

A GENERAL DATA MODEL FOR GEOGRAPHIC INFORMATION SYSTEMS

Guaraci Erthal
Gilberto Câmara
Diógenes Alves

Institute for Space Research
P. O. Box 515
12201 - São José dos Campos (SP)
Brazil

Comission: IV (Intercom. III/IV)

ABSTRACT

Recently, a major trend in Geographic Information Systems is integrating remote sensing data in such environment. This work presents a GIS to be used in remote sensing applications, together with an image processing system, enabling satellite imagery to be combined with thematic maps, DTMs, and tabular data. In designing an integrated system, vector and raster data must be treated on a common database. In order to overcome such diversities, the design relied on abstract data types. The data base uses a geo-relational model.

1. INTRODUCTION

The technology of geographic information systems (GIS) has grown to be one of the important areas in the Digital Mapping field. One of the general characteristics of a GIS is the presence of a large number of different formats of data organization. More recently, trends in GIS include the integration of remotely sensed data, the use of a relational data base, and a topologically structured data organization.

A growing concern in GIS is proposing some model that takes into consideration the particular nature of geographical data and may account for its diversity (Camara and Erthal, 1987). This paper presents a general data model for building a GIS, whose aim is to provide a framework for a single environment that integrates the various data formats that make up a GIS. The model is based on the concepts of abstract data types, that have proven to be a most useful tool on the Software Engineering area (Shaw, 1984).

The motivation for this work is the development of a GIS in Brazil's Institute for Space Research, that would be able to integrate remotely sensed data in a GIS in a natural fashion. In the model's derivation, the typical characteristics of GIS applications such as Automated Cartography, Geology, Urban Planning, and Surveying were taken into account. The concepts of this model extend earlier work (Lauzon et al., 1985) in that remotely sensed data and digital terrain models are included as parts of the system.

2. GEOGRAPHICAL DATA FORMATS

The main objectives of a GIS are (Nagy and Wagle, 1979; Tomlinson and Boyle, 1981):

- to integrate, in a single data base, spatial information from thematic and topographic maps, remotely sensed data, geophysical and geochemical data, and census data;
- to combine the various types of information, by means of manipulation algorithms, to generate derived maps;
- to manage a geocoded data base, including the reproduction, combination and visualization of its content;

In a broad sense, GIS data have two components: (a) the geographical phenomenon (its identification and value); (b) its spatial location. This means that a GIS deals not only with graphical data, but also with non-graphical data (attributes of geographical entities).

The geographical phenomena may be represented digitally in various graphical formats, depending on its origin and on the nature of the data capture and manipulation facilities. An example would be the case of urban surveys, where the geographical area of a parcel is related to its description (including its owner and the taxes). An example would be the case of urban surveys, where the geographical area of a parcel is related to its description (including its owner and the taxes). An example would be the case of urban surveys, where the geographical area of a parcel is related to its description (including its owner and the taxes). An example would be the case of urban surveys, where the geographical area of a parcel is related to its description (including its owner and the taxes). An example would be the case of urban surveys, where the geographical area of a parcel is related to its description (including its owner and the taxes). Among these data formats are:

- arcs : 2D coordinates that separate geographical regions;
- nodes : intersections of two or more arcs;
- polygons : geographical regions in 2D space;
- points : sparse values, based on terrain measures;
- regular grids : 2D matrices of Digital Terrain Model (DTM) data;
- triangular meshes: irregularly-shaped regions that link sparse DTM data together;
- images: data in raster format, whose values are usually quantized to a fixed range.

Out of the formats cited above, the first four are usually associated with vector-based systems; regular grids and triangular meshes are alternatives for handling DTM data, and images is the preferred format for remotely-sensed data.

3. BASIC CONCEPTS FOR THE MODEL

Given the nature and diversity of geographical data, it is important to devise a conceptual model. Its objectives are to distinguish among the various kinds of entities and link graphical and non-graphical data, to use a relational data base, and to enable both vector, grid and raster data formats on a single system.

The general data model comprises a series of concepts, enabling a GIS to be broken into different levels of abstractions. These concepts, presented below, are: category, geographical object, relation, class, project, information plane and cartographical document.

3.1 CATEGORIES AND OBJECTS

The different kinds of geographical data were grouped together into categories, according to its nature and origin. Each category brings together geographical information of the same kind, and the model has the following categories:

- Areas (poligonal data): geographical regions separated by polygons;
- Digital Terrain Model (DTM): spatial distribution of physical variables (such as altitudes or mineral content);
- Networks: linear information, such as streets, electrical and sewage systems;
- Hydrography: representations of water bodies of linear (rivers) and areal (lakes) nature;
- Points: locational entities, which indicate the existence of a ponctual phenomenon in space;
- Images: data obtained by remote sensing satellites, in the raster format.

