NASA

Technical Memorandum 86113

GEOCODED DATA STRUCTURES AND THEIR APPLICATIONS TO EARTH SCIENCE INVESTIGATIONS

Michael Goldberg

JUNE 1984

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

TM 86113

GEOCODED DATA STRUCTURES AND THEIR APPLICATIONS TO EARTH SCIENCE INVESTIGATIONS

Michael Goldberg

Space Data and Computing Division

June 1984

NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

A geocoded data structure is a means for digitally representing a geographically referenced map or image. This document reviews the characteristics of representative cellular, linked, and hybrid geocoded data structures. It then outlines the data processing requirements of Earth science projects at the Goddard Space Flight Center, describes the basic tools of geographic data processing, and presents specific ways that new geocoded data structures can be used to adapt these tools to scientists' needs by expanding analysis and modeling capabilities, simplifying the merging of data sets from diverse sources, and saving computer storage space.

CONTENTS

ILLUSTRATIONS

ACRONYMS

AVHRR	Advanced Very High Resolution Radiometer
GIMMS	Global Inventory Monitoring and Modeling Studies
GOES	Geostationary Operational Environmental Satellite
HCMM	Heat Capacity Mapping Mission
ISLSCP	International Satellite Land Surface Climatology Project
MSS	Multispectral Scanner
SIR-B	Shuttle Imaging Radar-B
SMMR	Scanning Multichannel Microwave Radiometer
TIROS	Television and Infrared Observation Satellite
TM	Thematic Mapper
VISSR	Visible Infrared Spin Scan Radiometer

ACKNOWLEDGEMENTS

The author acknowledges Mr. William J. Campbell, Dr. Philip J. Cressy, Mr. J.P. Gary, Mr. Marc L. Imhoff, Dr. Donna J. Peuquet, Dr. H.K. Ramapriyan, Dr. Vincent B. Robinson, Dr. Paul H. Smith, Ms. Regina Sylto, Dr. C.J. Tucker, and Mr. Ron Witt for their technical advice and editorial comments. The author also thanks Ms. Zachary Pantazes for her stylish typing.

1. BACKGROUND

Satellite remote sensing has greatly expanded the data resources of Earth scientists. There are now vast computer tape archives which represent, in the form of digitally coded images, coverage of the entire globe in a variety of spatial and temporal resolutions and spectral bands. There are also growing stores of geographically referenced data in more conventional forms, such as survey maps (both analog and digital) and aerial photographs. Scientists use image processing systems, geographic information systems, data base management systems, and other related systems to handle these data. New methods for digitally representing geographically referenced data can improve these systems by expanding analysis and modeling capabilities, simplifying the merging of data sets from diverse sources, and saving computer storage space.

This report introduces the reader to geocoded data structures and discusses the applications of new geocoded data structures to Earth science investigations, such as those being conducted at the Goddard Space Flight Center.

2. INTRODUCTION TO GEOCODED DATA STRUCTURES

2.1 DEFINITION

A geocoded data structure is a means for digitally representing a geographically referenced map or image. Geocoded data structures can be generalized to apply to any spatially referenced data, not just those data associated with locations on the Earth's surface.

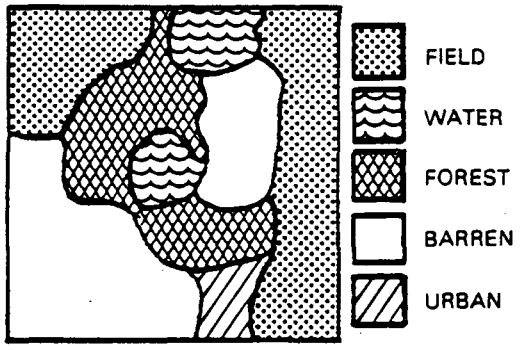
2.2 SOME REPRESENTATIVE GEOCODED DATA STRUCTURES

There are two major classes of geocoded data structures: cellular and linked. Recently, hybrid data structures have been developed which combine aspects of both the cellular and linked organizations. In addition, there are several other approaches to representing spatial data.

2.2.1 CELLULAR STRUCTURES

Cellular structures, also known as "grid" or "raster" structures, represent maps or images by subdividing them into subareas, also known as "cells" or "pixels". A computer stores the information associated with each cell.

The most basic cellular structure is standard grid structure (Figure 1a). In standard grid structure, a map or image is simply represented by a two-dimensional array of numbers. Each number corresponds to a uniform-sized rectangular cell on the original map or image. The storage address for the number implicitly defines the geographic location of the cell, based on the known position of control points and the assumed geometry of the map or image. The value of the number corresponds to the attribute associated with the cell. Standard grid structure is the most common format for representing maps and images in geographic information systems and image processing systems. It also corresponds directly to the format produced by raster-scanner input devices. It is easy to write manipulative software for almost any application using this structure, but



A CLASSIFIED IMAGE

1	1	2	2	1
1	3	3	4	1
4	3	2	4	1
4	4	3	3	1
4	4	4	5	1

1=FIELD
2=WATER
3=FOREST
4=BARREN
5=URBAN

THE IMAGE IN STANDARD GRID STRUCTURE

A.) STANDARD GRID STRUCTURE

3	4	5
2	5	5
1	3	2

BAND 1

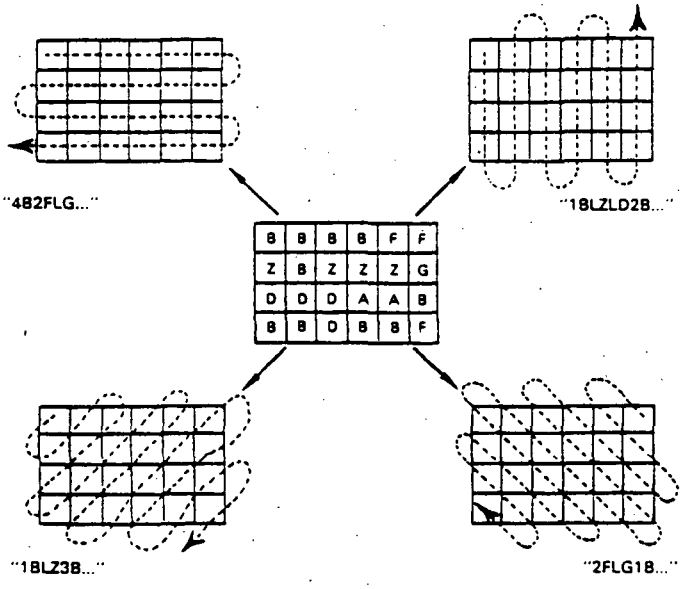
5	4	3
2	1	1
2	3	4

BAND 2

3	5	4	4	5	3
2	2	5	1	5	1
1	2	3	3	2	4

"BAND INTERLEAVED BY PIXEL"

B.) INTERLEAVED STRUCTURE



C.) RUN-LENGTH CODING SCHEMES (FROM AMIDON AND AIKEN, 1971).

FIGURE 1. CELLULAR STRUCTURES

for spatially sparse or highly redundant data, standard grid structure uses computer storage space inefficiently (Amidon and Aiken, 1971; Nagy and Wagle, 1979). Like all cellular structures, standard grid structure is not as accurate as linked structures for delineating region boundaries.

Sometimes, two or more standard grid-formatted maps or images which have been registered to a common spatial reference frame are placed in the same physical file in computer storage. These "multiple band images" are often organized in interleaved structure (Figure 1b). In this structure, cells which relate to the same geographic location, but which are from different sources, are stored adjacent to one another. Satellite data from multispectral scanners are transmitted in interleaved format to ground receiving stations and therefore are often stored in this structure.

