Title
THE CARTOGRAPHIC DATA BASE IN SEEDIS

Author
Holmes, Harvard H.

Publication Date
1975-05-01

Presented at Conference on Computer
Suppport of Environmental Science
and Analysis, Albuquerque, NM,
July 9-11, 1975

LBL-3848
c.1

THE CARTOGRAPHIC DATA BASE IN SEEDIS

Harvard H. Holmes and Donald M. Austin

May 1975

Prepared for the U. S. Energy Research and
Development Adminstration under Contract W-7405-ENG-48

LBL-3848
c.1

LBL-3848

THE CARTOGRAPHIC DATA BASE IN SEEDIS*

by

Harvard H. Holmes and Donald M. Austin

Mathematics and Computing Group
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

May 1975

*Work done under the auspices of the Energy Research and Development Administration.

ABSTRACT

A geographic information system is comprised of many parts: data base management systems, inquiry and modeling systems, and so forth. A cartographic data base (CDB) is an essential component of such a system. Such a data base provides the correspondence between geocoded data items and their geographic location. It provides the essential element in neighborhood analysis, that is, the analysis of effects which depend on adjacency or other geographical configuration. A CDB must be supported by subsystems for data base acquisition and editing, data base manipulation, display, and the use of the data base by analysis programs.

INTRODUCTION

A geographic data base (GDB) is any set of data containing a reference to a spatial region (i.e., points, curves, polygons, etc.). For example, census, air pollution, weather, land use, etc., data bases have an essential geographic content. Examples of data bases without essential geographic content are high-energy particle scattering cross-sections, bibliographic files and material properties. A cartographic data base (CDB) is a set of data defining geographic entities in some spatial coordinate system without regard to any other characteristics. These entities include points, cells, curves, polygons, polyhedra (essentially one, two and three dimensional spatial constructs). The concept can easily be extended to a fourth dimension (time) when appropriate. At Lawrence Berkeley Laboratory, this cartographic data base supports a Socio-Economic-Environmental-Demographic Information System (SEEDIS).

The primary purpose of a CDB is to allow analysis of geographic data bases. Included in the term "analysis" are the following functions:

1. Display
2. Integration (aggregation, disaggregation, combination)
3. Calculation of spatial variables (distance, area, etc.)
4. Modeling

The display function encompasses everything from straightforward thematic (or choropleth) mapping of data on a recognizable geographic basis to more exotic techniques such as boundary distortion, symbolic overlays and perspective rendering of 3-D representations.

The integration function allows the aggregation of data to larger geographic entities (e.g., states from counties), the disaggregation to smaller entities and the combination of data sets referred to different geographic entities (e.g., census tracts and transportation zones). Typically, most data sets are referred only to geographic codes instead of spatial coordinates. The CDB must contain the correlation between these codes and the geographic boundaries represented.

Calculation of spatial variables, such as distance and area, is possible with the CDB. Applications such as population density calculations, transportation network structures and land use studies depend on the CDB.

The modeling category is intended to describe the variety of applications possible with the combined sets of characteristic data and CDB. An example is power plant siting, which requires socio-economic and demographic, land use, topographic, and political boundary data. Such studies may also call for data on weather, water resources, and a variety of other variables, all of which must be referred to spatial coordinates.

In this context, an information system supporting the four functional operations on data bases with geographic content can be specified. In the following sections some of the components of such an information system will be delineated and descriptions of the implementation of these components at LBL will be given.

COMPONENTS OF A GEOGRAPHIC INFORMATION SYSTEM

In order to limit the realm of discussion to a practical domain, a Geographic Information System (GIS) will be restricted to mean a logical system containing the following components:

a. A set of data bases referred to spatial (and, where applicable, temporal) units. This set includes socio-economic, environmental, demographic and cartographic data bases.

b. A set of data base management programs for storing and retrieving these data bases.

c. A set of programs for performing basic operations on the data.

d. A set of programs for input and output, including graphic displays.

This is considered a logical system in the sense that different programs may be required for storing, retrieving, manipulating and displaying certain sets of data, but these functions must be present in order to comprise a Geographic Information System. One usually thinks in terms of program modules, each of which performs a specific task and for which an interface exists to the other modules, either in a hierarchical structure or in a network structure.

A variety of data structures may be required for such a system, including hierarchical structures (e.g., Region, State, SMSA, County, Census Tract), network structures (e.g., nearest-neighbor zones, next-nearest-neighbor zones, etc.) or relational structures (e.g., Area A $\supset$ (area B $\cup$ area C)). Rarely can one general data base management system efficiently handle large data bases containing all types of structures. Common implementations provide interfaces to specialized modules designed to handle subsets of the data efficiently.

