EXTRACTION OF AREA TOPOLOGY FROM LINE GEOMETRY

M. Visvalingam, P. Wade and G.H. Kirby
Cartographic Information Systems Research Group
The University of Hull, HULL HU6 7RX
England

ABSTRACT

Area entities are represented on a computer by geometric or locational information (which describe the course of their boundaries), topological information (which describe the areas to which boundaries belong) and object-related or geographic information (which describe the entities which map onto these areal units).

Most available systems for spatial data processing require from the outset object-related information, in the form of left/right or on-line references or area-seeds, for extracting the area topology.

This paper introduces the Disassociative Area Model (DAM) and contrasts it with other existing descriptions of areal entities. The primitive region (PR) forms the basic spatial unit for object modelling and acts as the link between the the geometric and geographic components. The derivation of spatial topology focuses on the boundary, which describes one extent of a PR. Geographic information may be input in a variety of ways and at any convenient stage using pragmatic models derived from DAM.

INTRODUCTION

One of the more challenging tasks in the development of Geographic Information Systems (GIS) is the derivation of an appropriate conceptual model for describing area entities. Verbal descriptions, internal representations via data structures, and data formats as perceived by the user of a GIS are adapted to match the requirements of different tasks, such as data capture, analysis, display and transfer. Both user and computer view areas as a set of polygons, describing uncut planar surfaces, when shading maps. In comparison, when displaying the linework of area boundaries, these outlines are perceived as boundaries between neighbouring areal entities, largely to avoid unnecessary replication of lines and associated problems of slivers.

In manual cartography, a user may trace the linework in any arbitrary way, since he can easily identify areal units and neighbourhood relationships even from spaghetti tracing. Most digital data models require topologically structured line data for automatic identification of polygons and their adjacency relationships.

This paper briefly reviews some of these models in the next section. It then presents the Disassociative Area Model (DAM) for describing area entities, the prime feature of which is that it separates the geography from the geometry. Some properties of this model are then explored and compared with those of others. An application of DAM highlights some further advantages of this model.

BACKGROUND

In the DIME system of the US Bureau of the Census, the basic element is the straight line, uncrossed by any other, called a segment. The segment is identified by its start and end points, called nodes, and the attribute codes for polygons on each side of it.

In the POLYVRT model for areas, the chain replaced the segment as the basic element. Like a DIME segment, a chain has two nodes at its end points, and is assumed to be uncrossed. It may, however, consist of many points. POLYVRT also keeps a list for each area of the chains which form its polygonal boundary. Although the data model remained similar to DIME, POLYVRT introduced the cartographic data structures which have continued to form the backbone of a variety of internal data structures used today.

Peucker and Chrisman (1975) reviewed DIME and POLYVRT and then described the GEOGRAF model. GEOGRAF introduced the Least Common Geographic Unit (LCGU) as another basic element and defined it as an area uncut by any other partitioning. The chain now became the boundary between two LCGUs and remained the unbroken unit of point retrieval. The boundary of each LCGU is constructed as a POLYVRT polygon directly from these chains. The identity or type of various features extending over this surface, e.g. districts, counties and countries in an administrative hierarchy, could then be associated with the LCGUs. Whereas POLYVRT represents such areas by hierarchical area codes, GEOGRAF can conceive of them as different data sets. The boundary between two objects of a set, e.g. between two counties, is described by a chain group, which is an ordered set of chains. The polygonal boundaries of these objects, in turn, consist of chain groups and will from now on be called GEOGRAF polygons.

The GIMMS segment format is an extension of POLYVRT. It allows composite codes for the extraction of GEOGRAF polygons given POLYVRT chains. However, GIMMS uses a unit line (POLYVRT) rather than a unit area model and is thus unable to combine the geometric and geographic descriptions in a flexible manner.