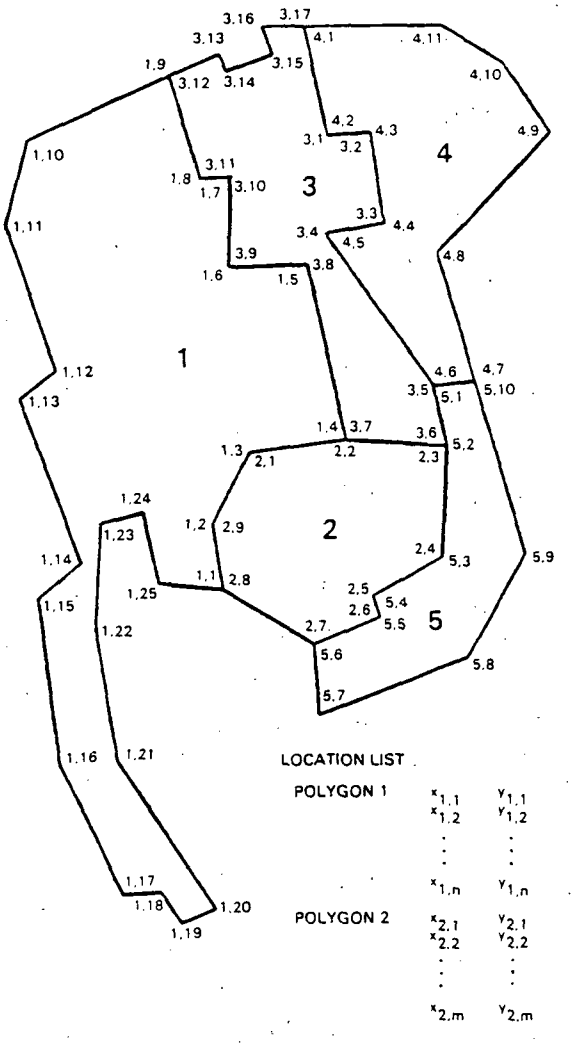
Run-length coding structure (Figure 1c) is a compressed version of standard grid structure in which strings of identical cell values are replaced by a single cell value and a repetition counter. For example, the string "AAAAA" is represented as "5A". Run-length coding structure has had practical application in data reduction of existing standard grid-structured data banks (Amidon and Aiken, 1971), and as an input structure for the manual encoding of maps.

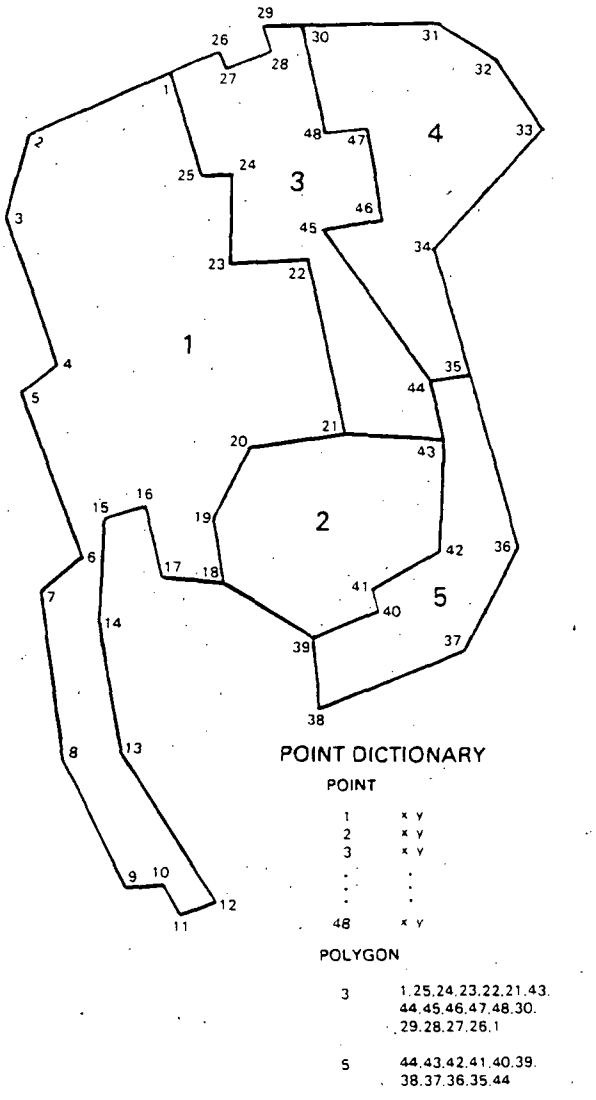
2.2.2 LINKED STRUCTURES

Linked structures, also known as "polygon" or "vector" structures, represent maps and images in terms of points, lines, and regions. Basically, points are described by their coordinates, lines are described by chains of points, and regions or "polygons" are described by chains of points which form a closed boundary. Each of these spatial entities is usually preceded in computer storage by a "header" which identifies it as a point, line, or region and which contains an associated thematic value. Linked structures are closely related to the data formats produced by coordinate digitizers and they can easily incorporate topological information. Except when extremely complex areas are digitized, all linked structures are quite storage efficient (Nagy and Wagle, 1979). Unfortunately, it is relatively difficult to write computer programs for editing and manipulating data stored in linked structures.

The most basic linked structure is the entity by entity structure (Figure 2a). In this structure, every point, line, and region is stored separately. Polygons are encoded without regard to adjacent or overlapping polygons. Lines are encoded without regard to intersecting or merging lines. While this approach is simple to understand and implement, it often results in lines along the boundaries of polygons being duplicated in slightly different positions. Editing is required to correct these "sliver lines" (Peuker and Chrisman, 1975).

The point dictionary structure solves the sliver line problem (Figure 2b). Each point is assigned a unique label. The label of each point and the point's coordinate are stored in a location dictionary. Each polygon is represented as a list of point labels from the dictionary. Since polygon boundaries are built from a common set of points, slivers cannot occur (Peuker and Chrisman, 1975).





A.) ENTITY-BY-ENTITY STRUCTURE (FROM PEUKER AND CHRISMAN, 1975).

B.) POINT DICTIONARY STRUCTURE (FROM PEUKER AND CHRISMAN, 1975).

FIGURE 2. LINKED STRUCTURES WITHOUT EXPLICIT TOPOLOGICAL INFORMATION

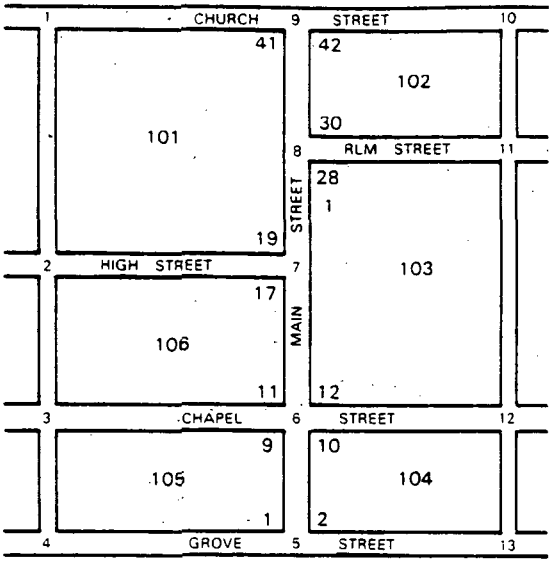
In line dictionary structures, lines are assigned unique labels. Each line label is associated with a list of points from a point dictionary. Polygons are in turn represented by lists of line segment labels from a line segment dictionary. This structure permits easier identification of points, lines, and polygons than the point dictionary structure (Peuker and Chrisman, 1975).

