include quantitative measures to help a user evaluate applicability. Since geographic information has attribute and temporal components, along with positional ones, each component should be evaluated. This conclusion of the working group rejects the findings of its Canadian counterpart, which saw fit to ignore all but the positional component:

> '... "up-to-dateness" has been interpreted by the Committee as 'date of cultural validity'. As applied to digital topographic data, 'completeness' was deemed impossible to quantify by the Committee; instead, it was proposed that the list of feature classes actually contained in the file be furnished.' (Canadian Council on Surveying and Mapping, 1982, p. 6)

In contradiction to these findings, temporal information can be subjected to tests (e. g., field checking photo-revisions). The more dramatic problem is the blindness to 'completeness'. It is not enough to list the feature codes used. It is necessary to evaluate how consistently features were assigned to classes and how exhaustive the classes were in the actual context. Contrary to the Canadian committee's statement, procedures to evaluate classification accuracy are widespread in remote sensing and other fields (e. g., Fitzpatrick-Lins, 1978; Turk, 1979), while evaluation of logical integrity of a data structure is a fundamental and valuable outgrowth of the topological model (Corbett, 1979). A broad coalition of disciplines must contribute to the components of quality assessment.

Arguing the relevance of temporal and attribute components does not reduce the importance of the positional component. The Canadian draft standards, as well as the efforts of the American Society of Photogrammetry (Merchant, 1982), provide a solid contribution. Still, most work has concentrated on 'well-defined' points, and the extension to more complex natural features may involve additional issues (Chrisman, 1982). Furthermore, estimates of error in position need to be converted into a form which relates to the user's application (e. g., bounds on areas).

The Working Group foresees a range of testing procedures, falling along a continuum of rigor, to evaluate quality in each component. The least rigorous 'tests' may merely represent deductive estimates. Under controlled circumstances (such as appropriate sampling applied to similar map sheets), a deductive estimate could provide the user with adequate information at a much lower cost to the producer. At intermediate levels of rigor, testing would compare the data to internal evidence or to the source document. The most rigorous test requires an independent source of data of higher accuracy.

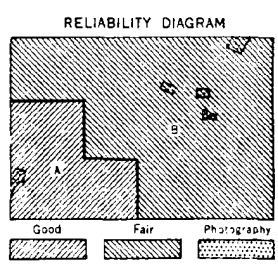
From this discussion it is clear that the National Committee for Digital Cartographic Data Standards is operating inside a charter from a traditional cartographic agency. The emphasis is on a data base product which largely replaces the map graphic product. Certainly standards are needed to ease the distribution of digital data. However, some of the largest impacts of investigating data quality will rebound on the producer.

Quality Information Serves Producers

Whereas the NCDCDS and other national standards efforts have focused on transmitting information to a user to evaluate aptness for an application, the same quality information should serve the producer as well. Recording how information was obtained is a normal cartographic function which has moved into digitial applications without great reexamination. For instance, the Houston METROCOM project creates a 'sheetless' map, but records source and some undefined quality assessment for the original sheets (Hanigan, 1983). While the input sheet correctly identifies the origins of the data, quality information will not remain forever tied to these units. In maintaining Houston's parcel map, updating will be sporadic and scattered. Each update has a different pedigree which should be recorded. Over the years, the process of maintenance will fragment the lineage and quality information.

Many map sheets show a reliability diagram as a part of the legend, displaying an important evaluation of quality variations (Figure 1). In a digital era, this 'diagram' should be an overlay, registered to the rest of the map and integrated into the data structure. Spatial variations in quality can go to the entropic extreme of a separate evaluation attached to each data item. In an application such as navigation or military intelligence with a high premium on reliability, this complete disaggregation is normal. At this limit, the storage of quality information expands from a negligible single figure per sheet to occupy a large fraction of the data base. Adding one word per coordinate, or fifty percent of file bulk, is a dramatic threat to system performance, but some sort of quality information may be fully justified.

This discussion has established the general nature of quality information. The



[503]
Edition 1-AMS (First Printing, 6-59)

Prepared by the Army Map Service (SNTT), Corps of Engineers, U. S. Army, Washington, D. C. Compiled in 1955 from: Bangka 1:50,000, Directorate of Military Survey, Sheets 35-XXVIII-B and 36-XXVIII-A, 1944; Sumatra, 1:100,000, Topografische Dienst, Batavia, 1918-25; Sumatra 1:200,000 Topografische Dienst, Batavia, Sheets A and B, 1924; Netherlands Hydrographic Chart 52, 1951; USHO Chart 1266, 1944; Indonesian Hydrographic Chart 104, 1951. Names processed in accordance with rules of the U. S. Board on Geographic Names. Road classification should be referred to with caution. The reliability of vegetation information is undetermined. Names for symbolized populated places are omitted where information is not available or where density of detail does not permit their inclusion.

