TILING LARGE GEOGRAPHICAL DATABASES

Michael F. Goodchild

National Center for Geographic Information and Analysis
University of California
Santa Barbara, CA 93106, USA

ABSTRACT

Geographical variation is infinitely complex, so the information coded in a spatial database can only approximate reality. The information will always be inadequate, in spatial resolution, thematic or geographical coverage. "Large" can be usefully defined as exceeding our current capacity to deliver. We provide examples of large geographical databases. Traditional stores partition geographical data by theme and geographically. We assume that digital geographical databases will be largely archival, and will be similarly partitioned. A general model of a large archival store is presented. We analyze the properties of a generalized Morton key as a means of indexing tiles, and illustrate its role in traditional systems of tile indexing. For global databases, we propose a tiling based on recursive subdivisions of the triangular faces of an octahedron using a rule of four.

1. INTRODUCTION

The theme of this conference on Large Spatial Databases raises a host of complex issues. The term "spatial" includes both two and three dimensions, and covers applications as diverse as medical imaging, CAD and remote sensing. The purpose of this paper is to provide a geographical perspective, and to enquire into the handling of large geographical databases, while noting that the geographical case is a somewhat limited subset of the larger set of spatial database problems. Although geographical data are complex, we propose in this paper to make and justify certain assumptions which may in fact simplify the issues. However it is clear that these assumptions do not apply to the more general spatial case.

The first part of the paper discusses the meaning of "large" in the context of geographical data, and provides a number of examples of current and planned databases. The next section looks at traditional methods which have evolved for handling large geographical datasets in the absence of digital technology. Section Four presents a simple model of access within a large geographical data store, and analyzes the model using a generalization of the Morton key. We illustrate the consistency of the model with traditional methods of map indexing using a simple example. Finally, Section Six looks at the extension of the technique to the global case of data for the sphere or spheroid.

2. THE NATURE OF GEOGRAPHICAL DATA

The spatial variation of many properties of the earth's surface is continuous and almost infinitely complex. The simple question "How long is a shoreline?" raises difficult issues as it is clear that the length of a simple cartographic object like a shoreline depends on the spatial resolution of the measuring device, or the scale of the map. Mandelbrot (1965) showed that the increase in length of a shoreline with increasing scale is often systematic, and continues without limit at least until the definition of shoreline breaks down. In general, the process by which increasing spatial resolution reveals more detail continues without limit for most forms of geographical data. Geographical surfaces and curves lack well-defined derivatives in many cases, and are better analyzed within the framework of fractals than with continuous functions.

There are some exceptions to this generalization, particularly when geographical objects have mathematical definitions. For example there is no difficulty in defining the length of a surveyed parcel boundary, or the area enclosed. Other cases are more marginal: the attribute "in the World Trade Center" is well-defined if the relevant object is taken to be the surveyed parcel boundary on which the World Trade Center is built, but not if it is taken to be the building itself. However many objects in spatial databases are generalizations of continuously varying fields (e.g. shoreline, contour, woodlot), and the precise form of the object is therefore sensitive to the approximation inherent in the generalization.

The relationship between a geographical database and the reality which it approximates is complex. Unfortunately the user of such databases is concerned with learning about reality, and often sees the approximation inherent in the database as an unwanted and often poorly understood filter. In a typical application of a geographical database, such as a resource management study, it is almost certain that the user will find the database inadequate in some respect - in its level of spatial detail, currency, geographical coverage, or in the range of geographical variables available. In essence, users have expectations which will almost always exceed the contents of the database, and an effectively infinite appetite for knowledge about geographical reality. With this assumption, we define "large" as beyond the current capacity of systems to deliver with reasonable efficiency. In this sense, large geographical databases provide an unsolved, but well-motivated set of problems.

2.1 Examples

The global context provides the best-known set of large geographical database problems. Current remote sensing instrument systems, such as Landsat and SPOT, generate data at the rate of 10^7 bytes/scene for 10^3 scenes every few days, producing around 10^9 bytes of data per day. The systems being planned for the 1990s, such as EOS, have much higher data rates to accommodate higher spatial and spectral resolution, and figures of 10^{12} bytes/day are anticipated.

At the national level, we find similar data volumes in the systems being developed to build and maintain digital versions of national topographic map series. The typical topographic map sheet requires on the order of 25 Mbytes of digital storage, so a full digital cartographic database of the order of 10^5 sheets covering the US will require on the order of 10^{12} bytes. Similar capacity will be needed to store the digitized versions of the 250,000 sheets of the UK 1:1250 base mapping series. At the county level, the parcel database currently being built by the County of Los Angeles is expected to reach 300 Gbytes. The fact that such large data volumes are anticipated even at the county level seems to reinforce the point made earlier: reduced geographical coverage is offset by increased geographical resolution, so that expectations are comparable at all levels.