None of the linked structures described so far are computationally efficient for overlaying digitized maps, computing the areas of polygons, or determining the distance from a point to the boundary of a polygon. These operations require the use of structures which contain explicit topological information.

One of the first structures to contain explicit topological information was the dual independent map encoding structure developed by the U.S. Bureau of the Census (Figure 3a). The purpose of this structure was to establish a correspondence between street addresses and geographic coordinates. The basic entity in this structure is a straight line segment defined by its endpoints. The straight line corresponds to a street and the endpoints or "nodes" correspond to street intersections. Curved streets are represented by straight line segments joined by "pseudonodes". A line segment is represented in a computer as a record containing the two endpoint labels, the endpoints' coordinates (relative to a predefined rectangular coordinate system), codes for the polygons on each side of the segment, and the range of street addresses between street corners (Nagy and Wagle, 1979). Dual independent map encoding is suitable for the purposes of the Census Bureau, such as address coding and matching. It is unsuitable for more complex applications, such as polygon overlay, because it has no means for representing polygons as separate entities, and because it is unwieldy for manipulating complex line networks (Peuker and Chrisman, 1975).

One linked structure that contains explicit topological information and is designed for the rapid processing of many-sided polygons (and complex lines) is the binary searchable polygon representation (Figure 3b). Polygons are divided into "sections". A section is a set of line segments represented in a computer as a chain of coordinates. "Primitive sections" contain a single line segment. "Simple sections" contain a set of line segments that are monotonic in the x and y directions. "Basic sections" contain local extremes in the x and y directions. All other sections are called "compound sections". All of the different types of sections are organized into a binary tree hierarchy. Points of intersection of polygonal lines can be determined by what is essentially a binary search. Testing for the inclusion of a point in a polygon can also be performed quickly (Burton, 1977).

Another linked structure that is particularly suited to efficient computation of polygon intersections, distances from points to boundaries, and other geometric properties is the tightly closed boundary or parallel scan structure (Figure 3c). In this structure, all polygon boundary coordinates are sorted and partitioned into sets. Each set contains only points with the same y-coordinate. Computations on data stored in this structure utilize the property that a line crossing a closed boundary contour will intersect the contour an even number of times (Freeman, 1974; Merrill, 1973).



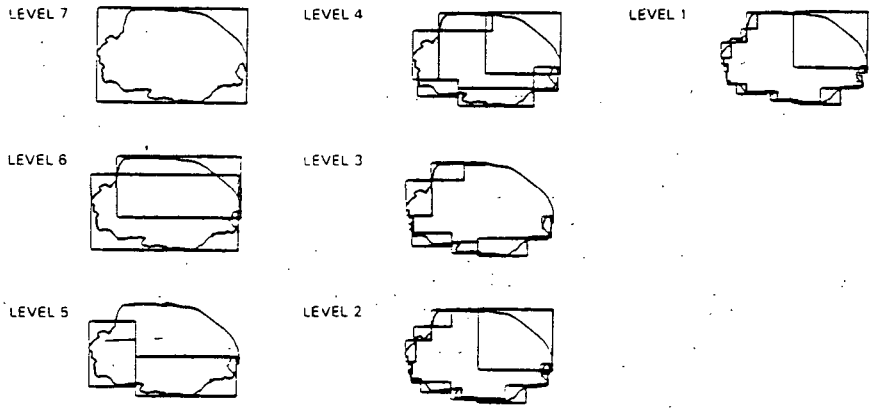
CENSUS ADDRESS CODING GUIDE RECORDS

STREET	TRACT	BLOCK	LOW ADDRESS	HIGH ADDRESS
MAIN	1	102	30	42
MAIN	1	103	12	28
MAIN	1	104	2	10
MAIN	1	105	1	9
MAIN	1	106	11	17
MAIN	1	101	19	41

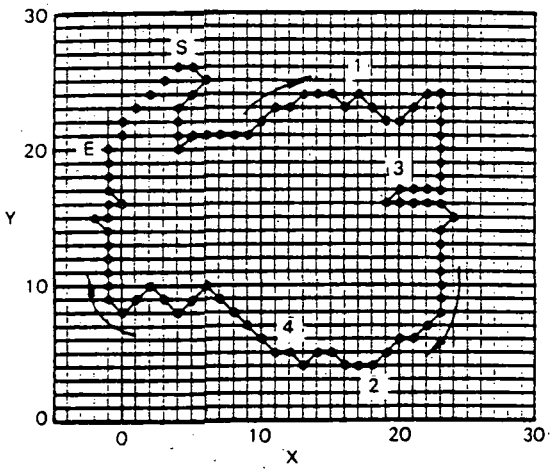
DIME STREET SEGMENT RECORDS

STREET	NODE START	NODE END	TRACT LEFT	BLOCK LEFT	TRACT RIGHT	BLOCK RIGHT	LOW ADDR.	HIGH ADDR.
MAIN	5	6	1	105	1	104	1	10
MAIN	6	7	1	106	1	103	11	17
MAIN	7	8	1	101	1	103	19	28
MAIN	8	9	1	101	1	102	30	42

A.) DUAL INDEPENDENT MAP ENCODING (FROM PEUKER AND CHRISMAN, 1975).



B.) BINARY SEARCHABLE POLYGON REPRESENTATION (FROM BURTON, 1977).



ORIGINAL BOUNDARY

(4,26), (5,26), (6,25), (5,24), (4,23), (4,22), (4,21), (4,20), (5,21), (6,21), (7,21), (8,21), (9,21), (10,22), (11,23), (12,23), (13,24), (14,24), (15,24), (16,23), (17,24), (18,23), ...

AUGMENTED BOUNDARY

(4,26), (5,26), (6,25), (5,24), (4,23), (4,22), (4,21), (4,20), (*4,20*), (5,21), (6,21), (7,21), (8,21), (9,21), (10,22), (11,23), (12,23), (*12,23*), (13,24), (14,24), (15,24), (*15,24*), (16,23), (17,24), (*17,24*), (18,23), ...

C.) TIGHTLY — CLOSED BOUNDARY STRUCTURE (FROM MERRILL, 1973).

FIGURE 3. LINKED STRUCTURES WITH EXPLICIT TOPOLOGICAL INFORMATION

2.2.3 HYBRID STRUCTURES

Hybrid structures combine characteristics of both the cellular and linked approaches.

Quadtree structure (Figure 4) is a means of storing cellular information in a linked format in such a manner that a map or image which has subareas of differing detail need not waste computer storage space by being divided into numerous tiny cells whose size corresponds to the highest required spatial resolution. Quadtree structure is a tree structure. The root node of the tree is associated with an entire (square) map or image. Besides the root, each other node is associated with one of the four quadrants of its parent node's square. If a node is associated with an entirely homogeneous quadrant, it has no descendant nodes. The value of a node which does have descendants is some aggregate of the descendant nodes' values (Hunter and Steiglitz, 1979). Thus, deeper levels on a quadtree represent successively finer subdivisions of a map or image, and the "leaf" nodes store the contents of individual cells. For maps or images which are mostly homogeneous but which contain subareas of high spatial variability, quadtree structure saves computer storage space. It whould be noted, however, that for highly heterogeneous spatial data such as unclassified satellite imagery, quadtree structure is much less storage efficient than standard grid structure.