FIGURE 1. *Lineage and reliability information from* AMS *1:250,000 sheet of Sumatra, Indonesia.*

following sections will provide some examples of the reasons why the quality component is necessary to the long-term functioning of a digital data base.

Centrality of Control to Positional Quality

A cartographic data base is distinguished from other computer applications mainly due to the representation of physical space. The special focus on spatial properties does not deny the relevance of tabular (attribute and temporal) data; it merely accepts the spatial problems as peculiar and critical.

In the construction of a map, the nature of geodetic control has a direct impact on positional quality. To some extent, control is an eternal verity akin to motherhood and apple pie. Yet, no agency can invest in first order control for all coordinates of interest. Control is expensive and must be used parsimoniously. Though new technology for geodetic surveying (e. g., Counselman, 1982) may revolutionize the field, it will still require hard economic choices. Some advocates of the multipurpose cadastre place the geodetic network as the initial phase. This grand densification of control can be demonstrated in a few current projects (e. g., Bauer, 1976; Hanigan, 1983). Massive investment does help bring a project up to a higher standard of quality, but it does not avoid a fundamental problem. Change in control is inevitable. No matter how complete the original network, new coordinates will trickle in due to normal progress and unrelated projects. In the long-term management of a GIS, it will be necessary to readjust coordinates to account for changes in control. Data structures and procedures for these adjustments must be developed.

The impact of change in control will be a distortion of the preexisting coordinate system. This distortion can be seen as a displacement field or surface. This field represents the displacement distance and orientation as if measured directly from a 'rubber sheet'. Rubber sheet distortions might be recognized as a form of witchcraft or as a pragmatic necessity in automated cartography, but there is little discussion of alternative algorithms and data structures to perform them. Petersohn and Vonderohe (1982) demonstrate that the choice of adjustment model (affine versus Helmert's projective) makes a difference in the result. Usually a programmer picks a method for numerical ease, not specific relationships to systematic errors.

Beyond numerical properties, there is a need for a data structure to manage the distortion surface and the control network. Hybrid data structures, such as those proposed by Brassel (1978), may provide the most likely alternatives. However, in many cases, the distortion of new control is not a simple surface effect. Many measurements are made relative to others, such as the linkage of property lines to section corners in the Public Land Survey. In the data base, an absolute coordinate may be recorded, and the relationships would not be recorded. A full GIS must find a method of dealing with dependencies between data items. Some relationships may be spatial and properly handled by surface data structures, while others may require explicit encoding.

To summarize, control is a foundation for positional accuracy, but it is bound to be readjusted from time to time. Any long-term information system must have procedures and data structures to carry out the readjustment in a manner which fits the nature of the measurements.

Quality in Classification

Quality in attributes can take many forms, but it is representative to restrict attention to the case of nominal attributes – the problems of classification. Apart from terrain and geophysical applications, the overwhelming majority of GIS applications concern some type of discrete phenomena. Topographic feature codes, place names, geocodes, parcel identifiers, land use types, all fall into the same broad group. The discussion above of the NCDCDS work mentions some procedures to examine the accuracy of attributes. These methods have been developed for the one-shot application so typical of current projects, particularly those using remote sensing. In addition to these procedures there is a need to develop methods applicable to the multi-layered environment of a full-fledged GIS.

For example, a GIS may have a topographic component showing rivers and streams. It may also include a floodplain determination (usually from a different or derivative source). The data structure of the GIS should be informed of the set relationship implied between a stream and its floodplain. This should ensure that each stream has a floodplain (or an explanation for not having one). In addition, logical impossibilities, such as rivers meandering in and out of their floodplain, should be detected. A GIS should be able to check for many attribute errors by using one layer to check another. Multi-layer comparisons demand *efficient* polygon overlay procedures which are not available in many systems.

Some elements of quality in classification have a map form, which can be most clearly demonstrated in the practices of remote sensing. A remote sensing classification can be unsupervised where only statistical parameters are used, but often supervised procedures are used. Supervision requires an operator to select some areas as typical of a target class. In order to document the derivation of a supervised classification, the locations of these areas, or training sets, is necessary. Once a classification is developed, it can be verified by a testing procedure such as a 'ground truth' sample. In general, for any classification procedure, it is important to know where it has been developed and validated. Training sets and ground truth samples may be acquired to perform a hidden function, but they should become another layer in the complete GIS.

Temporal Effects

Cross-validation of sources provides a powerful tool, but it demonstrates a major difficulty in quality assessment. Many have commented that polygon overlay leads to spurious results, such as the mismatch of river and floodplain mentioned above. The problem may not be the fault of the overlay process, but in the original sources. Many layers which are fed into a GIS are not fully comparable with the others, yet the comparison has to be made somehow. Some problems of comparability can be assigned to positional inaccuracy or differences in classification, but many also involve time. The most likely explanation of the river/floodplain inconsistency is that the two maps represent different, valid maps from different years. After ten or fifty years a river may move far enough to create the logical impossibility. Time, then, is an important component of quality information. Proper use of temporal reference could help explain these anomalies and ensure a reasonable resolution of the problem. Fur-

thermore, the long-term maintenance of a GIS should lead to simultaneous updating of features so that inconsistencies are avoided.

