# REGIONAL LAND INFORMATION SYSTEM DEVELOPMENT USING RELATIONAL DATABASES AND GEOGRAPHIC INFORMATION SYSTEMS

Laurence W. Carstensen Jr.
Department of Geography
Virginia Polytechnic Institute and
State University
Blacksburg, VA U.S.A. 24061

## ABSTRACT

Currently in the United States, more and more small government agencies are turning to automated systems for handling required land information. The benefits of automation can be great, and, with improved technology and increased competition among vendors, the costs have been drastically reduced. This paper describes the thought process in the development of a Land Information System (LIS) for a regional planning agency in the United States. The needs and goals of the users of the system dictated that it be flexible allowing both geographical data (maps) and attribute data to be stored and analyzed together. A dearth of microcomputer-based turnkey packages for this purpose in late 1984 forced the creation of a hybrid system that could combine a commercially available software package for attribute data (a relational database) with a software package for map analysis (Geographic Information System).

## THE PERCEIVED NEED FOR A LAND INFORMATION SYSTEM

The Fifth Planning District Commission (PDC) is a regional planning agency of the commonwealth of Virginia located in the city of Roanoke, Virginia. The PDC jurisdiction includes four counties (1700 sq. miles), a Metropolitan Statistical Area of 250,000 people, and many small towns and cities (Figure 1). Land uses in the region range from urban-industrial to national forest. In January of 1983, the Fifth PDC convened a committee of representatives of constituent governments to determine those ways that the PDC could best assist in the orderly process of planning. By February, 1984, it was decided that the greatest single obstacle to effective planning efforts was the scarcity of readily accessible data. Costs for searching through courthouse records manually as each need arose were becoming prohibitive. A centralized, computerized data base was to be housed at the PDC for all to have access. Data could be required at many different scales, resolutions, and by various units of agregation, therefore the land parcel was determined to be common denominator for all the expected needs. The system would consist of thirty-seven criteria that would be collected for all land parcels in the four county area.

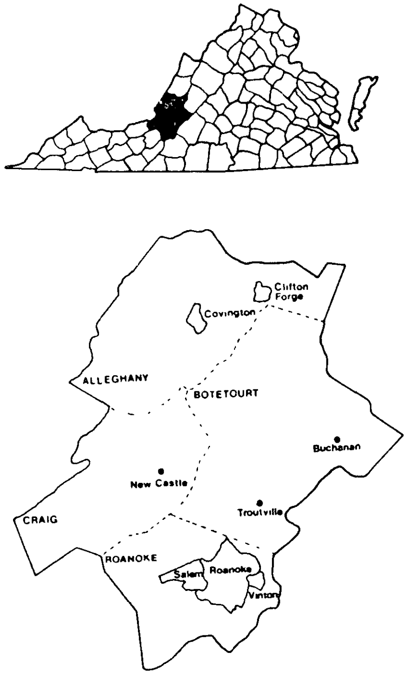


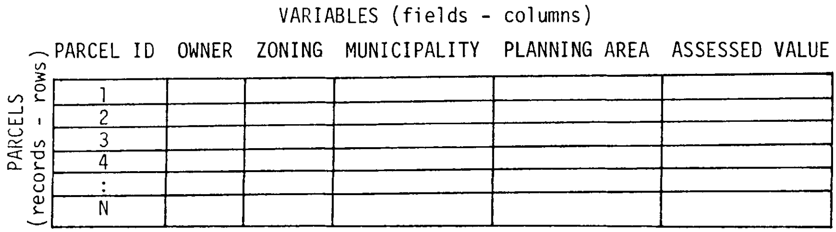
Figure 1. The Fifth PDC Region

## INITIAL PLANNING OF THE SYSTEM

The greatest obstacle to system development was a rather limited budget. Many extensive systems have been developed for minicomputers and mainframe computers, but the budget voided all such considerations (American Farmland Trust, 1985). The system had to be microcomputer based. Yet, available software for developing a land information system was very scarce in 1984 (it is only slightly more available today). Thus, developing a workable system for a set of needs which were not fully understood at the time of the search and are constantly evolving is a very difficult task. Even the meaning of the terms of the field: Geographic Information Systems (GIS), Spatial Information Systems, and Geographic Data Bases, seems to be a matter of opinion on the part of those who develop them, thus talking to vendors is not very helpful. Peuquet (1977) provided a broad definition of a GIS in a paper eight year ago, yet some vendors seem unaware of the effort. In this project, all agreed that a relational database was suitable to store attribute data for the land parcels. Yet, as the expense to support a GIS would be greater, the need was debated by a consideration of the needs of the PDC and its constituent governments.

