GIS DATABASE DESIGN CONSIDERATIONS: A DETERMINATION OF USER TRADEOFFS

Kristina M. Brooks
System Development Corporation
2500 Colorado Avenue
Santa Monica, CA 90406

ABSTRACT

This paper describes an experiment conducted by the author to identify the kinds of tradeoffs users are willing to make concerning levels of detail and cost in utilizing a geographic database. Two case studies were developed to measure the importance of several factors: resolution, land cover classification and cost in the utilization of a land use/land cover database. Case study participants were asked to evaluate the utility of a number of hypothetical databases to solve problems presented in the case studies. In a subsequent exercise, participants were given database development prices and asked to evaluate the cost effectiveness of each database in solving the problems. Although there was a great deal of individual variation in individual ratings, there was a consensus as to the relative ranking of databases in solving case study problems. Price had a definite impact upon participant ratings of effectiveness.

INTRODUCTION

A number of states have established or are considering the establishment of geographic information systems to support natural resource planning and management. Multiagency systems have been proposed, based upon the following assumptions: 1.) a number of state agencies are engaged in similar kinds of activities and require similar analytical capabilities; 2.) many agencies require similar kinds of information and could develop cooperative databases; and 3.) the sharing of a single GIS and cooperative databases would reduce information collection, analysis, storage and retrieval expenses. The determination of user data and analytical requirements is important to the development of any information system, but the identification of the requirements for a multiagency system is complicated by the number and diversity of potential users.

IDENTIFICATION OF USER REQUIREMENTS

User involvement in the design of information systems is stressed by a number of authors (Calkins & Tomlinson 1977, Dueker 1979, Kennedy & Gunn 1976), but specific techniques and instruments to identify user requirements are rarely

described. A number of of user surveys have been conducted in conjunction with the design of a GIS (Calkins & Marble 1978, Gordon 1979, Salmon et al 1977), but it has been difficult for users and system designers to evaluate potential GIS capabilities and databases through traditional interview and questionnaire techniques.

OREGON GIS REQUIREMENTS SURVEY

In 1979-1980, the author conducted a user needs survey for the State of Oregon in conjunction with a GIS design study (Brooks 1980a). Approximately 65 interviews were held, in person or by telephone, with potential users of or contributors to the proposed GIS in 13 natural resource agencies. Each program described in the interview was documented and the documentation was returned to the interviewee for review and comment. A ten page matrix of data characteristics was then developed and distributed to the interviewees to obtain additional information about data and analytical requirements. At the same time, a data inventory was conducted to identify data collections, automated or manual, which were being produced or used by the agencies included in the requirements survey.

The survey did provide useful information about the types of thematic data planners and managers require, their preferences regarding levels of detail for a number of data characteristics and their requirements for analytical capabilities. However, many of the participants were not intimately acquainted with the data they used and were not able to articulate what their requirements were or would be. This made it impossible for the project investigators to compare the results within and among departments and make judgements about the feasibility of developing shared databases, an important consideration in this design study. It was also not possible to rank the relative importance of data collections. Conversations with investigators who have conducted similar surveys have indicated that this is fairly common. Data users rarely think about the data they use in the terms a GIS designer or manager would use to determine the appropriateness of a given dataset for a GIS database (Brooks 1980b). Secondly, it is difficult for users to evaluate the utility of potential information products which they are not using at present. Finally, it is difficult for users to rate the effectiveness of potential databases or GIS products without considering development costs. GIS databases can vary considerably in terms of the level of detail at which the data are compiled and converted to machine readable form. The level of detail chosen will have a considerable impact upon system costs. Traditional interview and questionnaire techniques cannot assess the kinds of tradeoffs data users make concerning these factors.