Demands for very large volumes of geographical data are driven by a need for high levels of spatial resolution, which is costly to deliver: a doubling of spatial resolution produces at least a fourfold increase in data volume in many cases. Yet high spatial resolution is inescapable if analysis is to be useful. For example, effective forest management requires a resolution sufficient to see individual trees if harvest estimates are to be reasonably accurate; land ownership records must be as accurate as the survey data on which they were based; many military applications require resolutions of better than 1m; and management of elections in Los Angeles County requires resolution finer than the parcel level because of complex and detailed precinct definitions.

3. TRADITIONAL METHODS

Geographical information has been available in map form for a long time, and in many ways the digital problem of handling large amounts of geographical data is a simple extension of more traditional issues. The methods used to compile, reproduce, distribute, store and access maps have evolved through time as a compromise of conflicting objectives under a set of technological constraints. Although the constraints are different, it may nevertheless be useful to begin by looking at traditional methods.

Geographical information is traditionally partitioned thematically across different map series. The topographic map groups together elevation, roads, railways and a variety of cultural features. A soil map is created by overprinting a single variable, represented by a tesselation of irregular area objects with associated soil class attributes, on a base map. Since the topographic map already exploits the limits of the technology fully by showing as many themes as possible, the base map is usually generated by dropping one or more themes from a topographic map, often the contours or a green "wooded" tint. Land use maps are similarly produced by adding a single variable to a base map. Information must also be dropped if the scale is reduced, because of the limitations imposed by the technology on the size of lettering and density of information. Thus to make a road map by reducing the scale of a base map it is also necessary to drop certain themes or classes of features.

Geographical information is also partitioned geographically into map sheets or "tiles". Printing, distribution and storage all place technological limits on the sizes of map sheets, and these constraints have led to the familiar series of rectangular, equal-size sheets. Tilings at different scales often nest within eachother, which is another advantage of choosing a rectangular system of tiles.

This system of thematic and geographical partitioning seems to have arisen as a result of at least four issues: the next four subsections discuss these in turn.

3.1 The storage volume

The individual volume in the traditional map store is the map sheet. The physical size of map sheets is limited to approximately 2m by 1m by the printing process, handling and storage: the user would also find larger sheets difficult to handle. Cartography imposes a resolution limit of about 0.5mm for several reasons: pen construction, human vision and instability of the base medium, among others.

3.2 Storage devices

Consider the map library as an example of a traditional geographical data store. We consider the time required for the user to move from one object to another in the store. Clearly it is least when the two objects are shown on the same map sheet or volume, higher when the objects are on different sheets within the same drawer, and highest when the sheets are in different drawers.

3.3 Use patterns

Queries in a map library database often require access to more than one volume, within the same or different themes, and the grouping of themes onto volumes reflects these complex patterns of use. For example, roads are shown on soil maps because queries frequently involve both themes. Queries often involve access to adjacent map sheets, so it is common to group sheets according to some geographical hierarchy. The fact that most map libraries sort maps first by theme, and then geographically, suggests that transitions between adjacent sheets of the same theme are more common than transitions between different themes for the same tile or geographical area. Thus we can interpret the organization of data in a traditional store as the outcome of an informal analysis of transition probabilities between tiles and themes.

3.4 Data collection

The traditional system of data collection and map compilation divides responsibilities between agencies on thematic lines: in the US, for example, the Geological Survey produces the topographic series, while the Bureau of the Census produces much of the demographic mapping. This reflects the importance of interpretation in the mapping process, and the role played by discipline specialists, for example soil scientists, in the production of soil maps. It would be much less efficient to organize data collection geographically, by assigning agencies responsibility for all themes in one area.

4. A MODEL OF TILING

In this section we present a simple model of tiling for large geographical databases. We assume that data will continue to be partitioned by theme and geographically, based on the previous discussion. First, division of responsibility between data collection agencies will continue to lead to thematic partitioning. Second, geographical partitioning is administratively convenient in a system of varying currency and regular updating. Updating requires long transactions and is best handled by checking out partitions. Third, partitioning is essential as long as expectations over volume exceed capacity, and as long as fast access is desirable. However partitioning may be at least partially hidden from the user, who may value the ability to browse the database in an apparently seamless fashion.

We assume that geographical partitioning will be rectangular, as modified if necessary to take account of the earth's curvature in specific projections. We further assume that themes will therefore be grouped within tiles, and that transitions between groups of themes will be rare, as in the traditional organization.

