FORMATTING GEOGRAPHIC DATA TO ENHANCE MANIPULABILITY

Gail Langran
Mapping, Charting, and Geodesy Division
Naval Ocean Research and Development Activity
NSTL, Mississippi 39529

Barbara Pfeil Buttenfield
Department of Geography
University of Wisconsin
Madison, Wisconsin 53705

ABSTRACT

Geographic data tends to be exploited extensively and imaginatively once it becomes available. When standard data sets serve as input to applications software, however, the data must often be filtered or restructured. Given this likelihood, special attention should be paid to any distributed data set's manipulability. This paper discusses ways to organize sequential data sets to facilitate three major filtering tasks: windowing, categorical feature selection or aggregation, and resolution reduction Examples are drawn from current format standards.

INTRODUCTION

Users of mapping and geographic information system software can select their systems' input from a small but steadily growing assortment of digital geographic data sets. Since data is a far scarcer commodity than software, a seller's market has resulted. Not surprisingly, data vendors have chosen to distribute their data in standardized forms so resources may be directed to capturing data, rather than diverted to tailoring customized versions of data sets.

Since most applications will manipulate the standardized data format to meet user-specific needs, it follows that manipulability is a desirable data characteristic. The problem, then, is twofold: first, how does one predict which manipulations will be performed on a given data set? And once these are predicted, how can the data set be designed to facilitate the manipulations? The next section illustrates a method of extrapolating potential data set manipulations, followed by a discussion of ways to facilitate filtering of sequentially ordered data using current format standards as examples. The final section summarizes some of the points made here and suggests further work.

FORECASTING DATA MANIPULATIONS

This exercise takes four data set types--shoreline vectors, cartographic features, navigational chart data, and a digital elevation model--and envisions a set of applications for each. Once an application is forecast, its component operations are extrapolated.

World Shoreline Vectors

Shoreline vectors are the most venerable of all cartographic data sets. While commonly used in the past to sketch background maps, their current applications have broadened considerably (Table 1).

Table 1. Applications and (manipulations) of world shoreline vectors.

- Route planning: plot a route by air or sea that avoids a given country or region lying between its two endpoints.

  (create a land mask, apply topological constraints)

- Distance computations: compute the distance from nearest landfall to current position at sea; compute distance from port to current position at sea.

  (compute point-to-point or point-to-line arc distance)

- Merge data: add features; add bathymetric data.

  (transform coordinate system, translate feature codes)

- Repartitioning and windowing: group segments by oceans instead of by continents; extract an area of given dimensions around a given point; extract an area whose corners are given; extract an area that will fit on a given display device at a given resolution.

  (search, extract, clip)

- Restructuring: extract and use spaghetti data only; add adjacency information.

- Scale change: enlarge or reduce.

  (generalize or enhance lines; generalize small island and lake groupings)

- Display: create a map image.

  (label; symbolize; draw outlines only, color fill, merge features from a feature file)

Feature/Attribute Files

For our purposes, features and attributes are defined to include transportation and communication networks, political boundaries, drainage, hydrographic data, vegetation, and other mappable point, line, or areal data in polygon form. Possible uses for such data are listed in Table 2.

Electronic Navigation Charts

A growing family of electronic navigation charts share several properties: many functions occur in real time, some data is received in real time from sensors, and a default mode leaves few cartographic choices to the user. Table 3 extrapolates data manipulations for a shipboard electronic chart. Automotive applications are also possible.

Table 2. Applications and (manipulations) of feature data.

- Spatial comparisons: determine the adjacency, overlap, or distance between features.

  (compute point-to-point or point-to-line distance)

- Feature selection or aggregation: group all drainage features into one feature type rather than discriminating between rivers, streams, and canals; group deciduous, conifer, and mixed forest type into a single forest type; group individual hazards to navigation as "hazards" or different types of obstacles to aviation as "obstacles."

  (search for features; match features to segments; extract)

- Attribute selection, aggregation, or ordinal grouping: group all lighted harbor buoys into one category regardless of light color or strobe frequency; group all vertical obstructions over a given height into the category "hazard to aviation;" rank vertical obstructions into height categories; rank towns by population.

  (search for attributes; delete or assign new codes)

- Repartioning and windowing: see Table 1.

- Restructuring: see Table 1.

- Scale change: see Table 1; also, reclassify area features as point or line features.

- Display: see Table 1.

Table 3. Applications and (manipulations) of electronic navigation charts.

- Real-time computations: use current position, speed, bearing, chart features, and sensor input (e.g., radar, sonar).

- Feature selection and aggregation: see Table 2.

- Attribute selection and aggregation: see Table 2.

- Repartitioning and real-time windowing: see Table 1.