ADAPTATION OF THE INFORMATION INTEGRATION THEORY

The problems encountered in the Oregon needs survey illustrated the need for a technique to measure the tradeoffs users would be willing to make concerning database characteristics and to identify whether or not a consensus could be achieved by diverse users about the kinds of databases which should be developed. The information integration theory was selected as the basis for a GIS design tool which would meet the above objectives. The information integration theory was developed by N. H. Anderson, an experimental psychologist, as a methodological framework for analyzing how a variety of factors are combined or integrated in the decision making process (Anderson 1970, 1974). Particular levels of each factor can be characterized by two components: a scale value and a weight. Each combination of factors is rated on a continuous scale rather than rank ordered. The relative importance of individual factors can be assessed as well as the relative desirability of various combinations of factors or options. The options are presented in a factorial design; a case study which involves consideration of three levels of three factors (3 # 3 # 3) results in 27 possible combinations. The relative importance of each factor is determined using analysis of variance techniques. Unlike other decision analysis techniques, participants do not have to articulate the importance they place on each factor out of context; factor importance is inferred from their responses. The theory was developed by Dr. Anderson to study personal decisions, but it has been adapted to study a number of other types of decisions, including consumer purchasing decisions (Levin 1976a, 1976b).

DEVELOPMENT OF THE CASE STUDIES

Two case studies were developed to measure the relative importance of two factors, resolution and land cover classification, in the utilization of a GIS database for natural resources planning. The case studies had to be general enough so that planners from a number of agencies could relate to the problems, specific enough to be realistic, and appropriate for computer analysis techniques. The first case study involved the evaluation of potential impacts of a proposed timber harvesting plan on deer and elk habitats in a 50 square mile management unit. Some background information was given about the area and participants were given a number of tasks to complete. Twelve hypothetical databases were presented and participants were asked to evaluate the utility of each on a 10 point scale in solving the problems presented in the case study. In a subsequent exercise, the participants were asked to re-evaluate the databases on a "cost effectiveness" scale and were given database development costs. A time factor was introduced at this stage, so participants actually had to consider 24 hypothetical database options.

The second case study involved an assessment of the conversion of natural resource lands (agricultural lands, range and forests) to urban uses over a ten year period. The study would eventually include the entire state, but would be conducted on a county by county basis, beginning with County Y which contained 4,500 square miles. As in the first case study, twelve hypothetical databases were presented and participants were asked to evalute the utility of each in solving the problems presented in the case study.

In both case studies, the databases differed in two respects: land classification and minimum ground resolution. In both case studies, the USGS land cover/land use classification scheme was used as the basis for the classification scheme to be used for the databases. Levels I and II were taken directly from the USGS classification scheme (Anderson 1976) since they have been standardized for the entire country. Levels III and IV were developed by the author to illustrate the differences in detail between the hierarchical levels of the classification scheme. Levels II, III and IV were considered in case study 1 and levels I, II and III were considered for case study 2. Four levels of resolution were considered for each case study. Resolution was used rather than scale because participants in the preceeding GIS needs survey seemed more comfortable dealing with resolution. Minimum resolution levels of 160 acres, 40 acres, 10 acres and 2.5 acres were considered for case study 1 and resolution levels of 640 acres, 160 acres, 40 acres and 10 acres were considered for case study 2.

Participants in the study included Oregon resource managers and planners from the departments which participated in the GIS survey and Washington planners and managers who were involved in a similar project in the state of Washington. The level of familiarity with GIS systems varied; most participants were aware of the capabilities of such systems, although few had any practical experience with GIS or computer mapping systems. A total of 45 individuals participated in the study, 33 from Oregon and 12 from Washington. The case studies were administered in group sessions. Participants usually took one to two hours to complete both parts of each case study.

DEVELOPMENT OF DATABASE COST MODEL

It was necessary to develop a cost model in order to prepare the second exercise in each case study. GIS vendors were contacted to determine the average cost to digitize and edit polygons. USGS staff provided statistical information on the average number of land use/land cover polygons per square mile for urban and rural areas. The latter information was supplemented by some sampling

experiments done by the author on the reduction in number of polygons when data is reclassified at a higher level (i.e., Level I as opposed to Level II). The following assumptions were made:

1. $3.50 is the average cost of digitizing and editing a polygon;
2. For the most detailed level of classification (i.e., Level IV):
   a. 640 acre resolution = 1 polygon/sq. mile
   b. 160 acre resolution = 4 polygon/sq mile
   c. 40 acre resolution = 8 polygon/sq mile
   d. 10 acre resolution = 12 polygon/sq mile
   e. 2.5 acre resolution = 16 polygon/sq mile
3. Reducing the level of classification detail by moving up the hierarchy would reduce the number of polygons/square mile by the following factors:
   a. Level IV = 1
   b. Level III = .8
   c. Level II = .64
   d. Level I = .625

These assumptions were reviewed by several vendors and GIS users. The purpose of this model was not to predict the cost of a particular database, but to indicate the relative differences in cost at different levels of detail. The database options included in each case study are listed in Table 1.