The Socio-Economic-Environmental-Demographic Information System under development at LBL contains a variety of such modules specialized to access several sets of data bases in the most efficient manner. The Storage-File (STOFI) system handles large, hierarchical, variable length data bases in a manner optimized for the BKY operating system. The SIRAP system, based on QWIK QWERY, handles fixed-length data bases of medium size intended for interface to statistical packages. The INGRES system, developed at UCB, handles relational data bases for high-level language queries based on complex relations between data elements. The MAPEDIT system performs selective retrieval (and many other geographical manipulations) on cartographic data bases. There are several specialized interactive retrieval and reporting packages which serve particular needs.

Display modules include a wide variety of programs for representing the results of data retrieval and analysis in meaningful forms. Report generators, graphic table generators, mapping programs and 3-D surface display programs are examples of these modules.

The interface module which will tie all these specialized modules together is based on the Berkeley Data Base Management System (BKYDBMS). This is a relatively simple, hierarchical system designed to be machine-independent, that is, machine dependent routines for disk access, character packing and unpacking and terminal handling are isolated into easily replaced subroutines which interface to the rest of the system in standard ANSI FORTRAN conventions.

Figure 1 is a diagram of the way in which these modules interact with each other and with the user and LBL's archival data base.

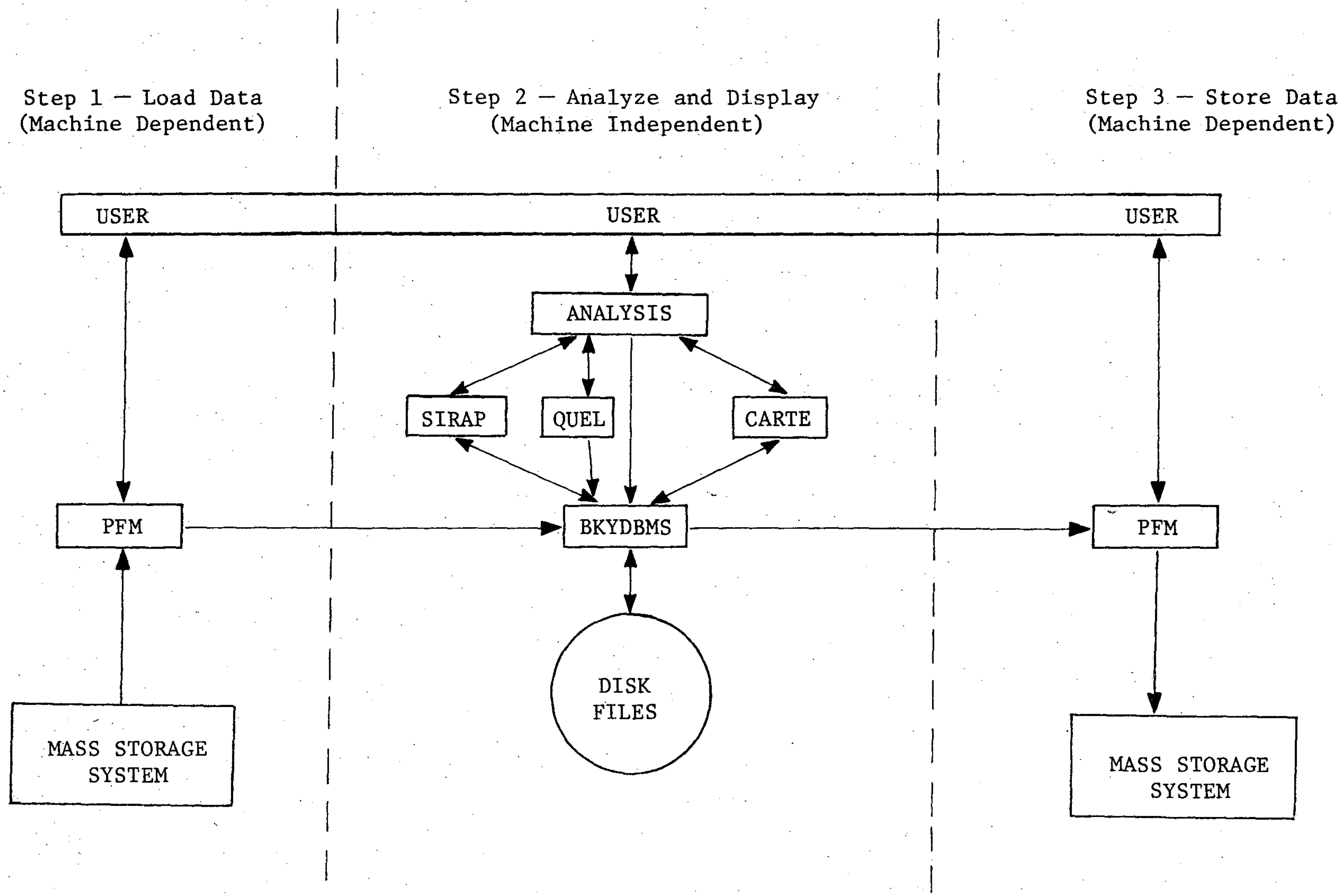

Figure 1. SEEDIS Block Diagram

CARTOGRAPHIC DATA BASE ACQUISITION AND EDITING

The MAPEDIT subsystem is a comprehensive system for digitizing and manipulating maps. It may be helpful to describe the largest project undertaken with the MAPEDIT system to date, the digitizing of census tract boundaries in the 241 SMSAs (as of 1970). The maps were obtained from the Bureau of the Census, where mylar overlays of the desired boundaries were photo reduced to 105mm film to fit the scanning area of the automatic digitizing equipment. The availability of the mylar overlays provided clean, uncluttered maps with only the boundaries needed. In addition, fiducials were put on the maps and a magnetic tape was made containing the location of each tract with respect to these fiducials. The tract locations were taken from existing MEDLIST tapes. The film and tapes were sent to a subcontractor, i/o Metrics Corporation, where they were digitized by a Laser-Scan digitizer. This digitizer uses a laser beam and a photomultiplier to scan the film image of the map. The laser beam is displaced by a spinning prism into a small circular orbit. The center of the orbit is set by