- Restructuring: see Table 1.

- Real-time scale change: see Table 1.

- Display: see Table 1; also, real-time animation and update.

Gridded Data

In the context of data manipulations, elevation, bathymetric, and other types of gridded data differ considerably from the previous three examples. The gridded format persists precisely because it is easily manipulable; far more space-efficient structures have been overshadowed by the programmability of the gridded format. Gridded data sets are therefore included in this discussion (Table 4) for contrast.

Table 4. Applications of gridded elevation data.

- o Windowing: compute location based on pole spacing, extract.
- o Scale change: eliminate elevations (i.e., reduce grid resolution); interpolate new points (raise grid resolution).
- o Display: compute contours, apply hillshading, layer tints, or other graphic effects.

Summary

Despite differences in application and content, a common thread runs throughout the data sets listed above: each may be filtered via windowing, categorical selection, and reduction of scale or resolution which, from this point on, will be referred to simply as reduction. Other possible manipulations include repartitioning and restructuring. Only the filtering process will be examined here; however, the methods and philosophy apply to all forms of data manipulation.

Given the importance of filtering to geographic data sets, the question arises: are current and planned data sets organized in a manner that is maximally conducive to such filtering? The next section uses some of today's standard formats to explore this topic.

MANIPULABILITY OF TODAY'S FORMAT STANDARDS

The discussion that follows suggests ways to facilitate the three major filtering operations and compares how each is addressed by today's standards (Table 5). A fourth performance factor, programmability, is considered last.

Table 5. Acronyms of referenced formats.

Acronym	Full name and (sponsor)
CEDD	Committee on Exchange of Digital Data (International Hydrographic Organization)
DEM	Digital Elevation Matrix (USGS)
DLG-O	Digital Line Graphic - Optional (USGS)
FGEF	Federal Geographic Exchange Format (Federal Interagency Coordinating Committee on Digital Cartography)
GDIL	Geographic Data Interchange Language (Jet Propulsion Laboratory)
GIRAS	Geographic Information Retrieval and Analysis System (USGS)
MCDIF	Map and Chart Data Interchange Format (Ontario Ministry of Natural Resources)
NCDCDS	National Committee on Digital Cartographic Data Standards (sponsored by the same)
SDDEF	Standard Digital Data Exchange Format (NOS)
SLF	Standard Linear Format (DMA)

Before continuing, however, it must be emphasized that some of the formats referenced in this section (CEDD, FGEF, SDDEF, SLF) were designed as vehicles for the exchange of mapping data among or within map production agencies. While these formats were never intended to be manipulable, it is yet instructive to examine them. Other formats (DLG, GIRAS, DEM, WDB II) were designed for more general use. A final class of format standards will not be discussed here. Such formats are essentially virtual envelopes into which data is sealed for dissemination. The envelopes describe the characteristics of the data contained within via coordinate transformation parameters and format statements that facilitate data loading. While extremely useful, the virtual envelopes (GDIL, NCDCDS, and MCDIF) are not relevent to this discussion and will be excluded.

Repartitioning and Windowing

Most commonly, windowing is a straightforward process of subtraction: find and collect only the segments that overlap a given area, then clip from those segments the pieces that are outside the window. Processing is proportional to the number of line segments being searched. Thus, timing problems arise as file size grows. The amount of data packed within a given file unit is of obvious importance in windowing or partitioning efficiency. DLG and GIRAS files reflect this constraint. The 1:100,000 DLG files are subdivided into 15' or 7.5' cells depending on data density. GIRAS files are subdivided into sections not exceeding 32,000 coordinate pairs or 2500 arcs. While the subdivisions were adopted due to memory constraints, their effect is improved windowing performance.

While a slight reduction in file unit size improves the efficiency of subtractive windowing, systematically subdividing the file into small rectangular cells allows users to adopt an additive method of windowing or repartitioning. For optimum results, cell size should match that of the smallest area to be windowed. To window, all cells that comprise the desired window (or partition) are assembled and adjoined, thus avoiding arduous segment searches and clipping. This method brings two space-saving bonusus: if computer memory is constrained data can be loaded in small pieces, and if coordinates are stored relative to cell origins file size is reduced.

To structure a file into cells, segments are clipped and nodes are formed at cell edges. Information concerning cell dimensions is recorded in the volume header, and short cell headers are constructed to provide cell origins (cell coverages are computed using the volume header information). To expedite the search for desired cells, a cell code can be computed and placed in the header to allow spatial hashing based on latitude and longitude (Connor and Langran, 1987). Alternatively, users have the option of aggregating to larger cells or to a quadtree cell representation (Jones and Abraham, 1986).