References to objects or entities to the left and right of chains allows these various models to cope with detached parts and holes (one or more uncut parts completely within another uncut part) without having explicit knowledge of their existence. Edwards et al (1977) proposed a hierarchical data structure (HDS) for representing areal data which has holes, holes in holes etc. There are two points to note with respect to HDS. Firstly, HDS utilised the concept of directed polygonal boundaries (DPBs) for analysing the area topology. The definition of DPBs does not invoke the concept of an enclosed polygonal area (p 3). Secondly, HDS is an extension of POLYVRT in that the DPBs ultimately refer to POLYVRT chains which refer to objects to their left and right. Since HDS is similar to our model in some respects, further description and discussion of HDS is postponed until later.

Although it expedites computer processing, the chain concept presents a poor human-computer interface. Left/right tags pose an unnecessary burden on users since this information can be derived by computer processing. The GIRAS structure (Mitchell et al, 1977, p 5) is described as topologically similar to that introduced by Peucker and Chrisman (1975). The input to GIRAS consists of arcs and polygon labels. Arcs, unlike chains, do not carry left/right tags at input time. A polygon label is an arbitrary point within each polygon with which is associated a not necessarily unique integer attribute. This suggests that GIRAS does not use the GEOGRAF model in its purity, but that it attempts to cope with detached parts of area objects directly. Further, by associating a composite feature code with this polygon label it is possible to encode a hierarchy of area entities.

GIRAS also explicitly recognises islands (GIRAS term for holes) and compound islands by clockwise ordering of arcs around the perimeter of a polygon and counterclockwise ordering of arcs around interior islands of the polygon. Thus like HDS, GIRAS uses DPBs. Polygon labels are used to fix the relationships between sets of boundaries describing complex equivalents of the LCGU (see Mitchell et al, 1977, p 11). The hierarchic relationship between holes within holes etc. remains implicit in GIRAS. GIRAS therefore uses a number of ad hoc procedures to circumvent problems in spatial data processing without seeking to accommodate them within an underlying conceptual model of areal entities.

The Level 3 Digital Line Graph (DLG) format, as the name implies, is line-oriented; only line elements contain explicit topological references (Allder and Elassal, 1984, p 7 & 8). Lines refer to the user identified unit areas on either side and only indirectly to area objects, which are encoded as attributes of unit areas. The unit area concept is thus used for extraction of boundaries of given area objects; it is not used for vertical integration of datasets. Line elements, which form the boundary between different categories of areas, are instead replicated in relevant datasets as 'coinciding' features (p 11).

The representation of locational data in ARC/INFO is said to be based on DLG (ESRI, 1985, p 2-9); descriptions of formats and some procedures correspond instead to GIRAS. ARC/INFO allows spaghetti digitising and vertical integration of datasets, called coverages, based on a variety of topological criteria. This adds procedures for pre-processing data into arc form prior to the building of polygons for new (input or derived) coverages. Illustrations suggest that it copes with islands but that it does not utilise DPBs.

The Working Group on Terms and Definitions of the American National Committee for Digital Cartographic Data Standards held that "holes in cartographic objects constitute a gap in our knowledge" (Moellering, 1984, p 24). HDS and GIRAS addressed this problem but both rely on the input of object-related information for extracting the relationships between boundaries. Only HDS makes explicit the hierarchy of geometric polygons.

DISASSOCIATIVE AREA MODEL (DAM)

This model disassociates the geometric and geographic components of areas with holes for separate academic consideration. This section briefly outlines the essential features of this model and then compares it with its precursors.

Geometry

Using only the geometry of the bounding lines, the areal map can be dissected into a set of uncut parts called primitive regions (PRs). At this stage, PRs exist only in concept; their representation hinges only on the boundary. Each boundary is a closed loop with direction, which can be subclassified as being either an enclosing boundary or a hole. The outer boundary of a PR is known as an enclosing boundary, and any inner boundaries are known as holes. Thus each boundary forms an extent of one, and only one, PR and each PR is bounded by one enclosing boundary and zero or more holes.