deflection mirrors under the control of a PDP-15. As the photomultiplier detects changes in intensity, the angular position of the beam in its orbit is read to deduce the exact location of a line being crossed. The laser beam then follows the line under control of the PDP-15. This PDP-15 uses the tape of tract locations to position the beam inside a tract. The beam then scans sideways until a boundary is found; it then follows the boundary until it closes on itself. The digitized coordinates are returned to LBL for processing.

Many steps are required to produce clean, error free maps. First the data is cleaned up and compressed by searching for straight lines and removing extra points; at the same time, corners are squared. The maps are now checked for accurate fiducials. After converting all the pieces of a map to longitude and latitude, according to the fiducials, the pieces should fit together when plotted. Most of the time, minor corrections must be made to bring the pieces into alignment. This operation is carried out interactively using a storage scope to display the map. When this operation is complete an automatic editing program compares each boundary with the boundaries of its neighbors and makes minor adjustments to match the boundaries exactly.

The tracts are now edited by hand with an interactive edit program. Each tract is examined and compared with a paper copy of the map. Any part of a tract can be changed with the editor, including adding or deleting tracts. A data tablet can be used to add tracts. When the editing is complete, the tracts are compared with a master list of tracts for completeness and to fill in geocodes. The MAPEDIT system provides modules for converting the map into a variety of formats. Facilities are also provided for scaling and rotating maps and for conic transformations. These facilities allow the cartographer to position the map areas to achieve a pleasing composition. For example, Alaska and Hawaii may be moved close to the continental U.S., and reduced in size to achieve a pleasing map for display purposes.

CARTOGRAPHIC DATA BASE MANIPULATION

To manipulate a cartographic data base, a uniform and convenient coordinate system must be used. Geographic location can be described in a wide variety of ways. People find it convenient to use terms of everyday reference such as "Berkeley", "next to the bank", or "1324 East Seventh". These systems reference locations in terms which are of interest to people. A greater degree of uniformity has been provided by the use of numeric codes to represent location, for example, Zip Codes, trucking industry codes, and FIPS codes. These codes provide a uniformity of format, but are still rather arbitrary with respect to the geographic locations to which they refer. There is no direct way to know which Zip Codes are next to code 94720 for example, or the distance from Berkeley to Albuquerque.