We further assume that the predominant role of geographical data stores will be archival. Geographical data is inherently static, as evidenced by the long update cycles of most map series. Databases which provide access to administrative records are often transactional, but their geographical content is not - customer accounts change frequently but their locations do not. However we should clarify that archival as used here does not necessarily mean static, but refers to the comparative absence of updates. Landsat scenes are recollected every 19 days, which makes them much less static than topographic maps, but it is rare for transactions with a specific Landsat scene to include updates. WORM devices are eminently suitable for geographical data because of the overwhelming dominance of read access over write. Where update is required, we assume that the user will check out a section of the database for intensive transactions.

The model which follows is intended to apply to a variety of storage media, including optical WORM, optical jukeboxes, cassette tape, and other devices in the terabyte range, as well as the traditional map library store.

4.1 Notation

Let the generalized device be positioned at tile i, and assume that we wish to access tile j. The probability of this transition from i to j will be written as p_{ji}. The "cost" or delay in accessing tile j from i is c_{ij}.

To organize the tiles on the device in the most efficient manner we should minimize the expected cost. The expected delay in accessing j from i is given by:

$$C \quad = \quad E(c_{ij}) \quad = \quad \sum_{i} p_i \sum_{j} c_{ij} \, p_{ji} \qquad (1)$$

where p_i is the marginal probability that the previous access was to tile i.

Unfortunately almost nothing is known about the marginal or transition probabilities for tile access. They clearly depend heavily on the nature of the application: the oil and gas industry, for example, would have very different patterns of map use from recreational hikers. In some fields surrogate information might be useful. For example, where street network databases have been used to support vehicle navigation systems the transition probabilities between tiles have been determined at least in part by the presence of major traffic arteries.

In the temporary absence of better information we make simplifying assumptions. We assume that the p_i are constant, and that the p_{ij} are equal to 1/4 when j is a 4-neighbor of i, and 0 otherwise. We thus restrict transitions to the set of four adjacent tiles, and assume equiprobability. The optimization model (1) has the general form of a quadratic assignment problem (Koopmans and Beckmann, 1957).

Let x_i represent the location of tile i on the device. If tiles are of equal size and stored sequentially then x_i will be proportional to the ranked position of the tile, which we can equate to a key. Note that this assumption will be reasonably robust since transitions will be evaluated only between 4-neighbors. Thus in the first instance we look at the case where c_{ij} is proportional to the absolute difference in positions:

$$c_{ij} = k\,|x_j - x_i| \qquad (2)$$

Thus our problem is to minimize the expected difference in keys between a tile and its 4-neighbors.

Different devices have different versions of (2). Our work on optical WORMs has shown that seek time c_{ij} is effectively zero within volume, in relation to transfer time, but is directly proportional to difference in position between volumes in a jukebox device. Tape has highly asymmetrical characteristics within volume, where (2) holds for $j>i$, but for $i>j$ c_{ij} is approximated by $k_1x_i+k_2x_j$: k_1 is the rate of rewind and k_2 the rate of forward motion of the tape device.

4.2 Generalized digit-interleaved keys

Goodchild and Grandfield (1983) computed the expected difference between neighboring keys in certain orderings of rectangular tiles. For an n by n square matrix, the expected difference is $(n+1)/2$ for the Morton order (note that Morton is well-defined only for $n=2^m$ where m is integer), and also for row by row order, and a modification of row by row order with every other row reversed. This last order, which Goodchild and Grandfield termed "row-prime" and is also known as "boustrophedon", allows every pair of tiles which are adjacent in the key to be adjacent in space. They also analyzed the expected squared difference between neighboring keys, and more recently Mark (1989) has looked at this statistic over a large number of alternative orderings of a 2^m by 2^m matrix.

The Morton key can be generated by numbering rows and columns from 0 to n-1, and interleaving the bits of the binary representations of the row and column numbers. For example, row 2 (10) column 3 (11) has Morton key 13 (1101). Now consider a generalization of this concept to digit interleaving. Suppose row and column numbers are represented to any collection of bases, and interleaved. Morton is a digit-interleaved key, as is row order (row and column as single digits to base n). Row-prime can be generated by applying an operator which complements the column digit if the preceding row digit is even.

Theorem: *For any digit-interleaved key of an n by n matrix, the expected difference of 4-neighbor keys is (n+1)/2.*

Proof: See Goodchild and Yang (1989a), who obtain general expressions for expected differences in rectangular arrays.

Although there have been frequent references in the literature to the efficiency of the Morton key at preserving adjacencies between tiles, it appears that no digit-interleaved key has advantages over any other, at least under the assumptions made here.