The definition of the categories took into account the topological content of geographical data. Areal (regions) data, networks and hidrography are better represented on a vector format, and have distinct topology. Areal data usually uses an arc-node-polygon topology, whereas networks are restricted to arc-node topology, and hidrography is a mixture of areal and network topologies.

In the case of images, the topology is usually non-existent, and for DTMs it is obtained by grouping data of the same value. The topology for DTMs is materialized when isolines are drawn.

Whithin each category, the geographical entities to be manipulated are similar, and, in order to integrate a relational data base, the model uses the abstraction of a geographical object (or object, for short). An object corresponds to an instance of a geographical variable, and objects are the entities manipulated by the relational data base. For example, an electrical network map belongs to the network category, and elctrical trunks are examples of objects in such a map.

A more detailed discussion on the categories of the model is presented in Table 1. For each category, some examples of geographical data and of possible objects for each map are given, and the data formats supported are presented.

TABLE 1
CATEGORIES OF THE GEOGRAPHICAL DATA

| Category | Map Examples | Geographical Objects | Data Formats |
|------------------------------|--------------------------------------|------------------------|--|
| Area1 Data (AREA) | Land Use, Political, Territorial | Farms, States, Parcels | Vector, Raster |
| Digital Terrain Models (DTM) | Topography, Geophysics, Geochemistry | Terrain Element | Samples, Grid, Raster, Vector, Triang. |
| Networks (NET) | Road maps, Sewage systems | Network trunks | Vector, Raster |
| Hydrography (HYDRO) | Rivers, Lakes. | Water Bodies | Vector, Raster |
| Mulitespectral Images (IMAG) | Remotely sensed Data | Band | Raster |
| Points (PTS) | Churches, Airports | Isolated entity | Vector |

3.2 CLASSES AND RELATIONS

The geographical objects of the general model are characterized by the following attributes:

- IDENTIFIER: a cardinal attributed to the object that serves as unequivocal identifier;
- SPATIAL REPRESENTATION: includes the TOPOLOGY and the GEOMETRY. The topology of the object describes the spatial relationship of the object to its neighbors. The geometry of an object is its geographical location.
- RELATION: non-graphical object properties, expressed as a tuple (instance of a relation) on a data base management system that uses the relational approach.
- CLASS: attribute of a group of objects that have a common characteristic, as chosen by the user. The class of an object may result from the application of a pattern recognition algorithm on satellite imagery, or from the choice of one of the relational attributes.

3.3 PROJECT AND INFORMATION PLANE

Once the notions of category and object are established, the working structure of the system may be defined. The basic unit of the system is that of a project. A project corresponds to analysis over a region, which use the same data set. Usually, each user of a GIS will have his own project. The projects are not directly related to each other, but parts of a project's data may be used by another.

A project contains one or more information planes (IPs), and each IP belongs to a given category of the model. The IP is a fundamental abstraction of the system, in the sense that it enables the data present in it to be seen by the user in a unified fashion. This is possible because of the fact that the topology and the geometry of each category are fixed. Therefore, the validity and the effects of the system operators are easily determined by the system.

Examples of IPs are:

- Topography: an IP of the DTM category, which may have the grid (regular grid), vector (isolines) and raster (classified image) formats;
- Land Use: an IP of the Areal category, which may have the vector and raster formats;
- LANDSAT TM Band 4 : an IP of the Image category, which is available only in the raster format.

Each IP is composed by a set of geographical elements (objects) that have a unique identification. For each IP, a set of relations may be defined. Objects with the same set of non-graphical attributes are seen as instances (tuples) of a given relation.

The objects in a IP map are only available if there is a vector representation of the geographical variable being analysed. The instantiation of each object occurs only at the moment there is information enabling the system to individualize it. For example, a thematic map (of the Areal category) on the raster format (corresponding to a classified satellite imagery) will have its objects instantiated only when a raster-to-vector transformation is applied.

Each IP has a CLASS TABLE, that indicates, for each class, the colour chosen for visualization, and the associated height, in the case of DTM. This is important in the raster representation of an IP, where each image element ("pixel") is usually associated with its class.

A formal definition would view an IP as a n-ary relation (C, O, R^*, F, G) , where:

- C: set (c_1, \dots, c_n) of classes;
- O: set (o_1, \dots, o_n) of objects;
- R^* : list (R_1, \dots, R_n) of relations;
- F: function of class-attribution, $O \rightarrow C$ (to each object, associates its class);
- G: function of relation-attribution, $O \rightarrow R^*$ (to each object, associates its relation);

3.4 INTEGRATION TO A RELATIONAL DATA BASE

The integration of a relational data base to the model requires the definition of relations to each IP. In the above formal definition, each object o_i may be seen as a tuple belonging to the relation R_i , which is the result of the function $G(o_i)$.