The hierarchical nature of quadtree structure is another one of its assets. It has been suggested that a network of processors called a "processing cone" could take data in pyramid structure (a variation of quadtree structure) and efficiently perform edge finding, region growing, texture analysis, and other sophisticated spatial analysis functions (Shapiro, 1979). The advantages of quadtree and pyramid structures are offset in part by increased software development costs.

Line segment data and grid cell data are encoded together in a compact and versatile form through the use of vaster (from vector + raster) structure (Figure 5). In vaster structure, a map or image is divided into horizontal "swaths" consisting of groups of parallel scan lines. The "leading edge" of each swath (e.g., the scan line with the minimum y value) is represented in parallel scan structure which has been enhanced to contain information about map line intersections. The rest of the data in the swath are encoded using chain coding (see section 2.2.4 below) which has been compressed through a run-length coding scheme. Vaster structure has been envisaged as a means by which large spatial data sets can be digitized by raster scanners, partly converted into a vector structure, and then utilized by a wide variety of existing raster-oriented and vector-oriented algorithms (Peuquet, 1983).

Topological grid structure (Figure 6) maintains the simplicity and transportability of standard grid structure while providing the capability, previously limited to linked structures, to treat contiguous regions as distinct spatial entities. A map or image represented in topological grid structure consists of two equal-sized standard grid-structured data planes and an ancillary file. The first data plane is identical to the original map or image. Each cell in the second data plane contains a number which uniquely defines the contiguous region to which its corresponding cell in

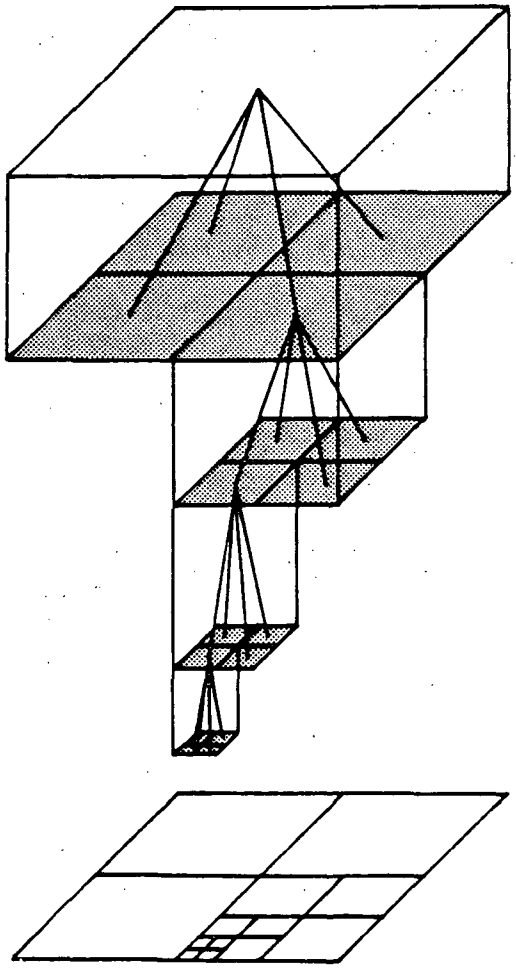


FIGURE 4. QUADTREE STRUCTURE (FROM HUNTER AND STEIGLITZ, 1979)

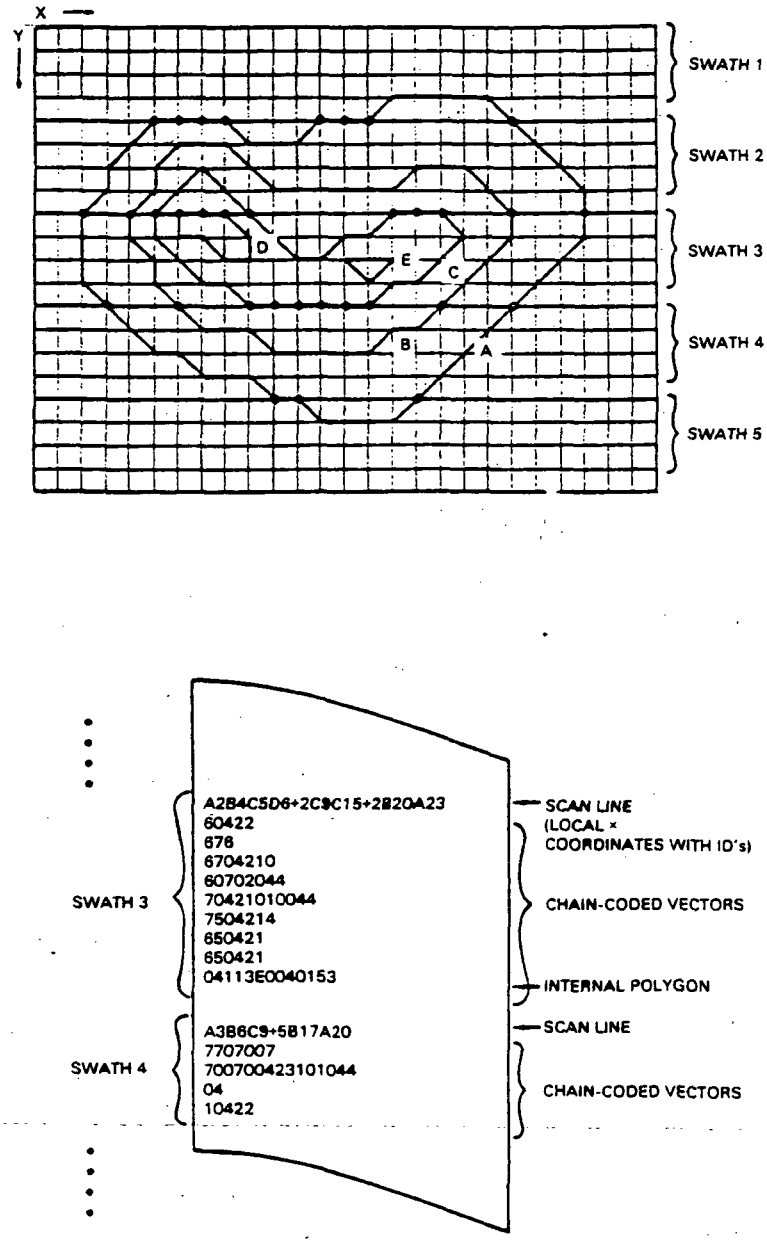
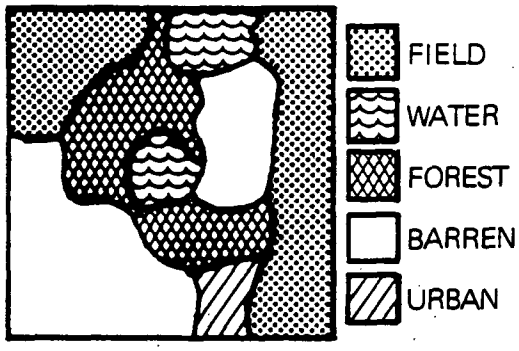


FIGURE 5. VASTER STRUCTURE (FROM PEUQUET, 1983).