5. TRADITIONAL MAP INDEXES

In this section we compare the concept of generalized digit interleaving with the structure of traditional map indexes. The GEOLOC grid (Whitson and Sety, 1987) is typical of many such systems in its use of rectangular tiles at each level of a hierarchical structure, and the nesting which it imposes between levels. We summarize the system below:

Level 0: The continental US is divided into 2 rows and 3 columns of tiles, each 25 degrees of longitude by 13 degrees of latitude, numbered row by row from the top left using the digits 1 to 6.

Level 1: Each level 0 tile is divided into 25 columns of 1 degree each and 26 rows of 0.5 degrees each. The rows are lettered from the left from A to Y, the columns from the top from A to Z. These tiles are the 1:100,000 quadrangles used by the USGS topographic series.

Level 2: Each level 1 tile is divided into 8 columns and 4 rows, to form 32 tiles numbered row by row from the top left from 1 to 32. Each tile is 7.5 minutes square, and corresponds to the 1:24,000 quadrangles of the USGS topographic series.

Level 3: Each level 2 tile is divided into 2 columns and 4 rows, lettered A through H row by row from the top left.

Level 4: Each level 3 tile is divided into 5 rows, lettered A through E, and 10 columns, numbered 0 through 9.

The row and column references of the system use 5 bases each, {2,26,4,4,5} and

{3,25,8,2,10} respectively. Interleaving produces a 10-digit reference with bases {2,3,26,25,4,8,4,2,5,10}. However the system combines row and column bases in some cases, so the full reference contains only 7 digits: {6,26,25,32,8,5,10}. Of these, the first and fourth are expressed using decimal notation starting with 1, the last using decimal notation starting with 0, and the remainder using the alphabet. Note that to avoid ambiguity, the fourth digit which uses decimal notation for a base greater than 10 (the only case where the base of the digit exceeds the base of the notation) is positioned between two alphabetic digits. At level 4, the GEOLOC reference is unique to an area about 200m square on the earth's surface. For example the reference 4FG19DC6 identifies an area just north of Los Angeles.

As an example of digit interleaving, the GEOLOC system is relatively complex, due to the desirability of nesting the system within the 1:100,000 and 1:24,000 quadrangles, and the unequal sizes of latitude and longitude divisions.

6. TILING THE GLOBE

Rectangular tilings are difficult to apply at global scales without significant distortion. For example if a rectangular tiling is superimposed on a simple cylindrical projection such as the Mercator or Cylindrical Equidistant, three problems arise: the tiles are unequal in area and shape, adjacencies are missing at the 180th meridian, and adjacencies at the poles are severely distorted. Analyses based on such tilings may be severely biassed. Instead, it is desirable to find a tiling which is hierarchically nested, preserves adjacencies, and in which tiles are roughly equal in area and shape at a given level. It is clearly impossible to satisfy all of these requirements perfectly and simultaneously, so the optimum tiling must be some form of compromise, as in map projections themselves.

Dutton (1984; 1988; 1989) has proposed a tiling called Quaternary Triangular Mesh or QTM. The sphere or spheroid is first projected onto an octahedron aligned so that two vertices coincide with the poles, and the remaining four occur on the equator at longitudes 0, 180, 90 West and 90 East. Each triangular facet is then recursively subdivided by connecting the midpoints of its sides. Goodchild and Yang (1989b) have implemented this scheme and provide algorithms for conversion between latitude and longitude and tile address at any level. A tile address at level *k* consists of a base 8 digit followed by *k* base 4 digits. At each level of Goodchild and Yang's scheme the central triangle is assigned digit 0, the triangle vertically above (or below) is assigned 1, and the triangles diagonally below (above) to the left and right are 2 and 3 respectively. The system provides a single address in place of latitude and longitude, and the length of the address determines its precision, or the size of the tile. For example, the US has address {null} because it spans more than one level 0 tile, while the block bounded by 3rd and 4th Streets, Fulton and Broadway in the city of Troy, New York has the level 13 address {10223022113013} (the Broadway and Fulton faces are entirely within level 16 tiles, but the block as a whole spans two level 14 tiles).

7. CONCLUSIONS

Because much geographical data is comparatively stable over time, we have argued that the normal case is archival. While short transactions may predominate in many administrative records, the basis of their geographical access is stable and is likely to be updated by long transactions carried out after checkout. The problems of large geographical databases are therefore issues of data gathering, compilation, assembly in archives, and distribution. We assume that analysis and query will be carried out on extracts which are moved from the archival store in blocks or tiles.