## Relational Databases as Land Information Systems

A relational database is one in which the data are arranged in a matrix. Each row is referred to as a record (or observation), a group of individual data items that are logically related together. Within each record are fields. A field contains one item concerning the record and is analogous to a column in the database (Ashton-Tate, 1984, p. 1-14). In an LIS application, the rows (records) of the matrix represent individual land units (of any desired description), and the columns (fields) represent various characteristics of those land units. For example, in a system based on land parcels, the records represent the different land parcels, and the fields represent various characteristics of those parcels such as ownership, zoning, municipality, planning area, assessed value, etc. (Figure 2).



| PARCEL ID | OWNER | ZONING | MUNICIPALITY | PLANNING AREA | ASSESSED VALUE |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| : | | | | | |
| N | | | | | |

Figure 2. A Relational Database Structure for LIS

This database structure makes certain operations very simple to accomplish. Searches for land parcels having particular characteristics are simple because the database maintains all information in consistent fields (columns). For example, all parcels over 10 acres in size that are zoned for agriculture may be identified by searching only two columns in the data file. For each parcel that meets the criteria, the parcel identifier can be printed giving a listing of all qualifying parcels. Because all the data values across a row refer to the same land unit, many fields can be combined in a single search. By their internal structures, relational databases can produce and process files of nearly unlimited numbers of records and more than 100 fields. Total file sizes may be nearly as large as the mass storage devices used by the hardware system; they are not limited by the relatively smaller memories of microcomputers. The major advantage of a relational database is its simplicity. Adopted for land information, the system can provide some analysis capabilities, and can produce analyses very quickly. Due to its simple structure, it is easy to learn and use. It can be operated on very basic hardware requiring only a computer and a monitor (a printer is helpful but not required).

### Geographic Information Systems as Land Information Systems

A Geographic Information System (GIS) is one in which the data to be analyzed are stored in a cartographic data structure (usually raster or vector). Maps each cover exactly the same ground area, and are referred to as "overlays", single-factor maps", "images", and "data elements," an element containing a map of one and only one particular thematic distribution (Berry, pp. 1-2). For example, if a GIS were to cover a county in the United States, one element might contain a map of political boundary information, another a map of the highway network, another of population density, and another of the land use for the county (Figure 3).

Because the data are all stored as maps, two major functions of a GIS are the entry and display of maps. Other functions typically include varied techniques for comparing maps (elements) by overlay, and for analyzing individual maps. For example, an overlay of an element containing existing land use polygons upon an element containing proposed zoning polygons would yield a third element depicting the areas that would be in violation of the new ordinance (Figure 4). Analysis functions within single elements, either original elements or those created by the GIS during overlay, would minimally include the ability to measure areas. An area analysis function could be called to determine the number of acres that would be in violation of a new zoning ordinance by measuring the new element created during the overlay of existing land use and the proposed zoning ordinance. Additional analysis functions should include neighborhood searches in which distances are determined from specified map features. For example, one could determine the distance of each land parcel in a jurisdiction from the nearest shopping center. Some GIS packages contain far more analysis functions than others, thus care should be taken in deciding which functions are necessary for a proposed implementation.

## A CARTOGRAPHIC VIEW OF THESE TWO LIS STRUCTURES

From the planner's standpoint, the system was to be used as an electronic file cabinet to support data entry, retrieval, and analysis. The problem with the file cabinet analogy is

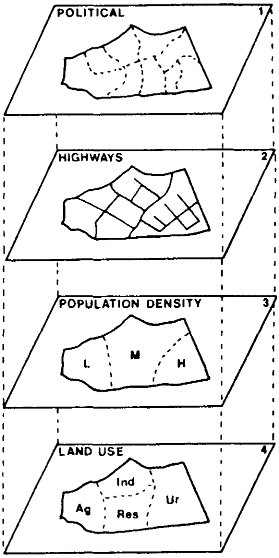


Figure 3. A GIS Database Structure

that it is very limiting. Because data analysis is clearly a geographic process, a spatial-cartographic analysis of needs was also carried out during deliberations.