Categorical Feature Selection

Considerable machination may be needed to extract from a standard file the particular feature and attribute

classes desired for a given application. Conceptually identifying the necessary features can, in itself, be a problem, since three feature coding standards exist in U.S. mapping agencies alone (DMA, 1985; NOS, 1985; and USGS, 1985) and a fourth is being recommended by NCDCDS (1986).

The NCDCDS recommends a hierarchical classification scheme for features and attributes that casts major feature types as nouns that are modified by attribute "adjectives". Both USGS and DMA's coding schemes reflect this sentiment to some degree. The USGS' 3-byte major code is a broad category--e.g., water bodies, political boundaries, rivers and streams--while its 4-byte minor code is descriptive: single-line perennial stream of length 50-60 km, perennial lake or pond, etc. DMA's 5-byte coding scheme describes category and subcategory in the first and second characters, respectively. The broad category represented by the first character (e.g., hydrography) leads to a more specific subcategory (e.g., ports and harbors," "dangers and hazards," "bottom information"). The final three characters are assigned sequentially to features in alphabetical order.

Once the user transcends terminology differences, he must write software to extract from the sequential file the feature subset he needs. The general procedure is:

1. Encounter a feature.
2. Determine the feature's processing needs.
3. If processing is needed, process the feature.

Step 2 stands out as an area where data adaptation could be helpful. Tabular and hierarchical methods of determining processing needs are possible. The tabular method constructs a look-up table containing the codes and processing needs of features to be included. The algorithm is:

1. Search for the feature code in the look-up table.
2. If found, reference and perform the required processing.
3. Get the next feature.

This procedure would be facilitated if feature codes were available in digital look-up tables, which could be edited as necessary by the programmer. Lacking digitized tables, programmers nationwide must do a great deal of duplicative typing.

A cascading method is possible for hierarchical coding schemes. A user may wish to extract from DMA's Hydrography category all port and harbor information, to exclude all bottom type information, and to aggregate all hazards into one "Dangerous" class. The algorithm is:

1. Read the first digit of the feature code.
2. For a hydrographic code, read the second digit.
3. For a port and harbor code, continue reading digits to obtain the rest of the data detail.
   For a hazard code, call the feature "Dangerous" and load it into the data base.
4. Get the next feature.

This method is particularly useful when elimination or regrouping occurs at the categorical level. Without hierarchically assigned feature codes, however, it cannot be used.

Reduction

The previous subsection described categorical, or qualitative, filtering. Reduction implies that features are eliminated based on spatial and quantitative factors: the feature is not important enough to crowd the map at the intended display scale.

Two major operations occur in reduction: points are eliminated from lines and areal boundaries, and features are eliminated based on space available and relative importance. We could find no evidence that any standard format has incorporated ways to facilitate either generalization operation. Since none are in use, this section discusses the feasibility of several data adaptations.

Line segments can be stored hierarchically, although the referencing system would add to data set size. Ideally, hierarchies would be based on geographic features so critical points are preserved in node values. To date, only rudimentary methods of recognizing linear feature types exist (Buttenfield, 1987). Assuming a tolerance-based line generalization strategy similar to the Douglas algorithm (Douglas and Peucker, 1973), tolerance values could be stored in feature look-up tables to avoid the poor results of generalizing all features uniformly (Buttenfield, 1986). Where positional integrity is required, flags could be embedded in segments to denote points that must not be altered due to navigational or other importance.

How to package sequential data to facilitate the second type of operation is problematic. WDB II stores ranks with island and lake groupings, so smaller islands can be deselected as scale is reduced to avoid coalescence. A more flexible alternative might be to store areas or population values with such features so users can determine their own rankings.

Processing efficiency

Processing efficiency can be defined as a rational balance in use of space and time. Programmability, a third factor, is gaining in importance as human resources grow more expensive relative to computer resources.

The physical and logical arrangement of data upon media has a major impact on processing efficiency. A good example can be drawn from logical and physical blocking of tapes. Table 6 shows the impact of block size on a tape's storage capacity. Since blocks must be physically separated on tape by interblock gaps, large blocks with few separa- tions are far more space-efficient. Large blocks are also more time-efficient, since it reduces the number of times the input program must access the tape.

Table 6. The impact of block size on storage capacity. Values are computed for a typical 2400-ft tape using a 0.75-inch interblock gap.

block size	#blocks	Tape capacity at 1600 bpi
8000 bytes	5008	40 MB
5120 bytes	7291	37 MB
1980 bytes	14490	28 MB
1280 bytes	18580	23 MB

The logical organization of records within blocks is a space and programmability issue. Small records are generally used, since these require less padding with spaces and are easily viewed at a terminal. Programmability becomes a further problem when logical records cross the boundaries of physical blocks, as is the case with SLF and CEDD (Table 7).