We have argued in the paper that much can be learned about the optimal ways of organizing large geographical archives by looking at traditional methods of geographical data handling. Traditional cartography partitions databases along thematic and geographical lines, grouping themes in ways which have been found to satisfy the user community. It uses rectangular tiles and concepts of digit interleaving to index map sheets, in order to provide hierarchical nesting and to ensure rapid access to adjacent tiles. However, we have seen that under certain assumptions the entire class of digit interleaved indexes, which include both Morton and row order, yields the same expected access time. Further research is clearly needed to determine the robustness of this result under different sets of assumptions.

Many of the assumptions made in this paper concern the ways in which users access large geographical databases. In order to optimize the organization of the database, we need to know much more about applications, about the types of access required to satisfy queries and support analysis, and about the geographical distribution of accesses in specific applications. The information available on these questions at this time is virtually nil.

In general, we have suggested that the best strategy for organizing large geographical databases at this time may be to emulate traditional practice, which has arisen as an optimum response to traditional constraints, and to consider carefully the impacts of new sets of technological constraints on traditional methods. Thus far, digital technology has been adopted with little change in traditional practice in such areas as data collection, map compilation, cartographic design, map distribution, thematic grouping and geographical partitioning. In the long run, the new technologies available in each of these areas may lead to radical changes in methodology, but such changes should be incremental; in the absence of evidence to the contrary, it seems better to assume that traditional methods are the best guide.

8. ACKNOWLEDGMENTS

The research on which this paper is based has been supported by the US Geological Survey, Digital Equipment Corporation and the National Science Foundation, Grant SES 88-10917.

9. REFERENCES

Dutton, G., 1984. Geodesic modeling of planetary relief. Cartographica 21:188-207.

Dutton, G., 1988. Computational aspects of quaternary triangular meshes. Unpublished.

Dutton, G., 1989. Modeling locational uncertainty via hierarchical tesselation.` In M.F. Goodchild and S. Gopal, editors, The Accuracy of Spatial Databases. Taylor and Francis, Basingstoke.

Goodchild, M.F. and A.W. Grandfield, 1983. Optimizing raster storage: an examination of four alternatives. Proceedings, AutoCarto 6. Ottawa, 1:400-7.

Goodchild, M.F. and S. Yang, 1989a. On average distance in Morton and generalized digit interleave tiling. Unpublished.

Goodchild, M.F. and S. Yang, 1989b. A hierarchical spatial data structure for global geographic information systems. Technical Paper 89-5, National Center for Geographic Information and Analysis, University of California, Santa Barbara.

Koopmans, T.C. and M. Beckmann, 1957. Assignment problems and the location of economic activity. Econometrica 25:53-76.

Mandelbrot, B.B., 1965. How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156:636-8.

Mark, D.M., 1989. Neighbor-based properties of some orderings of two-dimensional space. Geographical Analysis (in press).

Whitson, J. and M. Sety, 1987. GEOLOC geographic location system. Fire Management Notes 46:30-2.

Extending a Database to Support the Handling of Environmental Measurement Data

Leonore Neugebauer *)
IPVR, University of Stuttgart

Abstract

This paper presents an approach to handle environmental measurement data within a relational database. The approach has been developed within a cooperative environmental research project exploring the contamination of groundwater and soil. Considering the special properties of measurement data we suggest an extension to data retrieval that allows the user to query measurement attributes without knowing the exact spatial and temporal measurement points. Further on, the interaction of this extension with the relational operations is outlined and a system design for an implementation is sketched.

1. Introduction

Environmental research has seen a tremendous increase in recent years, and new measuring devices have been installed at many places, delivering a large amount of data. So the demand for DBMSs to support environmental research in terms of administrating, analysing, and interpreting this data is growing rapidly.

The enhancements to overcome the shortcomings of commercial DBMSs include modifications of data modelling and data definition as well as support for new user-defined predicate functions during query evaluation that allow the user to compare (spatially or temporally) parallel measurement series even if the single recording points are not identical. The extensions introduced in this paper include the embedding of procedures or stored sequences of DML commands.

Various approaches to extend conventional DBMSs to overcome shortcomings in handling spatial and temporal data have been made and were included into new DBMS designs and implementations: The EXODUS system [CDFG86] offers an open architecture that may be adapted to the special needs of various applications. DBMSs that are designed to be used within non-standard applications are Starburst [SCLF86], POSTGRES [Sell87], PANDA [EgFr89] for spatial data, and the proposal of [AMKP85] for VLSI/CAD support. Further approaches are discussed by [CaDe85], [LiMP87], [StAH87], and [WSSH88]. Our approach mainly deals with extensions denoted as 'data extensions' by [LiMP87].

*) The work of the author is supported by the PWAB Project of Baden-Württemberg under grant no. PW 87 045