ANALYTIC AND CARTOGRAPHIC DATA STORAGE: A TWO-TIERED APPROACH TO SPATIAL DECISION SUPPORT SYSTEMS

Lewis D. Hopkins
University of Illinois at Urbana-Champaign
1003 West Nevada
Urbana, Illinois 61801

Marc P. Armstrong
University of Iowa
Iowa City, Iowa 52240

ABSTRACT

Data can be analyzed from a data structure that is organized around the relationships that are required to support decisions. This approach is more efficient than analyzing from cartographic data structures. Especially for streams data, the sensitivity to fractal dimension makes direct encoding of analytical relationships more efficient than computation from digital geographic coordinates. Using an analog cartographic tier and a digital analytic tier is one good way to implement such a database.

INTRODUCTION

The purpose of a geographic information system (GIS) is to support data manipulation, analysis, synthesis, and display for decision making. The purpose of display is, in turn, to present data in such a way that the powers of human perception can be brought to bear on interpretation of data. The success of a GIS should, therefore, be evaluated not as a cartographic reproduction device with a fixed level of spatial and attribute resolution, but as a particular kind of decision support system. Bonczek et al. (1981:12) describe a decision support system as "an information-processing system embedded within a decision making system." A database should be designed so that data can be manipulated efficiently. Display of data in cartographic form need not be planimetric, if its purpose is to support interpretation. For these reasons it may be more efficient to organize data to be used for decision support into a separate tier in which they are stored and retrieved from an analytical structure, rather than directly from a cartographic structure.

This analytic tier must be supported in some degree by a cartographic tier, but the cartographic tier may be different in form from a traditional GIS. In particular, it need not be computerized, which may result in significant savings in initial costs compared to traditional approaches. An analytically structured system cannot do everything a cartographic system can, but it can support a prescribed set of tasks more efficiently than a cartographic system because the data can be collected and stored for efficient retrieval for particular analytical relationships.

The two-tiered approach is described here by using the Illinois Streams Information System (ISIS), an information system developed for streams data, as an example. The ISIS system is intended to support decisions on stream classification for management and regulation and to support decisions on permit applications for specific projects. The intended tasks are first identified in Section 2. The use of cartographic data structures to support these tasks is described in Section 3. A two-tiered approach is described in

Section 4. A preliminary evaluation of the two-tiered approach in Section 5 suggests that it can provide substantial savings in storage, retrieval, and initial investment costs.

The gradual development of a computerized cartographic tier, after development of the analytic tier, may be an effective strategy for GIS development. This apparently backward strategy may help to alleviate a problem identified by Dueker (1979): Many land resource information systems in the past were not effective because of mismatches between spatial/attribute resolution and the kinds of decisions the systems were intended to support. Developing the analytic tier first assures a satisfactory match because the database design starts from the decisions, not from the data.

STREAMS INFORMATION SYSTEM TASKS

A data structure designed to support streams management and planning should provide for efficient implementation of the following generic tasks: 1) retrieval of thematic information about a place, including its relationship to other places, and 2) selection of places meeting a specified set of characteristics. In the streams case the places may be points on or in relation to streams, reaches of streams, entire streams, or watersheds. The relationships frequently used, which should therefore be supported efficiently in the data structure, are:

1. upstream of
2. downstream of
3. tributary of
4. distributary of
5. flow distance between
6. right bank, instream, left bank

These relationships are simpler to support than more general geographic data relationships because streams form a nearly hierarchical network. Streams data are, therefore, an appropriate case to illustrate the potential advantages of a two-tiered system. Armstrong (1984) has addressed analytical data structures for other types of geographic data.

These generic capabilities can support a wide variety of streams planning and management tasks, as illustrated by the following examples. If a toxic spill occurs, water intakes within a given downstream distance can be identified. All streams with given percentages of bankside landcover can be identified. Flow at a point can be predicted as a function of arbolate sum (total length of channel) above the point. All characteristics surrounding the proposed location for a new bridge can be listed. The nearest water quality sampling station can be found. The above relationships are sufficient for analysis of stream charateristics given sufficient topological resolution and precision of measurement of flow distance. Hydraulic engineering studies, or detailed fish habitat studies require data with greater spatial resolution than stream classification based on aggregate characteristics, but the same data structure would support each task.

