GEOMODEL - INTEGRATED DATA STRUCTURES FOR REPRESENTING GEOGRAPHIC ENTITIES

Jay R. Baker
Programming Languages Research Staff
U.S. Bureau of the Census
Washington, D.C. 20233

Introduction

GeoModel is a system of data structures and access software providing a geographic base for a wide variety of planning functions. GeoModel consists of a data storage system with various input, manipulation, and output or display subsystems managed under the umbrella of a control language and special-purpose application packages. This paper will discuss the objectives, philosophical considerations, theoretical underpinnings, and implementation of the system. Not all of the ideas presented here are new, but the unique combination of elements in the design of GeoModel has led to a highly useful system.

Design Philosophy

The design philosophy of GeoModel includes a commitment to draw from mathematical disciplines for suitable models of geographic relationships, to provide a system which does not get in the planner's way, to generate spin-off products throughout the course of the project and to create machine-independent systems, while always striving to avoid complexity in the implementation.

The use of existing mathematical tools is motivated by the consistency and depth which mathematics provides. Ad hoc alternatives to a mathematical design often have some intuitive appeal and may provide expedient implementations, but must eventually lead to patches in the design, further resort to ad hoc approaches, and ultimate lack of generality or integration. By faithfully building on a mathematical base, however, generality and extensibility are assured to the extent to which the mathematical theory has been developed.

A major consideration in designing a system is to decide what it is to be able to do. Systems are often designed by making an inventory of the services and options the user may want, choosing among these as to which are feasible to implement, and then providing the user a menu of commands to the system. To the extent that a system provides off-the-shelf facilities, it also delineates what the user is permitted to do, and channels the planner's intentions into the designer's image of planning. The planning system, in effect, defines what planning is.

Our approach is to not anticipate what the user will want to do, but to provide a means to combine primitive access functions in a great number of ways. Implemented as a language which is extensible and recursive, this facility would permit the planner to build complex access, edit, and display functions. Use may then be made of the system in ways which the designer never anticipated. In short, rather than identifying high-level services to provide, we have tried to identify low-level barriers to access which can be removed.

An important goal of GeoModel is to insulate the planner, to whatever extent desired, from census-defined geography and to operate exclusively in the units which the planner finds most useful. Data access is dictionary-based so that the user need never become familiar with actual record layouts. A planner may define traffic zones, school districts, or police precincts as sets of tracts, zips, or any other geographic entities already defined in GeoModel. Once the planner has described the relationship of his real-world planning zones to the census geography, he will immediately have access to aggregated population, race, income, miles of bus-lines, or any user-defined attributes in terms of the planning zones.

Mathematical Foundations

Fundamental mathematical contributions to GeoModel are found in topology and graph theory, as developed by Corbett (1) and White (2). Topology tells us that a line is bounded by exactly two nodes and is cobounded by exactly two areas. These are abstract relationships for which a great body of mathematical analysis has been developed. In spite of some obvious real-world entities which are related in this manner, the topological relationships are not a description of any particular physical structure, but rather serve to describe inherent properties of any system which can be modelled in terms of nodes, lines, and areas.

DIME is a well known topological model which encodes city streets as lines between street intersections (the nodes), with blocks represented as the cobounding areas. Within the topological constraint of planarity, which is satisfied in a general way by local patches on the earth, it is clear that a street segment always has exactly two sides and extends between exactly two points. (Of course, the Capitol Beltway is nowhere simply a line with two sides, and streets do not intersect at a point, but the model which makes this assertion is useful and powerful.) DIME thus represents a physical situation which is ideally suited for a topological model.

Any computer-based implementation of the DIME model must support algorithms which operate on the topological and graph theoretic relationships. At some basic level, the following four operations must exist:

1. Given a line, compute its bounding nodes;
2. Given a node, compute its cobounding lines;
3. Given a line, compute its cobounding areas;
4. Given a area, compute its bounding lines.

Planning zones are comprised of sets of atomic areas. Lattice theory is well suited to representing the overlapping and nonhierarchical nature of various administrative areas by their set relationships. For instance, the boundaries of ZIP codes and tracts seldom correspond. A lattice relates any two such zones covering some common area on the basis of set inclusion and partial ordering rather than any predefined precedence between all tracts and zips.

Data Structures

A GBF/DIME file from the Census Bureau carries all the necessary information to support the four topological operations listed above. However, its sequential structure, oriented around the line or street segment, causes the second and fourth to cost several orders of magnitude more than the other two. Each GBF/DIME record carries the complete set of 0-, 1-, and 2-dimensional information associated with a street segment, making any segment-based access very straightforward, indeed. The drawbacks of this layout, however, include the lack of control between records which carry the same information in several places, such as street names and coordinates, and the difficulty of updating such information. Edit operations using such a structure typically involve reformatting the file into a more suitable form, performing the operation, and restoring the data to its original format.

GeoModel stores the GBF/DIME information in a direct-access form which makes the four operations listed above all cost the same. This requires three basic record types, called 0-cells, 1-cells, and 2-cells, containing 0-dimensional, 1-dimensional, and 2-dimensional data, respectively.

The 1-cells are the keystone of the topological data structures. Each 1-cell record corresponds to a street segment record in the original GBF/DIME file and relates the segment to its bounding nodes and cobounding areas (henceforth to be called 0-cells and 2-cells). They are split into two parts, a topological relations part, and a features part.

Figure 1a. 1-cell Record

from 0-cell	to 0-cell	left 2-cell	right 2-cell

Figure 1b. 1-cell Feature Record

phrase pointer	non-street code	GBF/ DIME id	left addresses	right addresses

GeoModel reduces the costs of computing 1-cells (the second and fourth operations mentioned above) by copying the relation information contained in the 1-cell records and grouping it by 0-cell and 2-cell. That is, a 0-cell record references all 1-cells which cobound it, and a 2-cell record references all 1-cells which bound it.