As originally envisioned by the committee, the system was strictly choroplethic. That is, the data were to be collected by previously defined units of land, the values to be collected for each parcel had to be considered homogeneous throughout the entire parcel, and the values had to be treated as discrete by parcel. For such a system, one might devise other data structures, but the microcomputer software marketplace offers many relational database structures. For most of the variables to be placed into the system, a relational database of choroplethic parcels is appropriate. Most of the variables are discrete and homogeneous within each parcel.

However, choroplethic structures have many weaknesses. First, choroplethic maps are the most generalized of all maps, often to the point of significant data loss. For example, a region of people having low income levels may be split among several enumeration units such that none of the units appear to have low income levels (Figure 5, after Sinton, 1977).

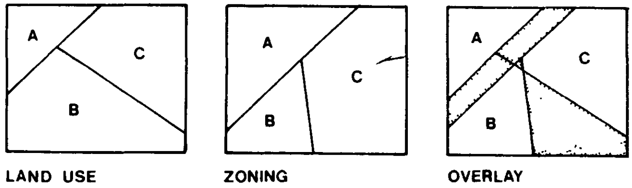


Figure 4. Map Overlay

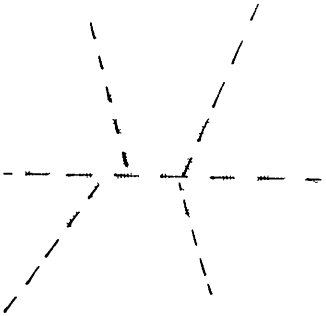


Figure 5. Choroplethic Shortcomings

Data must be fit into land units that have been defined with no regard to any spatial components of the data. That is, the spatial structure of the enumeration procedure takes precedence over the spatial structure of the data themselves. Data values are discrete within each enumeration unit and as such values can change only at enumeration unit boundaries. Choroplethic structures are useful only for area data, thus the inclusion of data related to points or lines is not feasible in a single data file. Also, many analyses are very difficult to accomplish using strictly a relational approach because of a lack of spatial topology.

As an example of the ability of a choroplethic-relational structure to deal with a possible request, suppose that an industry becomes interested in locating in the Roanoke area. This industry is well aware of its needs for land, and presents the following criteria to the PDC for site evaluation:

1. The land must be at least five acres (2 hectares) in extent.

2. The land must be zoned for industry.

3. The land must be currently vacant or for sale.

4. The land must not be subject to flooding.

5. The land must not be over one mile from a heavy-duty highway allowing truck traffic.

6. The land must have no slope over ten percent.

How well could a relational database structure using parcel land units deal with this request? The question is one of simplicity (generalization) versus reality. It is possible to load any information one wishes into the columns of a relational structure, but one must be aware of the results of such decisions. A good understanding of the nature of the dimensions of spatial data (point, line, and area), and spatial data variation (discrete-continuous) is critical.

The first query is reasonably analyzed by a relational data structure. Available land is in parcel form, therefore areas for each parcel can be stored as items in the data base. Such a value is clearly discrete as it relates only to the parcel under consideration. The second query can also be answered well by a relational database providing that the zoning ordinance doesn't cross-cut the parcels such that there are several different zoning designations for various parcels. In the PDC database such was not the case, though some parcels were designated as mixed. The third query is also simple for the choroplethic-relational structure. A parcel is either vacant or not and is either for sale or not. This logical dichotomy is both homogeneous and discrete to each parcel. Query number four is not well handled by a relational structure. The concept of flooding is related to linear data (streams and rivers) not to area data, and the incidence of flooding is certainly not constrained by the locations of parcel boundaries. To generalize a flood plain to a land parcel system, one must determine whether a parcel is in a flood prone area or not. Clearly, for the sake of safety (and to avoid disasterous legal consequences) a parcel must be coded as in the flood plain if even a slight portion of it is subject to inundation. A relational search for parcels that are not in the floodplain may exclude some very good sites for this industry because of small portions of parcels that might flood. The generalization of these data is unrealistic and too crude to support a valid analysis. The fifth query, the distance to the nearest major highway, is also inappropriate to be answered by a relational structure. As the structure is set up for the inclusion of information only on areas, the use of linear data is difficult. Certainly there is no way that the system could produce the distance estimate needed, as it does not contain the spatial elements of topology or scale. Relying on manual measurements to create a field for the database is possible, but is incredibly inefficient. The PDC could hire staff to perform this task, yet their time would prove far more costly than the costs involved in designing a more flexible system. The final query involves additional spatial problems concerning both generalization and spatial variation, and is innappropriate for a relational data base structure. Slope is based on elevation difference; elevation is a point phenomenon. The assignment of a single slope (or even several) to an entire parcel of land is a vast simplification of reality which may cause major innaccuracies in the data base. Further, slope is a continuous as opposed to a discrete phenomenon, thus it cannot be described well by either a measure of extremes (as needed in this request) or of central tendency. Its entry into the system would have to rely on manual measurements which, again, would prove very costly to compile.