A clear understanding of the relationship between resolution and valid interpretation, which is different from resolution and apparent quality of display, is necessary to compare a cartographic tier based on hard copy manuscripts and field notes, as used in ISIS, to a digitally encoded cartographic tier. The difficulties of determining relationships that depend on the scale of the original source, which determines fractal dimension, are especially critical in streams data. The primary relationships of flow length and right vs. left bank are much more sensitive to fractal dimension than is the area of a polygon or error in whether points are in or out of a polygon. Consider a square, as illustrated in Figure 1. The area will remain almost constant, regardless of the fractal dimension of the lines forming its edges. The location of most points as in or out of a polygon is also relatively insensitive to fractal dimension, especially for large polygons. The length of an edge, on the other hand, increases systematically with an increase in fractal dimension. Also, a point immediately adjacent to the edge, such as on its bank, is likely to switch from one side of the edge to the other as fractal dimension changes.

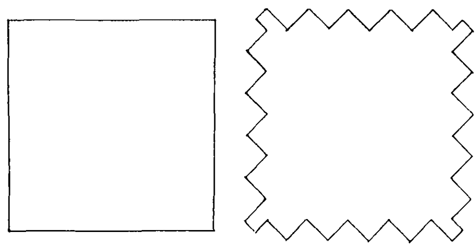


Figure 1: Fractal dimension and accuracy in areas and lines

In the next two sections we describe how the streams analysis tasks can be supported, first by a cartographic data structure then, by an analytically data structure.

CARTOGRAPHIC DATA STRUCTURES TO SUPPORT STREAMS DECISIONS

A widely accepted approach to organizing cartographic data is to use a chain-node data structure. Chains are ordered sets of identifiers for coordinate pairs, and nodes are defined as the meeting of two or more chains. Pointers are used to identify the nodes at which a chain ends, the chains emanating from a node, and the polygons bordered (left-right) by a chain (see Peucker and Chrisman, 1975). Streams in such a system are described as chains of coordinates representing reaches, which connect nodes located at confluences, stream ends, or some other noteworthy break in the stream network. In streams data, chains have among their attributes the direction of flow.

Consider the implementation of Tasks 1 and 2 — Upstream of and downstream of. To find a data item, such as a sampling station, upstream of point A, which is identified by geographic coordinates:

1. Determine which chain (reach) point A is on.
2. Check each successive pair of coordinates in the chain to see if any sampling station coordinates fall within the line between the two points.
3. If the upstream end of the chain is reached, continue search on upstream chains.

Note that at any given resolution, the determination of whether a point is on a line will be ambiguous. Points near confluences or very sinuous reaches will be especially problematic. For consistent analysis, the users must know and apply a set of rules for resolving these ambiguities. For these reasons most geographic information systems store the relationship "point on reach" directly, rather than computing it from geographic coordinates. This relationship is not only stored directly for more efficient retrieval, but is also frequently encoded directly based on determination by a human that the point is on the line by definition.

Next, consider the implementation of Tasks 3 and 4 — Tributary of and distributary of. Given a chain, the following steps are required to identify a tributary:

1. Determine flow direction of the chain from attributes.
2. Identify chains linked to upstream end using pointers.
3. Determine which chain is the mainstem from attributes.
4. All other chains are tributaries.

To identify all tributaries of a given stream the process is repeated. An analogous process can be used for identifying distributaries, by looking at the downstream end of the given chain.

Task 5 — Distance between two points, A and B, requires several distinct operations. For each point:

1. Determine which chain (reach) the point is on.
2. Compute the distance from the point to its nearest neighbor in the chain.
3. Sum this distance and the distances between all other succeeding points in the chain.