In some cases, a GIS records not just a single map layer, but its evolution over time. At any one time a traditional map coverage (as recorded by a topological structure) should be available. Basoglu and Morrison, for example (1978), constructed a hierarchical data structure which gave each boundary a time component. While this approach can be constructed to give a proper result, it requires very careful manual data entry.

The quality of temporal data can be subjected to the same analysis applied to spatial representation. Since time can be divided into many periods, it is impractical to test exhaustively. An alternative approach would create a polygon map using all lines from all times. This network will identify all the entities with a distinct history. By assigning temporal codes to these areal entities, there is only one map to check for completeness, plus a simple check for historical validity for each area.

SUMMARY

Space, time and attributes all interact. Quality information forms an additional dimension or glue to tie these components together. Innovative data structures and algorithms are needed to extend our current tools. No geographic information system will be able to handle the demands of long-term routine maintenance without procedures to handle quality information which are currently unavailable.

ACKNOWLEDGEMENTS

Funds to produce this paper were provided by the University of Wisconsin Graduate School. The ideas germinated through participation in the ACSM National Committee for Digital Cartographic Data Standards as chairman of the Working Group on Data Set Quality, but the views expressed are strictly personal.

REFERENCES

BASOGLU, U. and MORRISON, J. 1978. The efficient hierarchical data structure for the US historical boundary file: *Harvard Papers in GIS*, vol. 4.

BAUER, K.W. 1976. Integrated large-scale mapping and control survey program completed by Racine County, Wisconsin. *Surveying and Mapping*, vol. 36, no. 4.

BRASSEL, K. 1978. A topological data structure for multi-element map processing. *Harvard Papers in GIS*, vol. 4.

CANADIAN COUNCIL ON SURVEYING AND MAPPING 1982. *Standards for quality evaluation of digital topographic data*, Energy, Mines and Resources, Ottawa.

CHRISMAN, N.R. 1974. The impact of data structure on geographic information processing. *Proceedings, Auto-Carto 1*, p. 165–181.

——— 1979. *A Multi-dimensional projection of* ODYSSEY, Harvard Laboratory for Computer Graphics, Cambridge.

——— 1982. A Theory of cartographic error and its measurement in digital data bases. *Proc. Auto-Carto 5*, p. 159–168.

——— 1983. Issues in digital cartographic data standards: a progress report, in *Report 3*, NCDCDS, Columbus, Ohio.

CORBETT, J.P. 1979. *Topological principles in cartography*, US Bureau of the Census, Washington, DC.

COUNSELMAN, C.C., 1982. The macrometer interferometer surveyor. *Proc. Symp. on Land Information at the Local Level*, Univ. Maine, Orono, Maine.

DUECKER, K.J. 1979. Land resource information systems: a review of fifteen years experience. *Geo-Processing*, vol. 1, p. 105–128.

FITZPATRICK-LINS, K. 1978. Accuracy and consistency comparisons of land use maps made from high altitude photography and Landsat imagery. *J. Research USGS*, vol. 6, p. 23–40.

HANIGAN, F.L. 1979, METROCOM: Houston's metropolitan common digital data base – a project report. *Surveying and Mapping*, vol. 39, p. 215–222.

——— 1983. Houston's metropolitan common data base: four years later. *Surveying and Mapping*, vol. 43, p. 141–151.

HUXOLD, W.E. and others 1982. An evaluation of the City of Milwaukee automated geographic information and cartographic system in retrospect. Paper presented at Harvard Computer Graphics Week.

LOELKES, G.L. 1977. Specifications for land use and land cover and associated maps. *Open File 77–555*, USGS, Reston, Virginia.

MERCHANT, D. 1982. Spatial accuracy standards for large-scale line maps. Paper presented at ACSM-ASP Annual Meeting, Denver, Colorado.

MITCHELL, W.B., and others 1977. GIRAS, a geographic information retrieval and analysis system for handling land use and land cover data. *U.S. Geological Survey Professional Paper, 1059*, U.S. Geological Survey, Reston, Virginia.

MOELLERING, H. 1982. The goals of the national committee for digital cartographic data standards. *Proc. Auto-Carto 5*, p. 547–554.

PETERSOHN, C. and VONDEROHE, A.P. 1982. Site-specific accuracy of digitized property maps. *Proc. Auto-Carto 5*, p. 607–619.

TURK, G. 1979. GT Index: a measure of the success of prediction. *Remote Sensing of the Environment*, vol. 8, p. 65–75.