A CLASSIFIED IMAGE

1	1	2	2	1	1 = FIELD
1	3	3	4	1	2 = WATER
4	3	2	4	1	3 = FOREST
4	4	3	3	1	4 = BARREN
4	4	4	5	1	5 = URBAN

THE IMAGE IN STANDARD GRID STRUCTURE

1	1	2	2	1
1	3	3	4	1
4	3	2	4	1
4	4	3	3	1
4	4	4	5	1

DATA PLANE # 1
ORIGINAL ATTRIBUTES:
IDENTICAL TO STANDARD GRID STRUCTURE REPRESENTATION

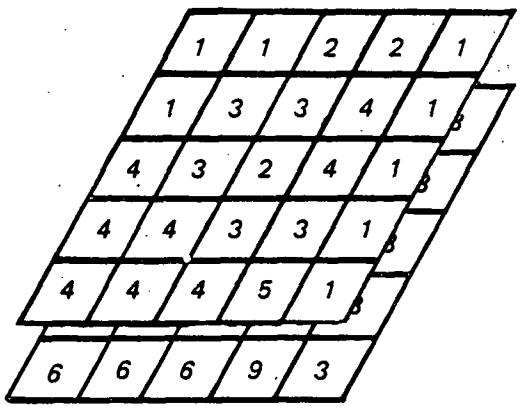
1	1	2	2	3
1	4	4	5	3
6	4	7	5	3
6	6	8	8	3
6	6	6	9	3

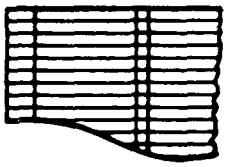
DATA PLANE # 2
CONTIGUOUS REGIONS UNIQUELY IDENTIFIED:
THE NUMBERS ARE THE CONTIGUOUS REGION ID's

(REGION BOUNDARIES HIGHLIGHTED FOR CLARIFICATION)

ATTRIBUTE	REGION ID	SIZE	OTHER . . .
1	3	5	
1	1	3	
2	2	2	
2	7	1	
3	4	3	
3	8	2	
4	6	6	
4	5	2	
5	9	1	

ANCILLARY FILE





THE IMAGE IN TOPOLOGICAL GRID STRUCTURE

FIGURE 6. TOPOLOGICAL GRID STRUCTURE

the first data plane belongs. Each record in the ancillary file contains an attribute identifier, a contiguous area identifier, and the contiguous area size (measured by cell count). The records are sorted in increasing numerical order of attribute identifiers. With each attribute, the records are sorted in order of decreasing contiguous area size. Any geographic information system or image processing system which can process multilayer data can be easily adapted to topological grid structure. Although topological grid structure enables the convenient implementation of region-based algorithms (Goldberg, 1984), it uses computer storage space inefficiently unless compressed through a run-length coding scheme.

2.2.4 OTHER STRUCTURES

There are several approaches to representing spatial data which are neither cellular, linked, nor hybrid.

Chain coding (Figure 7) can be used to represent lines at the most basic level in both cellular and linked organizations. Chain-coded lines consist of sequences of short segments connecting the centers of adjacent grid points. Lines are encoded as sequences of data nodes, where each data node in sequence coincides with one of the eight grid points that surrounded the previous data node (Freeman, 1974).

Non-rectangular structures (Figure 8) define point coordinates in terms of triangular or polyhedral coordinate reference systems (Gold, 1978; Thomas, 1978). Skeleton structures (Figure 9) represent regions as sets of "maximal neighborhoods" defined by their centers and radii (Pfaltz and Rosenfeld, 1967). Linguistic and complex recursive structures (Figure 10) represent spatial areas through language expressions (Freeman, 1974; Shapiro, 1979). None of these structures has yet been implemented on a large scale.

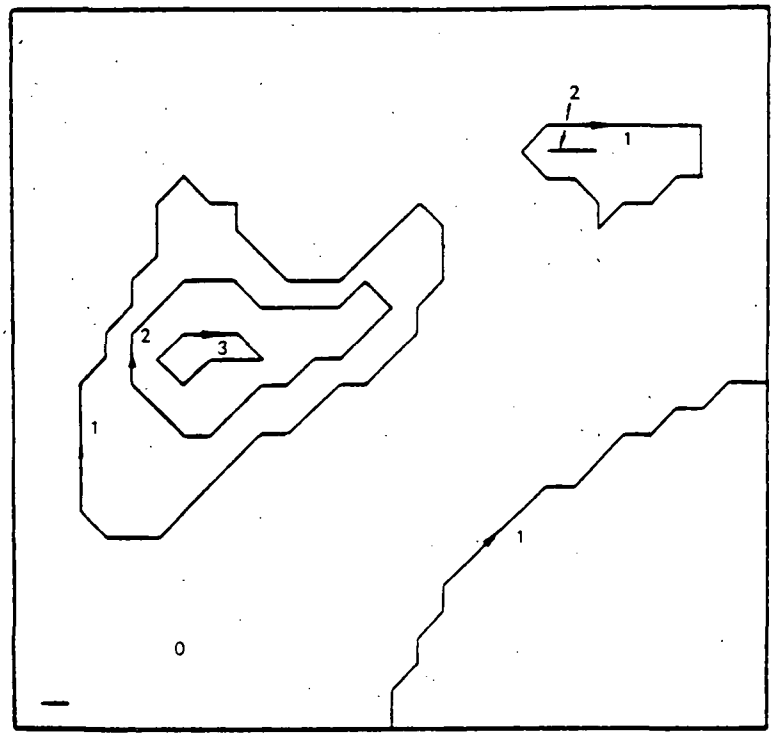
2.3 REMARKS

Review of the above information reveals that the design of a geocoded data structure is based on three factors: suitability for a particular class of algorithms (a function of software development costs and computing speed); correspondence to input/output device formats; and computer storage space efficiency. This leads to the conclusion that through intelligent use of geocoded data structures, one can: expand or improve analysis and modeling capabilities; simplify the merging of data sets from diverse sources; and save computer storage space.

There are many trade-offs involved in designing the best geocoded data structure for an application. No single structure is ideal for all needs. A discussion of the applications of geocoded data structures to Earth science investigations follows.

3. APPLICATIONS OF GEOCODED DATA STRUCTURES TO EARTH SCIENCE INVESTIGATIONS

Geocoded data structures affect all phases of geographic data processing. Their importance is particularly evident in Earth science investigations, which make intensive use of computer resources. Earth scientists can benefit greatly from the implementation of new geocoded data structures.



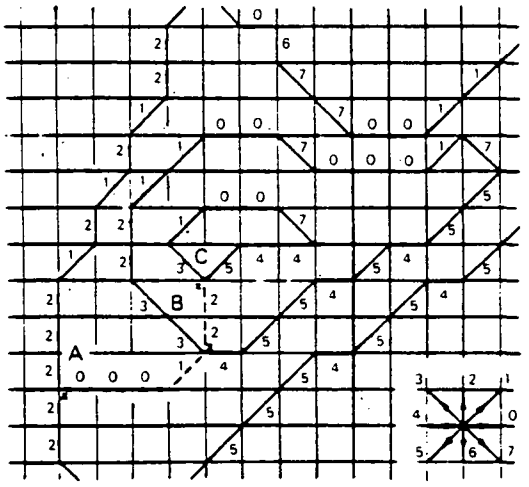


FIGURE 7. CHAIN CODING (FROM FREEMAN, 1974)