Then,

4. Determine which chains are between points A and B.
5. Sum distances for all points in the chains.
6. Sum chain lengths for these chains and distances within the two chains containing points A and B.

It is readily apparent that is is more efficient to compute each length once and store it as an attribute of the chain. Distance computed from the strings of coordinates is inherently at some given resolution (fractal dimension). Distance along a stream is an increasing function of fractal dimension at which it is measured; there is no true distance (See for example, Gardiner, 1982; Klein, 1982).

Task 6 — Determination of right bank, left bank, or in the stream, is very sensitive to resolution. Experience with trying to correctly position points that are near streams, suggests that it is generally necessary to specify the relative position directly. Very small errors in location can lead to incorrect results. If the point location is done first on a manuscript that also contains the stream centerline, then the placement of the point can be purposely moved to avoid ambiguity. Doing so, however, means that the coordinates of that point indicate only relative position and cannot be relied on for determining spatial relationships with items from other manuscripts. Although the actual positioning error is very small, the topological error with respect to streams is obvious. It is difficult to be confident in a dataset that places your town on the wrong side of the river.

Description of the cartographic approach has already suggested some aspects of the analytic approach. The basic idea is to encode directly and store directly the relationships that are frequently used or ambiguous in determination. There is little surprising in analytic data structures, but, importantly, they can be created without the expense of first creating a computerized cartographic GIS from which to compute these relationships. If stream relationships are only to be computed once (or even infrequently for updating), then it is more efficient and accurate to do so external to the computerized GIS environment. This conclusion is made even stronger if one recognizes the resolution required to accurately determine whether points are on lines and the lengths of lines at a fractal dimension sufficient for most data analysis needs.

In the analytic approach the six generic tasks are inherent in the data source. For flow distance between two points it is necessary to do one subtraction operation. If the flow continues to or from a tributary an addition operation will also be required. Relationships of right bank, left bank, instream are not only used to structure the data efficiently for retrieval, but are also the form in which data are encoded. The accuracy of field notes and fully interpreted cartographic manuscripts is thus maintained.