TABLE 1

CASE STUDY DATABASE OPTIONS

CASE STUDY 1		CASE STUDY 2	
Resolution/Class	Price	Resolution/Class	Price
160 acres/II	$ 480	640 acres/I	$ 7875
160 acres/III	$ 560	640 acres/II	$ 10080
160 acres/IV	$ 700	640 acres/III	$ 12600
40 acres/II	$ 896	160 acres/I	$ 31500
40 acres/III	$1120	160 acres/II	$ 40320
40 acres/IV	$1400	160 acres/III	$ 50400
10 acres/II	$1345	40 acres/I	$ 63000
10 acres/III	$1680	40 acres/II	$80640
10 acres/IV	$2100	40 acres/III	$100800
2.5 acres/II	$1790	10 acres/I	$ 94500
2.5 acres/III	$2240	10 acres/II	$120960
2.5 acres/IV	$2800	10 acres/III	$151200

DATA ANALYSIS TECHNIQUES

Although the information integration theory is usually analyzed using ANOVA techniques, several other statistical tools were used for this application. We were interested

in the degree of consensus as well as individual responses. If participants reached no agreement as to the ranking of the various options, this technique would not be very useful as a GIS design tool. The Kendall W Coefficient of Concordance (Siegel 1956) was used to identify the degree of association among the participants' rankings of the options. This is an appropriate test because it measures relative rankings rather than absolute ratings. Therefore, even if respondent A is a high scorer on a point scale and respondent B is a low scorer, the Kendall W Coefficient will still express the relationship between their ranking of options. The W score is not linear and its significance is tested by an associated chi square value. SPSS ANOVA was used to statistically determine the significance of the factors and their interactions in producing the ratings. Regressions were also run to identify other factors which might have influenced the variability in responses.

ANALYSIS OF RESULTS

There was a considerable amount of variability in individual responses as illustrated in Figure 1.

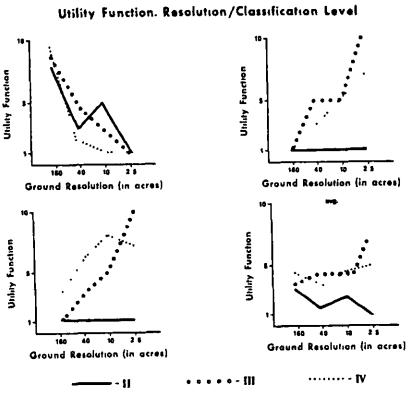

FIGURE 1

The rankings for utility and cost effectiveness for Case Study 1 are given in Table 2. As can be expected, the most detailed information was considered the most useful while the least detailed was considered the least useful. When cost was included as a factor in the cost effectiveness exercise, the order shifted slightly. This is to be expected since the most detailed databases are also the most expensive.

TABLE 2

CASE STUDY 1 - RANKING OF DATABASE OPTIONS

Utility	Cost Effectiveness
2.5 acres/IV	10 acres/IV
10 acres/IV	10 acres/III
2.5 acres/III	2.5 acres/IV
10 acres/III	40 acres/IV
40 acres/IV	40 acres/III
40 acres/III	2.5 acres/III
10 acres/II	40 acres/II
160 acres/IV	10 acres/II
2.5 acres/II	2.5 acres/II
160 acres/III	160 acres/III
40 acres/II	160 acres/IV
160 acres/II	160 acres/II

The interaction between resolution and classification was significant as measured by an ANOVA F score. In the information integration designs upon which this experiment was modeled, investigators were able to identify the relative weights of each factor in the assignment of ratings. However, this is not a meaningful statistic when the data is not linear. The stepwise regression indicated that classification may have a greater bearing on the scores than resolution, but there is a great deal of variability which is not accounted for by these factors. This individual variability is due to several factors:

1. individual differences in the application of a point scale
2. differences in the perception of the problem and its solution
3. different pricing thresholds.