The use of fixed or free format trades processing speed against flexibility. Since fixed formats are essentially read by template and free formats must be parsed, speed differences can be considerable. Fixed formats include SLF, DLG, SDDEF, CEDD, and WDB II. FGEF is of free format; users define a set of delimiters in the header to separate records, fields, and subfields. An interesting hybrid is proposed by NCDCDS, which would have the computer parse for format statements, which are then used to read N bytes of data in fixed format. GDIL suggests placing these format descriptions in a file header.

Space does not permit a full discussion of these data processing issues. However, further examples can be drawn from coordinate treatment, binary vs. ASCII storage, and media type. Often, the designer must choose between maximizing space, time, or programming efficiency. Since applications users may be constrained in all three areas, the right choice will require a careful deliberation.

Table 7. Size of logical records and physical blocks specified for standard data formats.

Format	Record size	Block size
CEDD	(1)	1980
DEM/DTED	(2)	-
DLG-O	80	multiple of 80
FGEF	80	1280
GIRAS	80	multiple of 80
SDDEF	80	5120
SLF	multiple of 1980	1980
WDB II	80	8000

(1) CEDD specifies four record types: the header (565 bytes), features (188 bytes), segments (42 bytes), and text (1972 bytes).

(2) DEM/DTED files have three record types: the header (864 bytes), the data (144 + (rows * columns * I6)), and data quality statistics (60 bytes).

SUMMARY

A broad range of topics have been discussed. Our original questions concerned how to adapt data so it is more amenable to reformatting by applications software. The paper's method is largely exploratory and expository, since few attempts have yet been made to design manipulability into sequential data sets. Since a number of sequential exchange formats are currently in formative states, however, such ideas could be incorporated with relative ease. If the future of geographic information processing includes data exchange with those outside the mapping profession, a wider range of applications, and a great deal of preprocessing, should be expected and planned for.

REFERENCES

Billingsley, Fred and Strome, W. Murray (1986). "Standardization of Remote Sensing and GIS Data Transfer." Paper presented at the ISPRS Convention, Baltimore, Maryland, May.

Buttenfield, Barbara Pfeil (1987). "Automatic Identification of Cartographic Lines." The American Cartographer (in press). January.

Buttenfield, Barbara Pfeil (1986). "Digital Definitions of Scale-Dependent Line Structure." Proceddings of Auto-Carto London, September. Vol. 1, p. 497-506.

Connor, Maura and Langran, Gail (1987). "Spatial Hashing to Facilite File Windowing." Naval Ocean Research and Development Activity, NSTL, MS. (in press).

Defense Mapping Agency (1985). "Standard Linear Format." Washington D.C.

Defense Mapping Agency (1985). "Feature Attribute Coding Standard." Washington, D.C., July.

Douglas, D. H. and Poiker (formerly Peucker), T. K. (1973). "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature." Canadian Cartographer Vol 10(2), p. 110-122.

IDON Corporation (1986). "MACDIF: Concept Definition." Submitted to Canadian Hydrographic Service, Document number ACN 3-R&D-068-CD, July.

IDON Corporation (1986). "MACDIF: Structure and Coding (draft)." Submitted to Canadian Hydrographic Service, Document number ACN 3-R&D-068-SC, November.

International Hydrographic Organization (1986). "Format for the Exchange of Digital Hydrographic Data." By the Committee on the Exchange of Digital Data, November.

Jet Propulsion Laboratory (1986). "General Data Interchange Language." Pasadena, California, May.

Jones, Christopher B. and Abraham, Ian M. (1986). "Design Considerations for a Scale-Independent Cartographic Database." Proceedings of the Second International Symposium on Spatial Data Handling, Seattle, July.

Langran, Gail; Connor, Maura; and Clark, R. Kent (1986). "Recommendations on DMA's Standard Linear Format." Naval Ocean Research and Development Activity, NSTL, MS, NR 146, July.

National Ocean Survey (1985). "Charted Features Data Base Categories: Feature Category Keys." Rockville, Maryland.

National Ocean Survey (1985). "Standard Digital Data Exchange Format." Rockville, Maryland, March.

U.S. Geological Survey (1983). "Digital Line Graphs from 1:2,000,000-Scale Maps." Reston, Virginia, Circular 895-D.

U.S. Geological Survey (1983). "Digital Elevation Models." Reston, Virginia, Circular 895-B.

U.S. Geological Survey (1983). "Land Use and Land Cover Digital Data." Reston, Virginia, Circular 895-E.

U.S. Geological Survey (1985). "Digital Line Graphs from 1:100,000-Scale Maps." Reston, Virginia, Data Users Guide 2.