IMPLEMENTATION OF A TWO-TIERED APPROACH

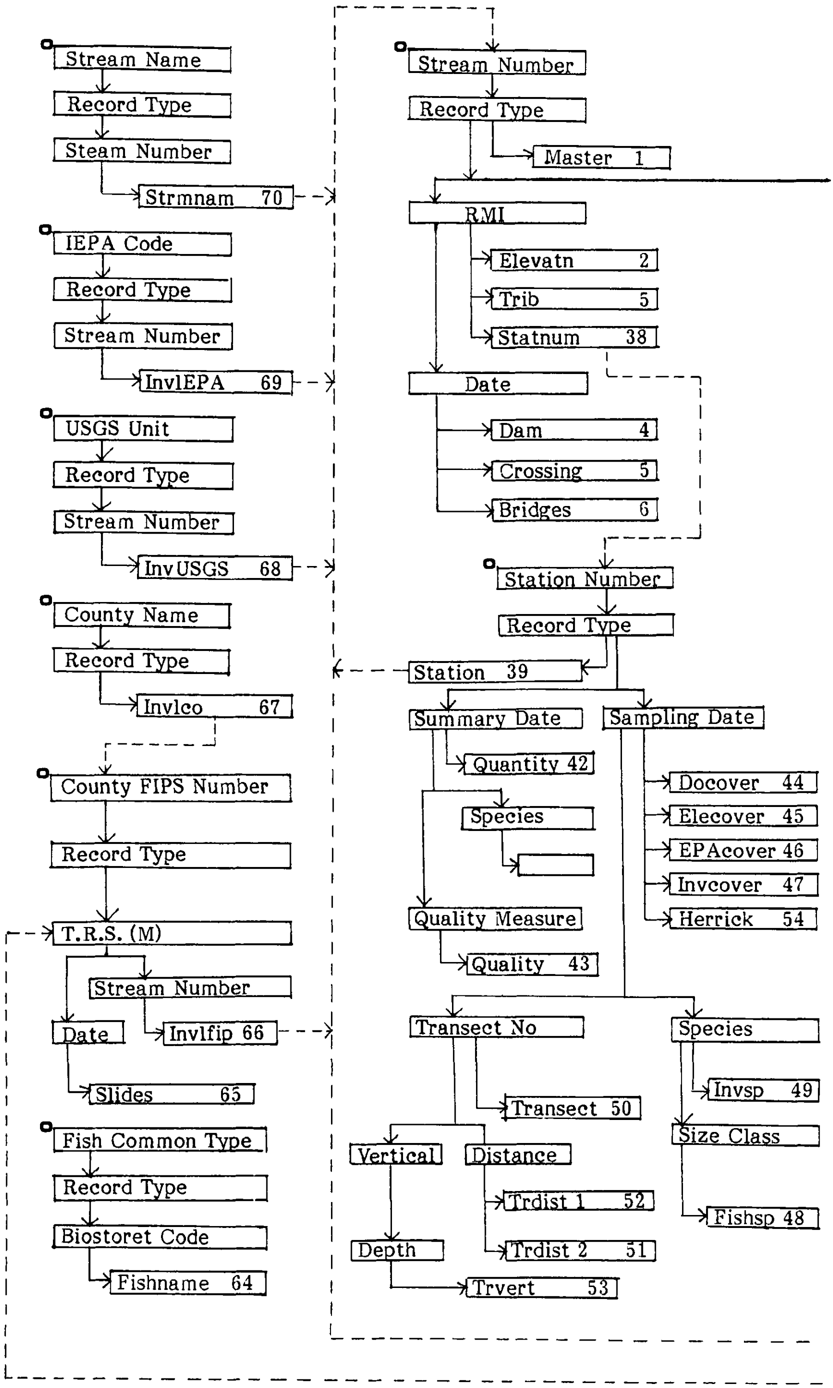
Both tiers, the cartographic and the analytic, must be present in some degree. The cartographic tier might consist only of field notes describing relationships; it might be detailed map manuscripts; or it might be a computerized cartographic GIS. The questions of implementation are:

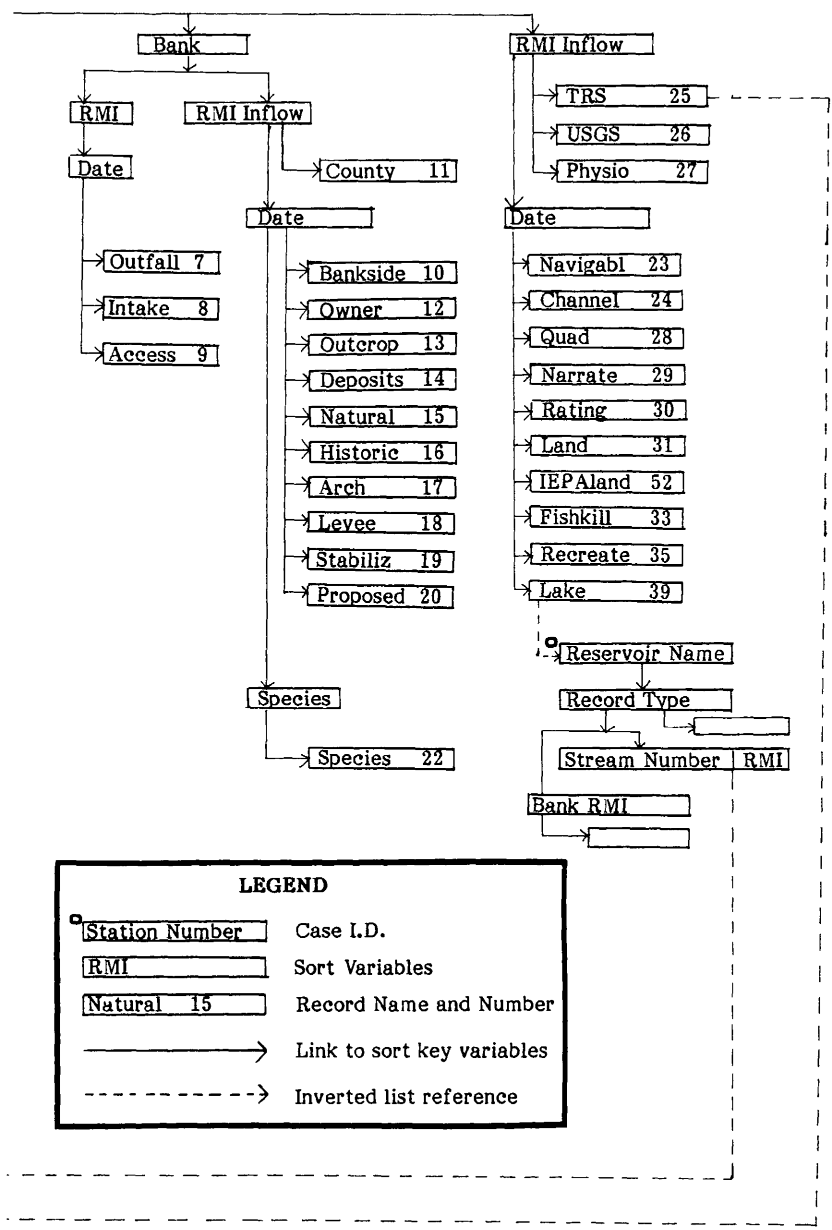
1. What is stored in machine readable form?
2. What data structure(s) are used?
3. What spatial and attribute resolution is supported?
4. What tasks - manipulation, analysis, and display - are supported from what stored data and at what initial and operating cost?