Any other information about a particular 0-cell or 2-cell on the GBF/DIME is stored only on the 0-cell or 2-cell record. Therefore, the latitude/longitude coordinates and census node number for a 0-cell are recorded only once, on a 0-cell record, rather than on each 1-cell which refers to it. So when a coordinate changes, it is updated in only one spot in the data base, with all references to that 0-cell implicitly corrected at no cost. Likewise, 2-cell records contain any 2-dimensional codes from the GBF/DIME such as state, county, tract, and block.

Figure 2a. 0-cell Record

GBF/ DIME id	lat/ long	count of 1-cells	list of 1-cells

Figure 2b. 2-cell record

pointer to CITY file	tract	block	ED	ZIP	count of 1-cells	list of 1-cells

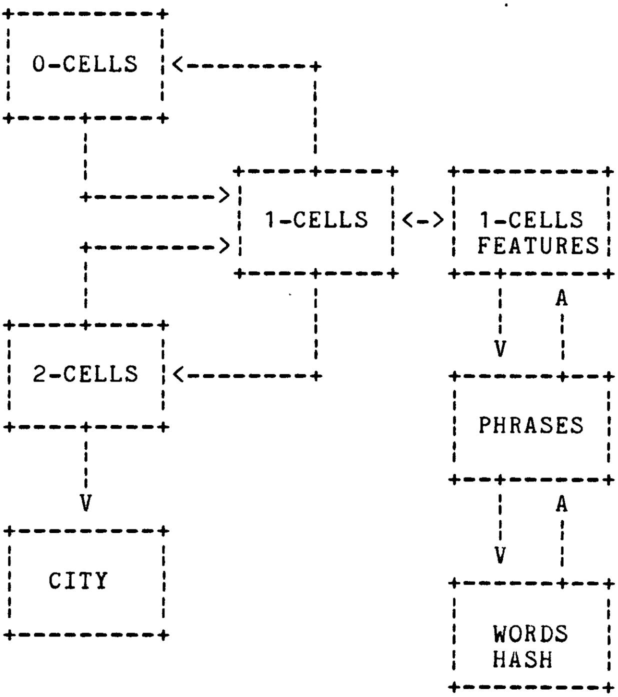
The CITY file is an interim form of the 2-dimensional set descriptions which will later be converted to a lattice.

Figure 2c. CITY Record

state	place	county	SMSA	MCD	CD	area	postal state	city name

After GeoModel has been initialized, the files have the following relationships:

Figure 3.



Implementation

The development plan calls for a series of products to be released throughout the course of the project. While only about one-third of the entire system has been implemented to date, there is already a stable and useful data base with several spin-off application packages and independent software tools. These include a data retrieval and manipulation language called L; the Basic Access Method (BAM); GMLOAD, a series of programs to restructure a GBF/DIME file into a set of GeoModel files; WORDED, a program to search and edit street names; and ADMATCH'80, an interactive address matching system. Two of these, BAM and L, are crucial to the implementation of GeoModel.

Early in the design of GeoModel it became clear that many files were going to be needed, and that the ultimate number of files was unknown. In the interest of portability (and hopefully, ease of use), the interface of GeoModel with the operating system had to be as simple as possible. This led to the design of BAM, the Basic Access Method, which stores all files used by GeoModel in a single direct-access file (called the "library") partitioned into any number of sub-files (called simply "files"). Records are then referenced by file name and record number, and may be read or written in any order. Records within a file must all be of the same length, but each file may have a different record length. BAM has the added benefits of centralized I/O handling, a reduced number of file handling tasks such as opens and closes, and a data access technique which is invariant throughout the system.

A language called L is being developed to access GeoModel. It is composed entirely of functions which may be built-in primitives or user-defined combinations of other functions. Any operand of a function may itself be another function. This allows pipelining of the results of one function into another. Every data item in GeoModel may be accessed by name and record key through a primitive called GM. Other primitives provide IF/ELSE structuring, iteration, and address matching.

New applications may be implemented by the planner at the terminal. As an example, suppose it were necessary to find the nearest cross street to a particular house address. The address must be matched to a particular 1-cell to make a 0-cell available. From the 0-cell, the intersecting streets are available, so via 1-cell and phrase (street name) pointers, the cross street may be found. Two lines of L can do this as follows:

```
PRINT GM 'PHRASE' GM 'C1P' GM 'C01' GM 'FROMO'
    HNO1 '202' PMT 'HAWAII AV NE'
```

Executing from right to left, PMT returns a phrase record key, HNO1 returns 1-cell record key, GM returns first a 0-cell, then a 1-cell, then a phrase record key, and finally the phrase itself. PRINT displays the street name on the terminal. While L requires a certain amount of training to use, it allows the composition of functions by the planner which the designer never could have foreseen. Some special-purpose applications of GeoModel have session control commands which are more user oriented, but less versatile than L.

GeoModel is being developed as part of an interagency agreement between the Census Bureau and the Urban Mass Transportation Administration (UMTA) to enhance the planner's interface to census data. It is implemented in portable COBOL, and both the software and the data files have been designed to allow transporting them from one computer to another. All functions supported by GeoModel can be executed in either interactive or batch mode. GeoModel is to be distributed by UMTA as part of the Urban Transportation Planning System (UTPS) to 300 planning agencies nationwide.

References

1. Corbett, James P., "Topological Principles in Cartography," to be published as a Census Bureau Technical Paper.

2. White, Marvin S., "The Cost of Topological File Access", *Harvard Papers on Geographic Information Systems*, 1978