For computers, a more useful system is provided by gridded or coordinate systems. In this case, a two or three dimensional system of coordinates provides a consistent system for geographic reference. The two most common coordinate systems are the Universal Transverse Mercator (UTM) and the geodetic (longitude-latitude). The UTM system attempts to represent equal distances on the ground by equal distances in the reference coordinate system. Since the earth is not flat, this is not possible and discontinuities are introduced at the edges of zones. These discontinuities require more work in computer codes and so the preferred system for computer cartographic data bases is the geodetic system. While the geodetic system is distorted with respect to earth distances, nevertheless it is in common use and the distortions are regular and uniform and can be easily compensated for by the computer.

One of the functions of a CDB is to provide a correlation between the many types of geocodes in use. This service is performed using two techniques: first, geocodes are converted to coordinate locations, second, a structure is created which allows rapid searches to find these correlations. Using these techniques we can answer the question, "Where is Berkeley", or "Is Zip Code 94720 in Berkeley?" Another function is to provide information about adjacency or other geographic configurations. For example, "What Zip Codes are next to 94720?" or "Does a certain set of Zip Code areas have any donut holes?"

A variety of cartographic data structures have been developed to serve these needs as well as to allow the simple display of geographic boundaries. The simplest approach is the grid or cell system. In this system, each point, line, or area is associated with one or more cells. Since the cells form a regular array, it is easy to find adjacent cells or to analyze the configuration of a group of cells. Cells provide poor resolution, however, since they are fixed in size over a given area.

Better resolution of geographic boundaries is achieved by directly measuring and storing the boundaries of geographic entities. The World Data Bank is an example of this type. These boundaries are essentially unstructured. More structure is achieved by collecting together all of the lines for a given boundary (polygon structure). This is one of the standard structures handled by the MAPEDIT subsystem of SEEDIS. Even with this level of structure, however, it is difficult to locate adjacencies and other geographic configurations. A variation of this structure is the point list structure in which the actual coordinate values are stored separately; each polygon is then a list of references to these points. This facilitates editing of the map and the detection of adjacent boundaries since the actual coordinates of a point are stored only once, even though it may belong to two or more boundaries.

Chain files code geographic boundaries as lines connecting nodes with codes for left and right entities associated with each line. Geographic subregions can be easily combined using this data structure. For example, a chain file of county boundaries can be reduced to a file of state boundaries by deleting all line segments which have the same state on both the left and the right. The Metropolitan DIME files created by the Bureau of the Census include street addresses for each segment. These files provide a data base to enable conversion from street addresses to geodetic coordinates. A number of improvements have been made to simple chain files [3,4].

A number of manipulations must be provided to use with the CDB. Coordinate transformations are among the simplest. Subroutines convert the geodetic coordinate to and from other systems such as Mercator, UTM, State Plane, and a variety of others. MAPEDIT also provides a utility transformation for making small adjustments to portions of the CDB. A selection facility must also be provided:

a. By geocode sets (Boolean)

   e.g., Get all tracts in County 1 not in Place 100.

   Retrieve Entity = TR where County = 1 and Place ≠ 100.

b. By geographic region

   e.g., Get all counties in box (X_1Y_2), (X_1,Y_2)

   Retrieve Entity = County where

   ANY POINT ON BOUNCARY $\subset$ Box((X_1,Y_2),(X_1,Y_2))

c. By calculated attribute

   e.g., Get all Counties where area > 10 square miles

   or Get all tracts in a radius of 10 miles from Berkeley.

The problem of overlapping geographic regions arises when data has been collected according to two or more different geographic partitionings. For example, data collected according to census tracts, and data collected according to hospital zones cannot be directly compared because the units of aggregation are different. A common solution to this problem has been to assign the data to an artificial grid of cells covering the area. This allows the comparison of data to take place at a common unit of aggregation, the cell. While this approach is simple and usually passably efficient, adequate resolution may require a very large number of cells. An alternative approach to this problem has been provided in SEEDIS. In SEEDIS, the geographic subsystem provides the capability to extract the least common geographic units (LCGU) from several sets of geographic regions. The LCGUs are the geographic regions created by considering all the boundaries in the geographic regions simultaneously. Thus, to pursue the example of census tracts and hospital zones, a map of LCGUs is created in which each unit is entirely inside a census tract and a hospital zone. Both census tracts and hospital zones are now composed of one or more LCGUs. The data analysis then proceeds at the level of LCGUs where the level of aggregation is comparable for the the data from the two sources. When the analysis is complete, the LCGUs may be combined where the data is identical. The data is then displayed at this level. If the data is mapped, any convenient set of reference boundaries may be shown on the map for user orientation. It is not necessary to use the coordinates associated with the LCGU although that is a common case.