The crux to understanding the two-tiered approach is that decisions are supported not by the digital (cartographic) representation of the map itself, but by interpretations of it. Careful specification of the generic tasks to be supported makes it possible to store the interpretations themselves. For tasks in which the means of interpretation are not well defined, so that a human must make the interpretation, we can still store only the minimum data required for a sufficient thematic map display. Data for thematic display may be much less costly to store and manipulate than planimetrically accurate data of equivalent resolution. If very detailed planimetric data are needed for interpretation by humans, it is now feasible to store the map or photograph itself on video disk at low cost.

The analytic relationships must be computed from field data or some accurate cartographic source. In ISIS, distances are determined from a set of USGS 1:24,000 quadrangle maps. The relationships tributary of and direction of flow are determined from these maps and related sources using many definitions and precedures to resolve ambiguities (see ISIS, 1983, 1984). Most other data relationships are taken directly from field reports or special purpose maps. Cartographic information added to the USGS base is recorded on mylar overlays to allow additional interpretations, and updating of information.

In ISIS, relationships determined from this analog, not digital, cartographic base, are encoded, and stored, using the data structure shown in Figure 2. ISIS currently operates on SIR/DBMS (SIR, 1983). No geographic coordinates are stored, only hydrologic relationships and positions. The stream numbers are hierarchical so that tributary relationships are inherent in the sorted records. RMI is River Mile Index, which identifies position along the stream centerline. ISIS uses streams labeling in which an entire named stream is indexed as a single entity, rather than labeling each reach (chain between two confluences) as an entity. The positions of points on a stream are then given in reference to its mouth and each tributary is referenced to such a point. All data about a named stream can, therefore, be retrieved directly without using pointers to link together reaches. If it is necessary to know the length of a river, it is stored as an attribute of an entity; it need not be computed by linking reaches and summing the lengths. There is no clear advantage derived from using a reach system, because the stream entity can be divided into either arbitrary or meaningful reaches by designating river mile break points. RMI INFLOW is the