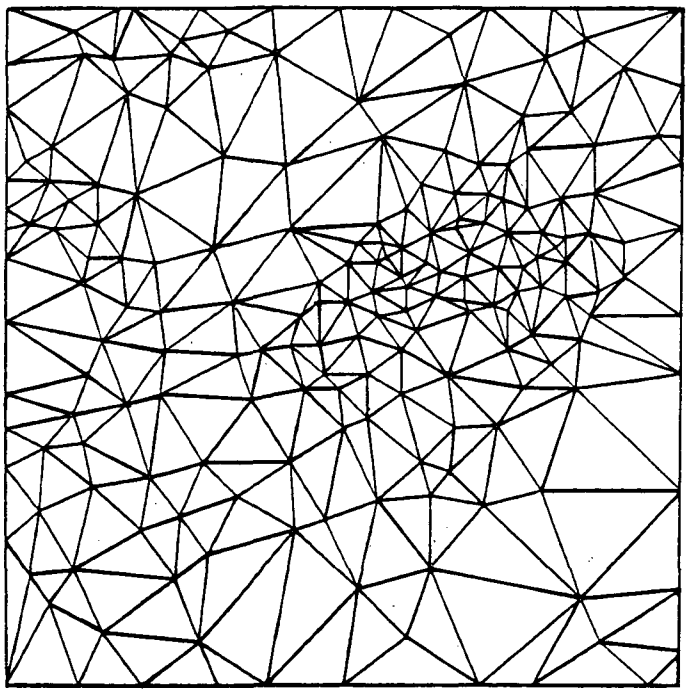


FIGURE 8. A TRIANGULAR STRUCTURE

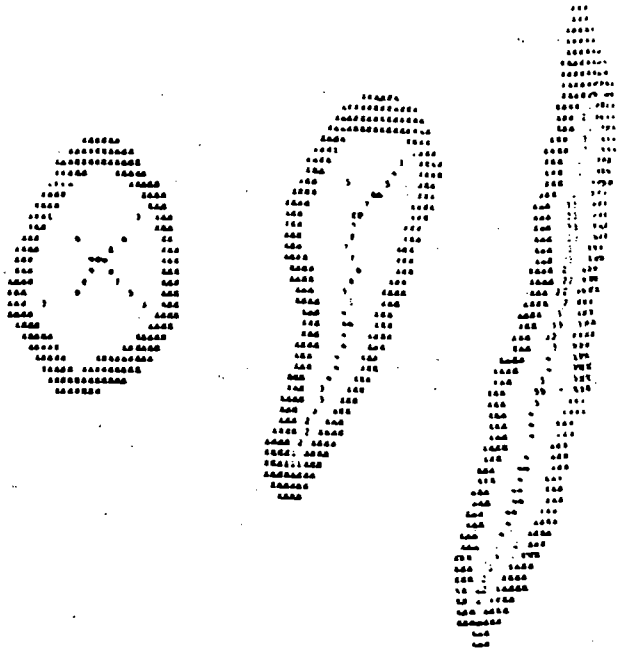


FIGURE 9. REGIONS AND THEIR SKELETONS (FROM PFALTZ AND ROSENFELD, 1967).

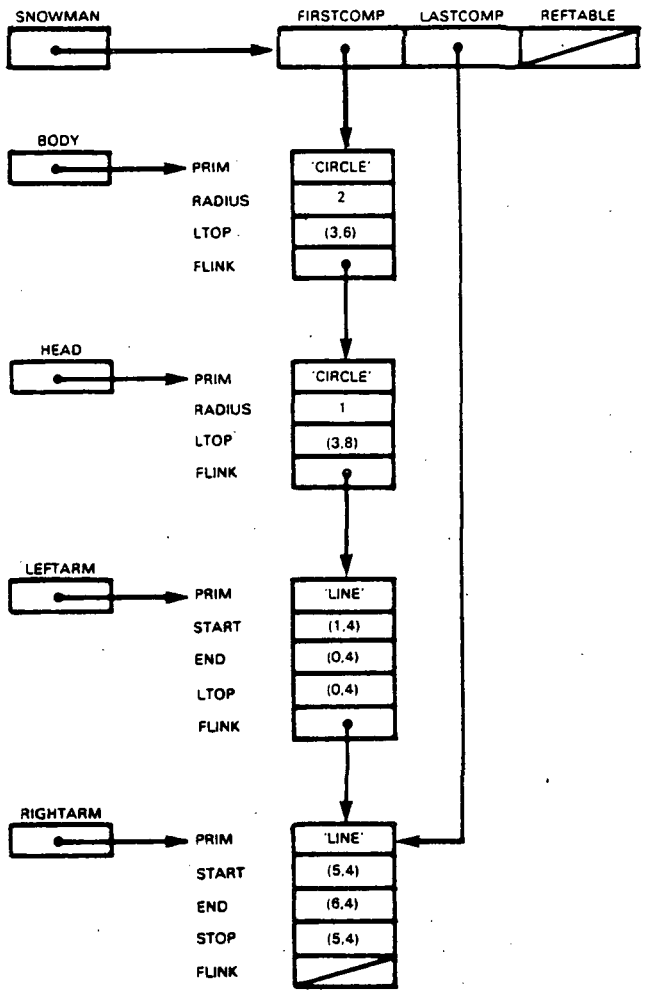


FIGURE 10. A LINGUISTIC STRUCTURE (FROM SHAPIRO, 1979).

This section outlines the data processing requirements of Earth science projects at the Goddard Space Flight Center, describes the basic tools of geographic data processing, and then presents specific ways that new geocoded data structures can better accomodate these tools to scientists' needs.

3.1 PROJECT REQUIREMENTS

Due to their varied and ambitious nature, Earth science investigations place unique demands on computer resources. The maps and images which are processed during these investigations contain enormous volumes of data, come from widely diverse sources, and are subjected to highly complex computer analyses. This combination of conditions exists in no other field of geographic data processing. The Earth science projects at the Goddard Space Flight Center contain, in varying degrees, elements of three basic themes: environmental and natural resources monitoring; physical process analysis and modeling; and sensor and image science. General descriptions of the data processing requirements of some representative projects are outlined below.

3.1.1 GIMMS

The Global Inventory Monitoring and Modeling Studies (GIMMS) project currently emphasizes environmental and natural resources monitoring. This project's purpose is to observe, measure, and interpret continental and global scale changes in vegetative landcover. The estimation of worldwide agricultural productivity and an increased understanding of desertification are some of the potential benefits of this investigation. The data used by GIMMS are characterized by relatively low spatial resolution (compared with Landsat), high temporal resolution, and large areal extent. The major source of these data is the Television and Infrared Observation Satellite (TIROS) Advanced Very High Resolution Radiometer (AVHRR) sensor. At present, the Goddard Space Flight Center receives (from the National Oceanographic and Atmospheric Administration) one four-kilometer resolution AVHRR data tape per day representing an 85 degree of latitude by 25 degree of longitude (at the Equator) swath of the continent of Africa. These swaths are remapped and composited by scientists at Goddard to form cloud-free vegetative index images of the entire Africa continent at the rate of two or three per month. Data from ground surveys, such as those conducted by the International Livestock Center of Africa in the Sahel region of Ethiopia and Niger, have been used to a limited extent to corroborate the remotely-sensed data. The most prevalent analysis performed on the AVHRR data is multispectral classification.