It is evident that a choroplethic-relational approach to this data base is valid for some analyses, but very poor for others. For data that are truly discrete and homogeneous according to the parcel system, the relational database is not only appropriate but also optimal. It is a simple system with which to work, very quick to perform analyses, and is inexpensive to purchase. However, when the degree of generalization required to include certain variables becomes too great to be of use for many analyses, other structures such as a GIS should be considered to <u>complement</u> the relational structure.

To consider the GIS approach, let's look at the same example. First, the GIS would need to store up-to-date, accurate maps of five themes:

1. Land Parcel Boundaries
2. Zoning Boundaries
3. Streams
4. Major Highways
5. Elevation

All of the queries could be satisfied by the system. Area calculations can be accomplished by most GIS packages. Overlay of the zoning map on the parcel map could determine those parcels zoned for industry. The stream element could be combined with the elevation element to determine the flood-prone areas. These areas could then be overlaid on the parcel maps to look for areas in each potential industrial site that are subject to flooding. The cartographic neighborhood search functions of the GIS could determine the distance from each major highway to each parcel, and the elevation element could be used to calculate slopes. The only difficulty might occur in tying the attribute values of vacancy/occupied and for sale/not for sale to the parcels, but this could be overcome by recoding the land parcel element in an appropriate manner.

Because GIS are generally far more flexible than realtional structures, the question in a GIS environment is not so often whether the system <u>can</u> satisfy a particular request, but whether the system <u>should</u> be used to answer particular queries. Three of the six queries in this example can be answered accurately and quickly by a relational database using far fewer computer resources than by a GIS. The other three are better answered using the GIS as it maintains a spatial topology, allows for multiple data dimensions (points, lines, and areas) to be included in an analysis, and has many different types of analyses available. It is clear that both structures could be used together more effectively than either one alone.

## SOLUTIONS FOR THE FIFTH PLANNING DISTRICT COMMISSION

The PDC required that the system be available to answer a variety of queries. There was a strong sentiment that the system should not be as severely limited as the choroplethic-relational structure, thus both a relational database (dBASE III) and a GIS (AE-GIS) were purchased. It was felt that the GIS could be used to perform some analyses, and to provide data for the relational database. For example, though it is feasible with the GIS software, recalculation of the distance from each parcel to the nearest major highway every time a query requests the information is a waste of computer resources. As such values are appropriate for inclusion in the relational data base (i.e. values are discrete), the GIS can calculate the values once, and they can be stored in the relational database. The same procedure is useful for area measurements, and has been used for determination of dominant soil types by parcel.

Though the GIS and relational database packages do not communicate directly with each other, they are each able to produce data files on the system's hard disk which may be reformatted by in-house software routines for use by the other package. The Fifth PDC LIS integrates both a relational database and a GIS into a very workable product. The PDC is able to better serve its members by providing statistical analyses and geographic images in support of planning efforts. The interface between the relational database and the GIS has helped to overcome many obstacles to the fulfillment of the initial goals. By using both a relational database and a GIS, the system has quick response to many sorts of requests for information through the relational database, yet it is not limited in its analysis capabilities to using only choroplethic data.

## REFERENCES

The American Farmland Trust (1985), "Survey nof Geographic Information Systems for Natural Resource Decision-Making at the Local Level," 1717 Massachusetts Avenue, NW, Suite 601, Washington, D.C.

Ashton-Tate (1984), _dBASE III User's Manual_, Culver City, CA.

Berry, Joseph (1984), "Cartographic Modeling: Computer-Assisted Analysis of Maps," _Papers in Spatial Information Systems_, Yale School of Forestry and Environmental Studies.

Peuquet, Donna J. (1977), "Raster Data Handling in Geographic Information Systems," in Dutton, G. ed., _Proceedings of the Advanced Study Symposium on Topological Data Structures for Geographic Information Systems_, Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.

Sinton, David F. (1977), "The Inherent Structure of Information as a Constraint to Analysis: Mapped Thematic Data as a Case Study," in Dutton, G. ed., _Proceedings of the Advanced Study Symposium on Topological Data Structures for Geographic Information Systems_, Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.