Stream Name
Record Type
Steam Number
Strmnam 70
IEPA Code
Record Type
Stream Number
InvlEPA 69
USGS Unit
Record Type
Stream Number
InvUSGS 68
County Name
Record Type
Invlco 67
County FIPS Number
Record Type
T.R.S. (M)
Stream Number
Date
Invlfip 66
Slides 65
Fish Common Type
Record Type
Biostoret Code
Fishname 64
Stream Number
Record Type
Master 1
RMI
Elevatn 2
Trib 5
Statnum 38
Date
Dam 4
Crossing 5
Bridges 6
Station Number
Record Type
Station 39
Summary Date
Sampling Date
Quantity 42
Species
Docover 44
Elecover 45
EPAcover 46
Invcover 47
Herrick 54
Quality Measure
Quality 43
Transect No
Species
Transect 50
Invsp 49
Size Class
Vertical
Distance
Trdist 1 52
Trdist 2 51
Fishsp 48
Depth
Trvert 53



Bank
RMI Inflow
RMI
RMI Inflow
TRS 25
USGS 26
Physio 27
Date
County 11
Date
Date
Outfall 7
Intake 8
Access 9
Bankside 10
Owner 12
Outcrop 13
Deposits 14
Natural 15
Historic 16
Arch 17
Levee 18
Stabiliz 19
Proposed 20
Navigabl 23
Channel 24
Quad 28
Narrate 29
Rating 30
Land 31
IEPAland 52
Fishkill 33
Recreate 35
Lake 39
Reservoir Name
Species
Record Type
Species 22
Stream Number
RMI
Bank RMI
LEGEND
Station Number
Case I.D.
RMI
Sort Variables
Natural 15
Record Name and Number
Link to sort key variables
Inverted list reference

upstream end of a section of stream with a given attribute. Data records sorted by RMI INFLOW identify attributes of sections of stream between two RMIs.

In an analytic data structure, multiple levels of positional resolution can be maintained. The position of a bridge can be specified to the nearest 0.1 mile. Points within a station sampling for fish can, however, be located relative to the bridge, and therefore to each other, to the nearest one foot using field data.

This approach allows detailed transect data to be stored. It is difficult to handle such multiprecision relationships within most available GIS's. Multiple precision analytic structures for streams, on the other hand, can be handled straightforwardly in most data base management systems.

Topological relationships and detailed relative postions (e.g., bank position of discharger) were encoded directly from field data, rather than computed from geographic coordinates, even though digital cartographic data were available. This digital cartographic tier could not be used because the digital data were at 1:250,000. Empirical comparisons of potential errors showed that 1:250,000 USGS stream lines were on the average .85 of the length of 1:24,000 stream lines. Also, some streams are not included on the 1:250,000 data. Because of severe resolution requirements, the streams case in particular is not effectively supported by computing topology and stream distance from a cartographic GIS.

COMPARATIVE EVALUATION

If all generic tasks are supported by the analytic tier, there can be no advantage in supporting decision tasks directly from a cartographic system. Indeed, the chain encoding approach can be viewed as an analytic tier for a particular set of tasks, which emphasizes the point that efficiency is gained by storing directly only the needed relationships at a necessary accuracy. If more elemental relationships are stored than are needed, storage requirements will increase and retrieval efficiency will decrease.

A comparison of tasks shows that for streams, the cartographic approach is inefficient both for storage and retrieval. Finding the characteristics of a place, finding the places with given characteristics, and determining the six relationships listed in Section 2 are more efficient in the analytic structure for both retrieval and storage. From the above comparison we can conclude that the analytic tier should be stored in the computerized database and used to support decision-making. Whether the cartographic tier should also be stored depends on its value in computing analytic relationships, the relative costs and accuracy of an analog vs. computerized cartographic tier, the justification of the initial cost of the cartographic tier for related decision support systems, and the value of the cartographic tier for computerized display.

The value of a cartographic tier in computing relationships is highly dependent on resolution. An evaluation for ISIS of a computerized cartographic tier at 1:250,000 indicated that none of the relationships being determined and recorded at 1:24,000 could be computed satisfactorily from the 1:250,000 data. Most of the relationships could be computed from a computerized cartographic tier at 1:24,000, but, at a cost on the order of two to five times what it cost to use a hard copy cartographic tier. A computerized cartographic tier might be justified if it supported a sufficient number of different analytic *data structures* designed to support efficiently different sets of decision tasks. In other words, if the full capabilities of the cartographic tier were used sufficiently frequently from the elemental level to create many analytic databases, the computerization might be justified.