In practical terms, the identification of each object on an IP being unique, it is easy to retrieve individual objects from the data base, using the identification as a bridge.

3.5 CARTOGRAPHIC DOCUMENTS

One of the central aspects of a GIS is the possibility of map generation containing a combination of information over the region of interest. It should be noted that the usual geographical map will combine data from various information planes. Therefore, the model defines a cartographical document as an output product to be generated by the system, combining parts of several information planes.

3.6 MODEL HIERARCHY

In a resumé of the model's general organization, the following hierarchy was defined:

- * Project
 - * Information Plane (of a category)
 - * Objects (tuples of a relation)

4. MODEL OPERATORS

The operators of the model are the usual GIS functions, which can be divided in five classes:

- Definition : environment set-up;
- Entry : data input to the system;
- Conversion : format change;
- Manipulation: operations on the geocoded data base.
- Retrieval : access to the relational data base.
- Output : generation of cartographic documents.

The definition operators include:

- Project initialization;
- Information plane description: includes the description of a new IP;
- Relation definition: the relations in an IP are defined, including its attributes (domains).

The conversion operators include both raster-to-vector and vector-to-raster functions, as well as contour line generation (which transforms a grid onto vector format) and regular grid refinement (which generates a raster image from a regular grid).

The manipulation operators generate new IPs from a combination of existing ones. These operators usually work on classes of a PI (for example, "show all corn fields within three miles of the river"). The operators include:

- Distances : generation of a map containing distances from a selected subset of classes in an IP.
- Slope : slope map generation (of the Areal category) out of a DTM.
- Reclassification : definition of a new IP by the union of several classes of an existing IP. The category of the output IP is the same as the input one, except for the case of a DTM, which outputs an Areal IP.
- Overlay : Logical operations over two (or more) IPs. This operation is not possible where DTMs are involved.

- Classification: pattern recognition operator, which finds the classes of a satellite image by a choice of algorithms (including statistical and deterministic techniques). The output IP is of the Areal category.

Most of the manipulation operators work both on vector or raster data, excluding the classification one.

The retrieval functions enable the realization of relational operations on the non-graphical attributes of an IP. Unlike the manipulation operators, the retrieval functions work on objects and their relations. The usual operations of relational algebra (selection, projection, junction and division) are completed with a set of special operations, which include:

- Spatial query display: the results of a relational operation are shown graphically;
- New IP creation: a new information plane is created, containing the objects that are the result of a query.

The two most important output operators are the definition and the generation of a cartographic document. The definition bring together all map generation parameters, such as the set of IPs involved, the legend, and the output transformation (scale, projection and device coordinates).

5. CONCLUSIONS

The model was applied as the conceptual framework in the development of a GIS (INPE, 1987), aimed at a large number of applications. The system has been tested in a number of practical problems, including the development of a prototype urban data base, urban data studies using satellite imagery, and digital terrain modelling applications. In all cases, it was verified that the abstractions use enable the user to combine all geographical data together easily.

Extensions of the model include the creations of data hierarchies, and the use of concepts such as aggregation and generalization (Smith and Smith, 1977). Another important extension would be the use of inheritance operators - typical of object-oriented languages, such as Smalltalk-80 (Goldberg and Robson, 1983) - in order to enable geographical objects to communicate properties and relations.

6. BIBLIOGRAPHY

CAMARA,G.; ERTHAL,G., "Integration of relational models in geographic information systems " II Brazilian Symposium on Data Bases, Porto Alegre, May 1987, pg 39-48 (in Portuguese).

GOLDBERG, A.; ROBSON, D. "Smalltalk-80: The Language and its Implementation", Addison-Wesley, 1980.

INPE (various authors), Specification of GIS version 2.0. Institute for Space Research, S. José dos Campos, Setembro de 1987 (in Portuguese).

LAUZON,J.; MCLAREN,R.; HARWOOD,C. "Database support for geographic information systems - the System 9 approach". Wild Heerbrugg report, 1985.

NAGY, G. and WAGLE, S. "Geographic data processing" ACM Computing Surveys, 11 (2), June 1979, 139-181.

SHAW, M. "Abstraction Techniques in Modern Programming Languages", IEEE Software, (1):4, pg. 10-27, October, 1984.

SMITH, JM and SMITH, DCP "Database Abstractions: Aggregation and generalization" ACM Trans on Database Systems(2):2, June 1977, 105-133.

TOMLINSON, R. and BOYLE, R. "The state of development of systems for handling natural resources inventory data", Cartografica, 18 (4), 1981, 65-95.