3.1.2 ISLSCP

The International Satellite Land Surface Climatology Project (ISLSCP) emphasizes the analysis and modeling of physical processes. Specifically, its goal is to understand those interactions between the Earth's soils system, biosphere, atmosphere, and hydrological cycle which influence climate over land surfaces. Taken in sum, these interactions affect the suitability of the land to support vegetation and human habitation. Studies will be conducted at several sites (such as the southern Great Plains region of the United States, the tropical rain forests of Brazil,

and the Sahel region in Senegal) which represent different climatically sensitive regimes. ISLSCP has two major subgoals. The first subgoal is a retrospective analysis of existing remotely-sensed data. For each study site, all relevant remotely-sensed data acquired since 1972 (when Landsat data first became available) will be analyzed to determine the extent to which climate-influencing changes in the land surface can be measured. The second subgoal is to establish the utility of current satellite-based information for understanding land/climate interactions on a continental or global scale. ISLSCP will process on the order of 10^{10} bytes of data. These data will come from a wide variety of sources. The sources of remotely-sensed data include the Landsat Multispectral Scanner (MSS) and Thematic Mapper (TM), the TIROS AVHRR, the Geostationary Operational Environmental Satellite (GOES) Visible Infrared Spin Scan Radiometer (VISSR), the Heat Capacity Mapping Mission (HCMM), and the Nimbus Scanning Multichannel Microwave Radiometer (SMMR). The sources of ancillary data include World Monthly Surface Station Climatology readings and manually collected "ground truth" measurements. The remotely-sensed imagery will be radiometrically corrected, registered (along with reformatted ancillary data) to a common spatial reference frame, and placed in a multi-layer composite data set for analysis. The manipulations performed on these data will include statistical cross-correlations, analysis of variance, principal components analysis, image classifications, computations of Earth science parameters (such as vegetative index and antecedent precipitation index), and special purpose modeling functions.

3.1.3 SIR-B

The Shuttle Imaging Radar-B (SIR-B) mission, scheduled to be flown in Fall 1984 on the seventeenth flight of the space shuttle, is an example of a project which emphasizes sensor and image science. SIR-B will be the first spaceborne imaging radar providing control of incidence angle. Further, it will provide digitally processed images at these selected incidence angles. This imagery, when subjected to newly-developed analyses (including mathematical models of radar backscatter), will greatly enhance traditional photointerpretive methods. Specifically, it will demonstrate expanded utility for the assessment of topographic relief, surface roughness, geomorphology, and land cover classification. One of the SIR-B investigations at the Goddard Space Flight Center will attempt, through the use of new stereo image processing techniques being developed on the Massively Parallel Processor, to derive terrain elevation information from overlapping radar images. If successfully produced, these elevation data will be used to further enhance the information content of Landsat Thematic Mapper data. Another SIR-B investigation at Goddard will assess the vegetation penetration characteristics of spaceborne imaging radar and then use SIR-B imagery for landcover mapping in tropical ecosystems. This investigation will require the development of mathematical models which relate vegetation canopy characteristics and surface geometry to the radar backscatter detected at different reflection angles. It will also require the development of new techniques for merging radar imagery with ground-truth data (and data from other remote sensors) in such a manner that distinct local sites occurring within the same radar image can be analyzed separately.

3.2 PROCESSING TOOLS

Geographic information systems, image processing systems, data base management systems, statistical packages, and high speed processors all contribute to the handling of geographic data. The basic roles of these systems are described below.

Geographic information systems: (1) provide for the conversion of analog maps and spatially-referenced tabular data into digital form; (2) register multiple multi-thematic map and image data planes to a common geographic reference frame; (3) perform spatial analysis and modeling using these data, including relabelling, overlay, distance, and neighborhood functions (Tomlin and Berry, 1979); and (4) display images and maps on a variety of graphics devices.

Image processing systems: perform general analytic operations on remotely-sensed images, such as geometric and radiometric corrections, coordinate conversions, classifications, principal components analysis, and enhancements.

Data base management systems: (1) perform recordkeeping functions, such as the cataloging and inventorying of information about maps and images (This is done through the creation of unified collections of interrelated data called "data bases". Data base management systems permit users to share the updating and querying of these data bases.); (2) permit retrieval of data (through generally not image-sized datasets); and (3) can include analytic processing and graphic display capabilities.

Statistical packages: perform standard statistical operations on tabular data (e.g., computation of mean, standard deviation, etc.; correlation and regression; hypothesis testing; analysis of variance; etc.)

High-speed processors: perform certain classes of operations (e.g., image convolutions and classifications) very quickly, and are often controlled by a "front-end" processor (a more or less standard computer which regulates data flow to and from the high-speed processor and which initiates process execution).

3.3 SPECIFIC APPLICATIONS OF NEW STRUCTURES

Existing geographic data processing systems can be better adapted to the needs of Earth scientists through the integration of new geocoded data structures, especially hybrid ones. Hybrid structures were expressly invented to deal with multisource data characteristics. In addition, they can be tailored to match specific storage and algorithmic requirements. Thus, they are the best structures for handling the simultaneous constraints of diverse data origin, large data volume, and complex computation which are imposed by Earth science investigations. Applications of three hybrid data structures are presented below.

3.3.1 TOPOLOGICAL GRID STRUCTURE FOR REGION-BASED MODELING

The automated analysis of land surface features is of increasing importance to Earth scientists. Through the use of topological grid structure, user-friendly algorithms for studing these features can be integrated into geographic information systems.

Many Earth science investigations now focus on selected land surface features such as fields, lakes, and forests. For example, the SIR-B study of tropical ecosystems being conducted at the Goddard Space Flight Center will evaluate radar backscatter from particular forest stands. Similarly, the ISLSCP project will compare multiple sensor interpretations of specific cultivated fields, desert areas, and rain forest sites. Studies of feature-type information were common when most geographic data were cartographic in nature and represented in linked structures. But now, classified remotely-sensed images have become a major source of land data. "Regions", which are groups of contiguous identically-classified pixels in these images, correspond to land surface features. Unfortunately, because classified images are represented in standard grid structure, they do not contain explicit regional information. For example, consider a typical landcover image. Each pixel of the image has been assigned a numerical value which tells whether it corresponds to "water", "bare soil", "forested land", or other landcover type. But, the digital representation of the image contains no explicit information which indicates a pixel's membership in a particular waterbody, barren field, or forest. If classified images are converted to topological grid structure, however, this type of information becomes available (see section 2.2.3) and "region-based" algorithms for studying features, which previously had been extremely awkward to implement and were rarely attempted, are easily designed.

Region-based modeling algorithms employing topological grid structure are both transportable and user-friendly. They are transportable because topological grid structure is almost identical in form to the standard multi-layer gridded image format supported by most current geographic information systems. They are user-friendly because the details of manipulating and accounting for groups of contiguous pixels are transparent to the scientific investigator. Five region-based algorithms are being developed as part of a geographic information system residing on the Land Analysis System at the Goddard Space Flight Center. The first, BIGTOP, converts classified images into topological grid structure. The four remaining algorithms (TOPSIZE, VENN, FINDNEAR, and REGADJ) act upon topologically grid-structured images to perform region-based versions of the four basic geographic information system modeling operations: relabelling, overlaying, distance searching, and neighborhood scanning. TOPSIZE performs relabelling based on size. An example of its use would be to label all waterbodies below a certain size as ponds and all waterbodies above that size as lakes. VENN performs feature-by-feature change detection. It might be used to measure the decrease in area of a particular forest stand due to logging operations. FINDNEAR computes regional proximity. For example, it can identify all barren fields within a specified distance of waterbodies (for soil run-off studies). REGADJ characterizes regional neighborhoods. It might be used to single out cultivated fields which are adjacent to snow-capped mountains as potential flood hazard areas. (Goldberg, 1984).

3.3.2 VASTER STRUCTURE FOR STORAGE-EFFICIENT AND VERSATILE CARTOGRAPHIC DATA SETS

Information from maps is essential for establishing the spatial accuracy and thematic content of remotely-sensed images. Vaster structure provides a means for digitally representing very large cartographic data sets in a form that is storage-efficient yet easy to correlate with satellite data.