Computerized display for human interpretation may be supportable directly from an analytic tier. A tree of hydrologic relationships could be constructed with straight lines. Such a display would be quite sufficient to scan for flow relationships. This display could be enhanced through fractalization to improve comprehension (Armstrong and Hopkins, 1983). In general, much lower resolution will be adequate for thematic map display than for computing relationships. Relationships stored in an analytic tier can be correctly displayed, regardless of the resolution or fractal dimension of forms used for thematic display. It may, therefore, be more efficient to create a map display system that is dependent on the analytic tier, than to use a cartographic tier for map display. Data capture, storage and manipulation will be much less costly because the resolution required will be much lower. For many uses of displayed data, video display of a high quality map manuscripts and aerial photographs will be more useful for interpretation.

CONCLUSIONS

Two tiers, an analytic tier and a cartographic tier should be distinguished. The decision to computerize each tier is distinct and must be made using different criteria. If the relationship between the two tiers is carefully structured, it will be possible to modify the two tiers independently. As additional decision support systems based on analytic tiers are created, eventually it may be worthwhile to computerize the underlying cartographic tier. Careful specification of a topological accuracy standard is required to assure valid computation of relationships from digital data. Relationships determined from field data and encoded as relationships can not necessarily be computed from the cartographic tier, even if it is at the same resolution as the hard copy cartographic tier. Data recorded on the original hard copy cartographic tier may have been based on topologic relationships determined in the field. Thorough documentation of data collection and interpretation procedures must be provided to assure valid use of data.

The two-tiered approach will encourage much more rapid adoption of computerized spatial databases to support decision making because it greatly reduces the initial cost of computerization, but still does not preclude evolution toward the full range of GIS capabilities. Very little effort is wasted because the analytic tier, implemented initially on the basis of an existing, hard copy cartographic tier and field data, will continue to be the most efficient storage of data for direct support for decision making. Even if a computerized, cartographic tier is created later, the design and implementation of the analytic tier is therefore justified.

ADKNOWLEDGMENTS

The work reported in this paper is being performed in part under contract with the Illinois Department of Conservation and funded in part by the U.S. Department of the Interior.

REFERENCES

Armstrong, M.P. 1984, An Extended Network Data Structure for Spatial Analysis, paper presented at Annual Meeting, Association of American Geographers, Washington, D.C.

Armstrong, M.P., and L.D. Hopkins. 1983, Fractal Enhancement for Thematic Display of Topologically Stored Data: Proceedings, Sixth International Symposium on Automated Cartography, Volume II, Ottawa, pp. 309-318.

Bonczek, R.H., C.W. Holsapple and A.B. Whinston. 1981, Foundations of Decision Support Systems, Academic Press, NY.

Dueker, K.J. 1979, Land Resource Information Systems: A Review of Fifteen Years Experience: Geo-Processing, Vol. 1, pp. 105-128.

Gardiner, V. 1982, Stream Networks and Digital Cartography: Cartographica, Vol. 19, No. 2, pp. 38-44.

ISIS, 1983. Data Manual, Illinois Streams Information System, Department of Landscape Architecture, University of Illinois, Urbana, Illinois.

ISIS, 1984. Stream Number and River Mile Index Manual, Illinois Streams Information System, Department of Landscape Architecture, University of Illinois, Urbana, Illinois.

Klein, P.M. 1982, Cartographic Databases and River Networks: Cartographica, Vol. 19, No. 2, pp. 45-52.

Peucker, T. and N. Chrisman. 1975, Cartographic Data Structures: The American Cartographer, Vol. 2, No. 1, pp. 55-64.

SIR, 1983. SIR/DBMS Users Manual, SIR, Inc., Evanston, Illinois.