Some information, however, is naturally available as cellular data, that is, based on a regular subdivision of the earth into rectangular pieces, e.g., ERTS data. If a cellular structure is not desired, then the data can be contoured and converted to a polygon representation. Both techniques have their place, and both will be available.

The INTERSECT module, part of the MAPEDIT [7] system, is responsible for intersecting two or more maps and extracting the LCGUs This is accomplished using a data structure and algorithm developed by Nicholas Chrisman [3,4]. The results of the module provide the foundation on which data is associated with its geographic extent. Thus, the unit of analysis is the LCGU, shown schematically below:

LCGU
IDENTIFIERS
COORDINATES
DATA

Figure 2. Geographic Information

There are a number of refinements which are made to this basic structure to make it more useful for analysis and modeling routines. A list of adjacent LCGUs is provided to allow modeling systems to easily incorporate the effects of adjacent areas. Provision is made for degenerate LCGUs to allow the convenient incorporation of line (transportation, power, etc.) and point (power plants, schools, etc.) data. For speed of access, the actual coordinates of each LCGU are stored separately from the data, since most analysis programs do not use the explicit coordinates of a geographic region. In order to conserve storage space, data may be stored at a level of aggregation which applies to several LCGUs. This would commonly be done as the data base was loaded from each of the data sources with different geographic subdivisions. As the analysis proceeds, new values would be stored in more detail, associated with each individual LCGU. A more realistic example is now presented, based on our earlier example of census tracts and hospital zones.