Nearly all Earth science investigations depend upon cartographic information. Multisource satellite images are registered to a common geographic reference frame using the exact latitude and longitude of ground control points supplied by survey maps. Similarly, thematic maps provide the "ground truth training sites" needed for automatic image classification. And, of course, thematic maps themselves are an important input into physical process models. But, because maps are generally converted from their original vector representation into raster form before they are used (due to the overwhelming predominance of raster-formatted satellite data), much of their versatility is lost. For example, information on exact polygon boundaries and region adjacency is no longer available. Also, raster structure is especially wasteful of computer storage space when used to represent large, sparse line maps. Vaster structure is the first attempt to solve both of these problems simultaneously. It combines the storage efficiency and information content of vector data with the convenience of raster data (see Section 2.2.3).

The implementation of vaster structure can improve the functioning of several types of geographic data processing systems. Because of its storage efficiency, vaster structure is a good candidate for representing spatial datasets in data base management systems, which can currently only handle entity records of restricted size. Because of its ability to associate the exact geographic coordinates of point and polygon data with particular grid cells, vaster structure could serve as the organizing principle for control point libraries in image processing systems. Vaster structure can also be considered the ideal archiving mechanism for geographic information systems, which often need to perform spatial modeling operation in both the vector and raster domains.

Map information is converted into vaster form by being "frozen" part way between vector to raster conversion. Details of this process can be obtained from the initial report (Peuquet, 1983). It should also be noted that raster to vaster conversions are also possible: these operations are analogous to conversions from standard grid structure into topological grid structure. The Goddard Space Flight Center is supporting the development of prototype software to build and manipulate vaster-formatted data sets through a grant with Professor Donna J. Peuquet of the University of California at Santa Barbara.

3.3.3 TREE STRUCTURES FOR MULTIPLE-RESOLUTION IMAGE DATA SETS

The creation of multilayer image data sets is a crucial component of Earth science investigations. The implementation of quadtree structure (and its logical extensions) can eliminate the requirement that images in these datasets be represented at a common spatial resolution.

In order to compare images of the Earth that were acquired at different times, to evaluate the relative utility of different satellite sensors, or to perform spatial modeling, scientists build multilayer composite data sets. For example, the GIMMS project is building multi-temporal data banks of AVHRR images. The ISLSCP and SIR-B projects will be merging data from a variety of sensors with differing spectral, spatial, and temporal characteristics. A factor which affects the scientific validity of all of these projects, however, is the artificial constraint (imposed by geographic information systems) that images be converted to the same spatial resolution (as well as areal extent, geographic projection, etc.) if they are to be overlayed. This means that once a particular resampling algorithm has been chosen, the resultant data layers forever reflect the bias of the scientist that created them. Tree structures, by representing areas at several spatial resolutions simultaneously (see Section 2.2.3), provide a way to preserve the original information. Each level of the tree can correspond to a layer in the data set and can represent the image with the corresponding resolution. For cases where the spatial resolutions of source data are in ratios of even powers of two (such as for 1 km AVHRR Local Area Coverage and 4 km AVHRR Global Area Coverage), quadtree structure may be used. For other cases, tree structures of different (or even varying) cardinality can be devised. The suitability of these structures for parallel computation is an added advantage. The practical application of the above concepts will be the subject of future research.

4. CONCLUSIONS

This report has reviewed the methods for digitally representing spatial data and has shown how the implementation of new geocoded data structures can better accomodate existing computing systems to Earth scientists' needs. At the Goddard Space Flight Center, two practical goals related to geographic data processing still remain. First, scientists must be provided with a comprehensive, consistent, and efficient means for registering multisource map and image data planes to a common geographic reference frame. Second, improved methods for managing very large image data archives need to be developed.

REFERENCES

1. Amidon, E.L. and G.S. Aiken. "Algorithmic Selection of the Best Method for Compressing Map Data Strings." Communications of the ACM. 14,12 (December 1971), 769-774.

2. Burton, W. "Representation of Many-Sided Polygons and Polygonal Lines for Rapid Processing." Communications of the ACM. 20,3 (March 1977), 166-171.

3. Freeman, H. "Computer Processing of Line Drawing Images." Computing Surveys. 6,1 (March 1974), 57-97.

4. Gold, C.M. "The Practical Generation and Use of Geographic Triangular Element Data Structures." Proceedings of the First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems. Volume 5 (1978), Gold/1-18.

5. Goldberg, M. "Region-Based Modeling Algorithms for Remotely-Sensed Data." Proceedings of the Tenth International Symposium on the Machine Processing of Remotely-Sensed Data. (1984), 205-208.

6. Hunter, G.M. and K. Steiglitz. "Operations on Images Using Quadtrees." IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. PAMI-1 (April 1979), 145-153.

7. Merrill, R.D. "Representation of Contours and Regions for Efficient Computer Search." Communications of the ACM. 16,2 (February 1973), 69-82.

8. Nagy, G. and S.G. Wagle. "Geographic Data Processing." Computing Surveys. 11,2 (June 1979), 139-181.

9. Peuker, T.K. and N. Chrisman. "Cartographic Data Structures." The American Cartographer. 2,1 (April 1975), 55-69.

10. Peuquet, D.J. "A Hybrid Structure for the Storage and Manipulation of Very Large Spatial Data Sets." Computer Vision, Graphics, and Image Processing. 24,1 (October 1983), 14-27.

11. Pfaltz, J.L. and A. Rosenfeld. "Computer Representation of Plane Regions by their Skeletons." Communications of the ACM. 10,2 (February 1967), 119-122.

12. Shapiro, L.G. "Data Structures for Picture Processing: A Survey." Computer Graphics and Image Processing. 11,2 (October 1979), 162-184.

13. Thomas, A.L. "Data Structures for Modelling Polygonal and Polyhedral Objects." Proceedings of the First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems. Volume 5 (1978), Thomas/1-42.

14. Tomlin, C.D. and J.K. Berry. "A Mathematical Structure for Cartographic Modeling in Environmental Analysis." Proceedings of the American Congress on Surveying and Mapping, 39th Annual Meeting. (March 1979), 269-284.

BIBLIOGRAPHIC DATA SHEET

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
TM 86113		

4. Title and Subtitle	5. Report Date
Geocoded Data Structures and Their Applications to Earth Science Investigations	June 1984
	6. Performing Organization Code: 630

7. Author(s)	8. Performing Organization Report No.
Michael Goldberg	

9. Performing Organization Name and Address	10. Work Unit No.
Space Data and Computing Division/Code 630 NASA/Goddard Space Flight Center Greenbelt, MD 20771	11. Contract or Grant No.
	13. Type of Report and Period Covered: Technical Memorandum
12. Sponsoring Agency Name and Address	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

A geocoded data structure is a means for digitally representing a geographically referenced map or image. This document reviews the characteristics of representative cellular, linked, and hybrid geocoded data structures. It then outlines the data processing requirements of Earth science projects at the Goddard Space Flight Center, describes the basic tools of geographic data processing, and presents specific ways that new geocoded data structures can be used to adapt these tools to scientists' needs by expanding analysis and modeling capabilities, simplifying the merging of data sets from diverse sources, and saving computer storage space.

17. Key Words (Selected by Author(s))	18. Distribution Statement
Data Structures, Geographic Information Systems, Data Management, Earth Sciences	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
	Unclassified		

*For sale by the National Technical Information Service, Springfield, Virginia

GSFC 25-44 (10/77)