There was considerably more variability in the second exercise in which price was introduced. Although the dollar range was not great for this case study, prices varied from $700 to $2800, price did have a dampening effect on the ratings and on the relative rankings of responses illustrated in Figure 2. The interviews in the GIS needs study and discussions generated during the case study sessions clearly indicated that planners and managers are not used to separating out data costs. They were very sensitive about the potential costs of a GIS, even though it is quite likely that their current data costs are high but hidden.

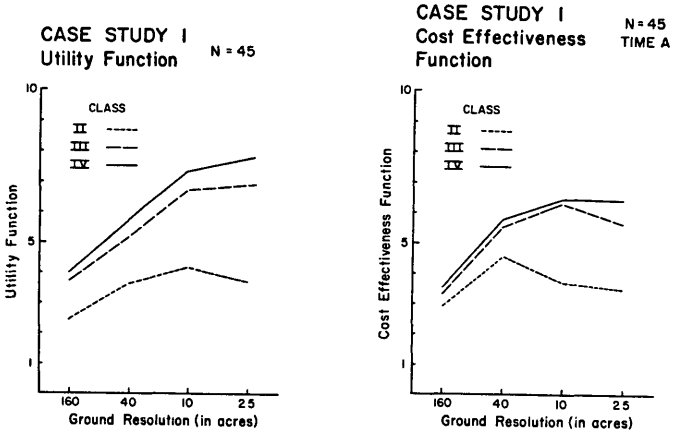

FIGURE 2

There was less agreement about the utility and cost effectiveness of database options in case study 2, although the Kendall W scores were still significant.

Participants expressed greater difficulty in dealing with the case study and were more ambivalent about the data requirements. Again, as shown in Table 3 there were interactions between classification and resolution in the utility exercise, and resolution produced a higher R value in the regression analysis.

TABLE

CASE STUDY 2 - RANKING OF DATABASE OPTIONS

Utility	Cost Effectiveness
40 acres/II	40 acres/II
10 acres/II	160 acres/II
10 acres/III	160 acres/I
40 acres/III	40 acres/I
40 acres/I	40 acres/III
160 acres/II	160 acres/III
160 acres/I	10 acres/II
10 acres/I	640 acres/II
640 acres/II	640 acres/I
160 acres/III	640 acres/III
640 acres/I	10 acres/I
640 acres/III	10 acres/III

The most significant aspect of these results are the differences in the relative rankings of database options in the utility and cost effectiveness exercises, illustrated in Figure 3. In fact, several of the most highly rated

options in the utility exercise were ranked very lowly in the cost effectiveness exercise. Some participants simply could not deal with the higher prices. One of the original case studies in the pretest dealt with a statewide problem, but the cost figures were so high that the pretest participants recommended that the study be scaled down.

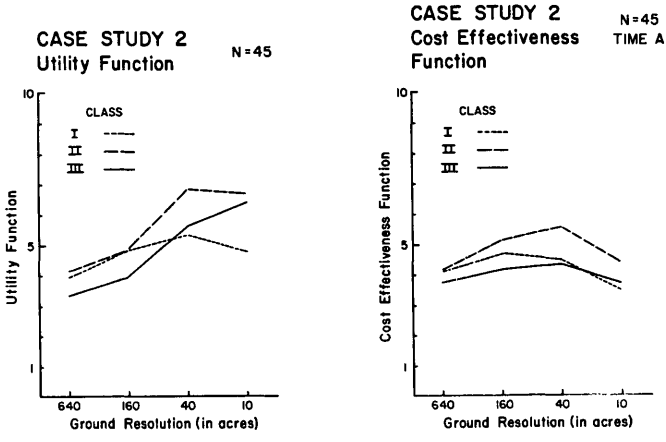

FIGURE 3

CONCLUSIONS AND RECOMMENDATIONS

The specific results of these case studies may not be be generalizable beyond several superficial observations:

1. cost has an impact upon user perception of the value of information.
2. there is a great deal of variability in the evaluation of data even among colleagues with similar backgrounds and responsibilities

The exercise was a useful one as an educational experience for the users. The sessions generated a great deal of discussion about data requirements and forced participants to think about data they use and their relationship to the tasks they perform. The case studies provided a common framework for participants from different departments to discuss the differences in their requirements, i.e. the foresters and wildlife biologists. Such an exercise may also serve as a method of generating consensus or, at least determining whether or not consensus can be reached. A followup Delphi session is planned in which participants will be given their scores and group scores and asked to submit a new set of ratings.