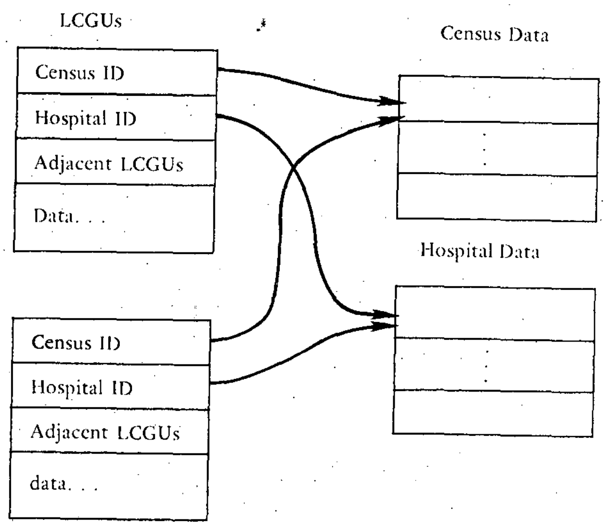

Figure 3. Geographic Data Structure

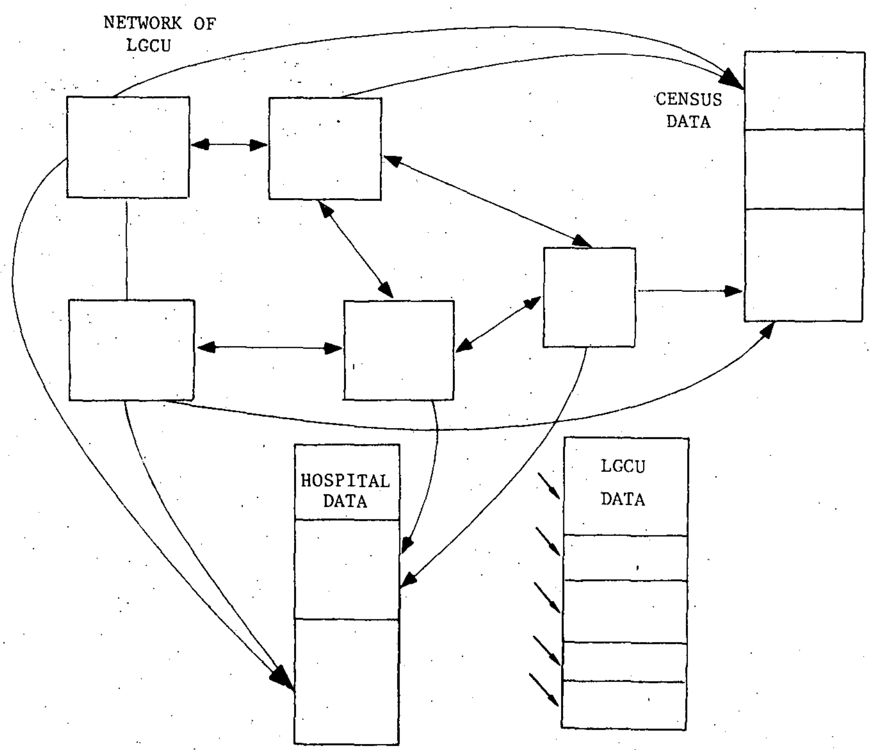

Figure 4. LGCU Data Structure

As illustrated in the example, two or more LCGUs can belong to a single geographic area of one of the maps. In this case, two LCGUs belong to the same census tract, although they belong to different hospital zones. Thus they can share data which applies to census tracts. This situation can be recognized when the data base is being created. As the analysis proceeds, new data is created. These values are stored with the individual LCGUs since they probably reflect both census tract and hospital zone data, and thus cannot be aggregated beyond the LCGUs.

This same data structure can be used with a grid of cells, of course. The only change is the use of a simpler INTERSECT routine and the replacement of LCGUs by cells.

DISPLAY OF GEOGRAPHIC DATA

SEEDIS has two subsystems for graphic output, a Graphic Display Program (GDP) and a mapping module CARTE. The Graphic Display Program provides a tool for visual analysis of tabular data. To aid in this objective, the data can be grouped, scaled, sorted or transformed by simple mathematical operations. A variety of display forms are provided such as line charts, bar charts, pie charts, etc., to allow several kinds of emphasis to be applied to the data. By a variety of transformations, patterns in the data can be made evident to visual inspection. Emphasis can be applied to negative values, positive values, or values in a particular interval. Hard copy is provided so that these analyses can be easily incorporated into reports.

CARTE [13], the mapping module, provides a variety of options. CARTE can produce both symbol and choropleth maps from point and polygon data. It is designed to produce publication quality maps at low cost. The usual output is color separation negatives, but cross-hatch maps can be produced for a quick-look viewing of the map. CARTE allows the user to compose the map to achieve a pleasing layout of titles, legends, etc. Extra boundaries can be included for emphasis or to help orient the viewer. Congested areas can be enlarged and replotted as extra maps to provide more detail. A set of default options is provided to give a convenient starting point for new users.

In addition to these two systems, there are a variety of routines for contour plots, scatter plots, 3-D representations and the like.

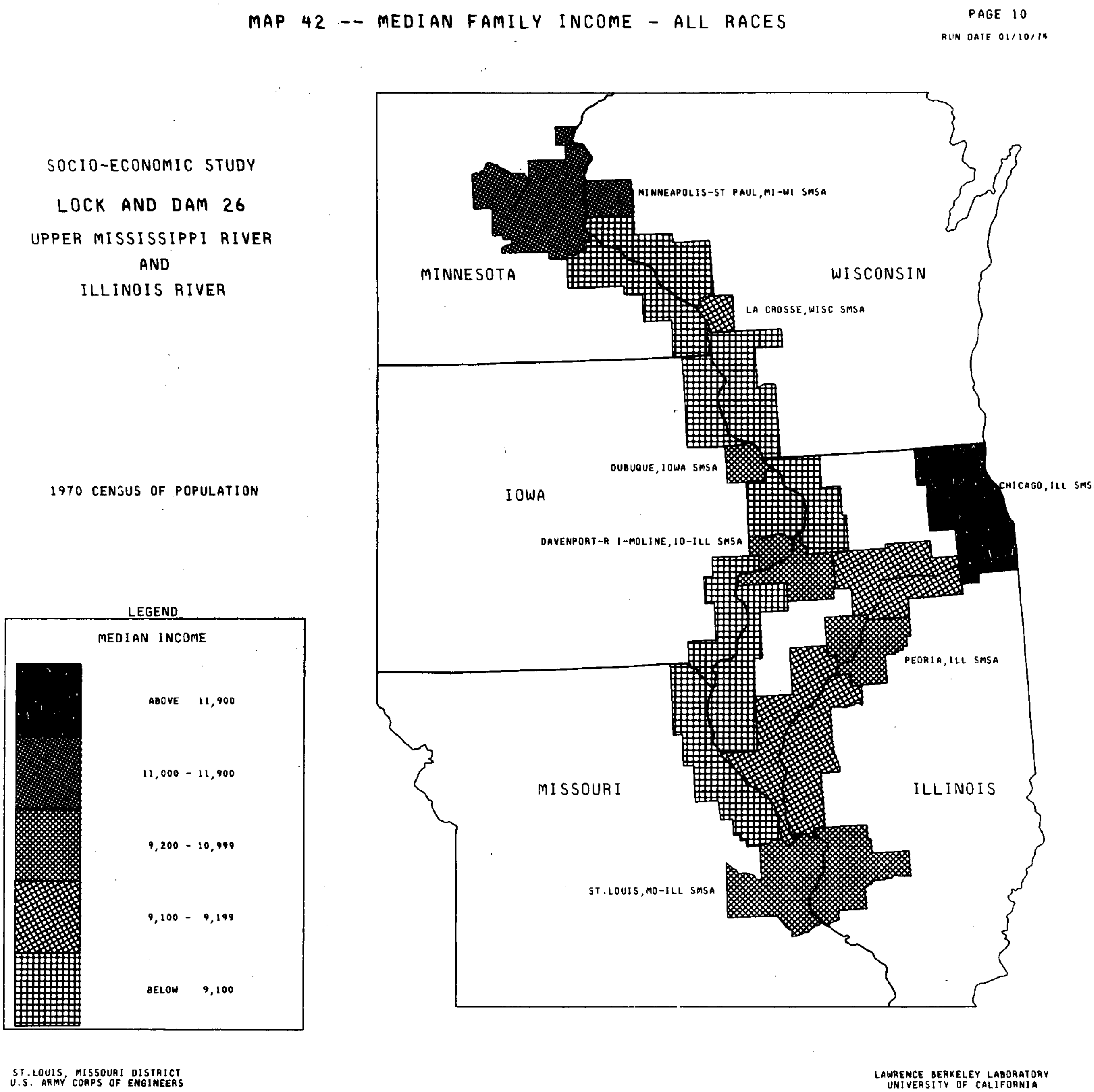

Fig. 5. A CARTE Map

INTERFACING TO ANALYSIS AND MODELING PROGRAMS

The data base management system must provide an interface between a using module and large amounts of information stored on disk or other mass storage device. The using module must be able to create structural relationships between data items and to use this structure to quickly and efficiently retrieve stored information based on user requests or on relationships within the data.

To perform this service, the data base management system must manage the physical disk space, requesting more space or releasing unused space, as necessary or desirable; it must accept new items for storage or update; and it must find and read information from the data base. In order to utilize the structural relationships among data items, the data base management system must provide selective access to the data.

A using module can request information according to a keyword related to some attribute of the information itself. Using such a facility, a module can build up and store relationships between data items. The data base management systems used by SEEDIS already provide a variety of schemes for relating data items to one another. These built-in organizational facilities allow items to be accessed sequentially, or in a predefined order as well as randomly under control of the requesting module. Further extensions allow searching for items on the basis of the nearest match with a keyword rather than an exact match. These features are generally supported by creating one or more indexes to the file based on one or more of the fields in each record. The user can specify which fields should be indexed.

In addition to the features described above, which are part of the "standard" data base management systems at LBL, the SEEDIS versions have been extended to provide some special capabilities which are commonly required in geographic modeling and analysis systems. These are (1) the efficient storage of time series data and (2) the efficient storage of geographic adjacencies. The efficient storage of time series data is accomplished by only storing <u>changes</u> in time series data. For example, if data were stored for 1964, 1968, 1972, 1976 and 1980, only <u>new</u> values would be stored for the years after 1964. Missing values would be automatically supplied by referring to earlier years. The storage of geographic boundaries and adjacencies is described in the section on geographic data structures.

The data management system also provides the capability to reorganize information to provide faster or more economical access to the data. The most common example would be the reorganization of a matrix or table so that elements were stored by row instead of column (or vice versa). This is done by adding another index to the file without rewriting the file itself.

Based on these capabilities the data management system serves as a repository of information which is used by the various modules in SEEDIS. The data management system includes modules to transfer information between the data base and the file formats used by various other programs. Many of these conversion modules are consolidated in the data acquisition modules.

The Berkeley Data Base Management System (BKYDBMS) is a machine independent module for storing and retrieving data from disk. It is used by other SEEDIS modules as a collection of subroutines. BKYDBMS stores and retrieves records consisting of data elements. These elements may be character strings, integer or floating point numbers, or nodes. Any element may be defined as keyed. The data is keyed in ascending sequence on each key. This automatic ordering requires that a routine to reorganize the key file must be run at intervals.

SEEDIS contains a variety of inquiry languages. Some of these inquiry languages (e.g., BKYDBMS) interface with the standard data base management subsystems and interface with the rest of SEEDIS via the file conversion modules. These languages are sufficient to do ordinary retrieval and update and also a few simple calculations such as sums, percents and some standard statistical analyses.