I would make several recommendations concerning the use of this case study factorial approach. Database options were presented in a random fashion, so as not to bias participants into setting up an empirical rule to rate factors.

Some participants actually set up a matrix on paper to develop such a rule while others attempted to do so mentally. The exercises would have taken less time if options had been presented in a matrix and it would have allowed participants to be more consistant in their application of whatever rules they chose to follow. Participants were not allowed to look at their utility ratings when they did the cost effectiveness ratings. More consistant results would have been achieved if this had been allowed. Even though it is rather time consuming to develop realistic case studies, it is far less expensive than conducting bench mark tests or demonstration projects. This approach cannot take the place of such exercises, but case studies can provide provide GIS designers information about user requirements and tradeoffs, resolving some of the problems in user needs assessment.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Kenneth Dueker, Dr. Irwin Levin and Dr. Jon Kimerling for their assistance and encouragement.

BIBLIOGRAPHY

Anderson, James R. et. al. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. Geological Survey Professional Paper 964 Washington DC: U.S. Government Printing Office. 1976.

Anderson, N.H. "Functional measurement and psychological judgement." Psychological Review 77:153-170. 1970.

_____. "Information integration theory: a brief survey." in Contemporary Developments in Mathematical Psychology vol. 2. D.H. Kranz, R.C. Atkinson, R.D. Luce and P. Supes eds. San Francisco: Freeman Press. 1974.

Brooks, Kristina M. Requirements for a Statewide Geographic Information System: A Survey of Agency Data and Analytical Needs. Salem, Oregon: Oregon Department of Forestry. 1980a.

_____. "Survey of data and analytical requirements: implications for the development of a geographic information system in Oregon" Annual Conference of the Urban and Regional Information Systems Association, August 17-21, 1980b. Toronto, Canada. pp. 101-110.

Calkins, Hugh and Duane F. Marble. Long Range Information Need Assessment (LNA) for the Resource Information Display System. Amherst, New York: Geographic Information Systems Laboratory, State University of New York at Buffalo. 1980.

_____ and R.F. Tomlinson. Geographic Information Systems, Methods and Equipment for Land Use Planning. 1977. Ottawa, Canada: Commission of Geographical Data Sensing and Processing, International Geographical Union. 1977.

Dueker, Kenneth J. "Land resource information systems: spatial and attribute resolution issues" in AutoCarto IV Proceedings of the International Symposium on Cartography and Computing. November 4-8, 1979 Reston, VA. Volume II. pp. 328-336.

Gordon, Kenneth. An Evaluation of Environmental Data Use in Computer Assisted Spatial Data Handling Systems: The Results of a Survey of Applications in the Pacific Northwest States. NASA-Ames Research Center, Moffet Field, California. 1979.

Kennedy, Michael and Charles Gunn. "Automated spatial data information systems: avoiding failure" Annual Conference of the Urban and Regional Information Systems Association, Chicago, IL. 1976. pp. 74-87.

Levin, I.P. Attitudinal Modeling of Travel Behavior: Application of the Information Approach of Experimental Psychology. Working Paper 17. Iowa City Iowa: Institute of Urban and Regional Research, University of Iowa. December 1976a.

_____. "Comparing different models and response transformations in an information integration task." Bulletin of the Psychonomic Society 7(1):78-80. 1976b.

Salmon, Larry,, George Nez, James Gropper, John Hamill and Carl Reed. User Needs Assessment for an Operational Geographic Information System. Report FWS/ORS-77/21. Fort Collins, CO: Federation of Rocky Mountain States. 1977.

Siegel, Sidney "The Kendall coefficient of concordance: w" in Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co. 1956. pp. 229-238.