However, it is anticipated that the majority of simulation and modeling systems using SEEDIS will be constructed to meet the special requirements of specific projects. SEEDIS is designed to provide an environment in which these special systems can be constructed with a minimum of difficulty and with a minimum of duplication of effort. As already described, SEEDIS provides a standard data base management system, a standard data structure for geographic data and a variety of existing data bases. In addition to these, as part of the language support, SEEDIS also provides syntax analyzers and other tools which are of use in constructing specialized analysis languages.

Several languages are currently available. QWIK QWERY [9] is a proprietary language which provides both a self-contained data management system and a self-contained report generator. It interfaces with other system components by passing files to them. QWIK QWERY can read most of the archival files via the archival retrieval modules and can supply output files for many of the report generators.

SIRAP [10] (System of Information Retrieval and Analysis for Planners) is a system including a comprehensive set of statistical analysis programs.

CUPID [11] is an interactive graphics language for the INGRES [12] system. The user is provided with a dictionary of the data base and a menu of query components, e.g., symbols representing retrievals, arithmetic operations and conditional specifications. This language is translated into QUEL, a lower level query language, and thus needs no interfaces with data management or report generator routines.

QUEL is the non-graphic inquiry language for INGRES. The QUEL processor analyzes the query and optimizes the sequence of retrieval steps. It interfaces with a data management system with the following level of capabilities: (1) get the next item, (2) get the item matching a certain key, (3) delete last item, (4) index by a given field, and so forth. The interface with the report generator is by file transfer. INGRES is an experimental system under evaluation for efficiency and user convenience.

CONCLUSION

The combination of a CDB together with a suitable geographic data structure provides a powerful base for creating geographic information systems. By constructing such a system in modular fashion, an excellent starting point is provided for specialized, user-written analysis packages. SEEDIS is an information system designed to provide such an environment.

REFERENCES

1. Donald M. Austin, Suzanne G. Kranz, and Carl Quong. "An Overview of the LBL Socio-Economic-Environmental-Demographic Information System (SEEDIS)", LBL-3699, Lawrence Berkeley Laboratory, March 1975.
2. Carl Quong, "Using SEEDIS", LBL Report, Lawrence Berkeley Laboratory (in preparation).
3. Thomas K. Peucker, and Nicholas Chrisman. "Cartographic Data Structure", ACSM, St. Louis, 1974.
4. Nicholas Chrisman. "The Impact of Data Structure on Geographic Information Processing", ACSM and U.S. Geological Survey, International Conference on Automation in Cartography, Reston, Virginia, December 10, 1974.
5. Dave Richards, and Gill Ringland. "Data Base Management Systems (DBMS) Manual", UCID 3728, Lawrence Berkeley Laboratory, April 1975.
6. Marilyn Mantei and Judith Sarna, "An Introductory Manual to STOFI A Data Management System", LBL 2649, Lawrence Berkeley Laboratory, January 1974.
7. H. H. Holmes, D. M. Austin, and W. H. Benson. "The MAPEDIT System for Automatic Map Digitization", LBL-3072, ACM SIGGRAPH First Annual Conference on Computer Graphics and Interactive Techniques, Boulder, Colorado, July 15-17, 1974.
8. Fred Gey, and Edna Williams, "REAP, The Teletype Retrieval Program for the Socio-Economic-Environmental-Demographic Information System", LBL Report, Lawrence Berkeley Laboratory, April 1975 (in progress).
9. QWIK QWERY Users Manual, Consolidated Analysis Center, Inc., Santa Monica, California, April 6, 1970.
10. Fredric Gey, "SIRAP Information Storage and Retrieval Facilities", Lawrence Berkeley Laboratory, March 1974.
11. Nancy McDonald and Michael Stonebraker, "CUPID - The Friendly Query Language", Mem. #ERL-M487, ERL, University of California, Berkeley, California, October 1974.
12. M. R. Stonebraker, and E. Wong, "INGRES - a Relational Date Base System", Mem. #ERL-M477, ERL, University of California, Berkeley, California, November 1974.
13. Peter Wood and Donald M. Austin, "CARTE: A Thematic Mapping Program", LBL-3073, ACM SIGGRAPH First Annual Conference on Computer Graphics and Interactive

Techniques, Boulder, Colorado, July 15-17, 1974.

14. R. C. Durfee, "ORRMIS - Oak Ridge National Laboratory, Oak Ridge, Tennessee, September 1974.

TECHNICAL INFORMATION DIVISION
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720