Boundaries are equivalent to a polygon in geometry, but they also have an associated direction to distinguish enclosing boundaries from holes. Where one PR completely surrounds another uncut PR, the polygons describing the enclosing boundary of the inner PR and the hole in the outer PR are identical in shape. The two boundaries, however, remain unique since they have opposite direction. The direction of a boundary thus relates it to one specific PR.

Boundaries are composed of links which are similar to arcs. DAM is unconcerned as to how the link geometry is represented but assumes that links are node matched. It also assumes that spaghetti digitising is topologically structured by a pre-process into a link and node structure in order to extract boundaries. Each boundary when formed has a separate existence.

When PRs all occur at the same level, i.e. when there are no holes, there is a one-to-one correspondence between boundaries and PRs and thus the latter may also be assigned the identity of the boundary. Links provide the adjacency relationships between boundaries and their corresponding PRs as in GEOGRAF.

When PRs are nested, i.e. when holes exist, there is a many-to-one correspondence between boundaries and those PRs containing holes. The links still provide the adjacency relationships between boundaries, and also between clusters of adjacent PRs. The problem is that the identity of PRs containing holes remains unknown. What exists is the distinct references to separate boundaries at the outer and inner extent of such PRs, and no single reference to the PRs themselves. Forerunners to DAM were constrained by the inability in practice to resolve the separate references to boundaries and derive a unique identity for each PR.

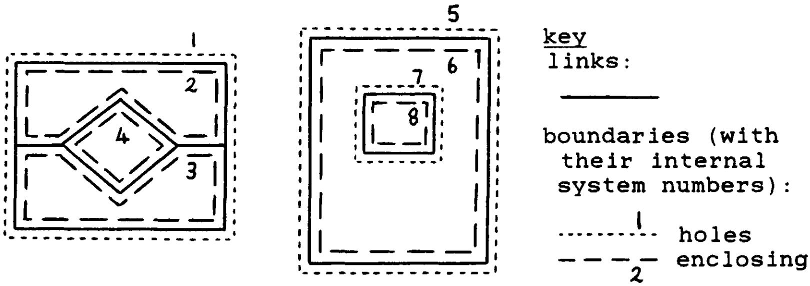
The complete set of boundaries can be viewed as forming a hierarchy which can be represented by a rooted tree. The root of the tree consists of a nominal reference to the part of the plane surface which surrounds all the other boundaries. Each boundary is enclosed spatially by every boundary which precedes it in the tree but no other. When two boundaries are identical in shape, the one which is a hole is considered as surrounding the one which is an enclosing boundary. Thus the level of each boundary in the tree is one greater than the number of other boundaries which enclose it. The set of holes at level one of the tree describe the holes in the outermost PR, whose enclosing boundary is undefined. This PR forms the complement of the union of all the other PRs on the plane surface. Also, boundaries at even levels of the tree will be enclosing boundaries and those at odd levels will be holes. Figure 1A illustrates the general case of the rooted tree. Note that the tree is not an ordered rooted tree as there is no set order to the edges leaving each vertex of the tree.

This hierarchical system can be applied to any set of boundaries irrespective of their complexity. If a map frame is digitised around all the existing boundaries, this would have the effect of creating additional boundaries and PRs (see Figure 1B).

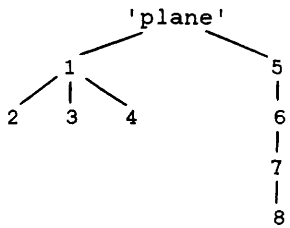
The derivation of such a tree _fully_ resolves the relationships between boundaries and thus the topology of the PRs, since the holes within each PR immediately follow the enclosing boundary for that PR in the tree. The nesting of PRs within holes is made explicit. The derivation of this tree for a set of boundaries is a one-off process

FIGURE 1 : THE HIERARCHY OF BOUNDARIES

A) Without a Map Frame



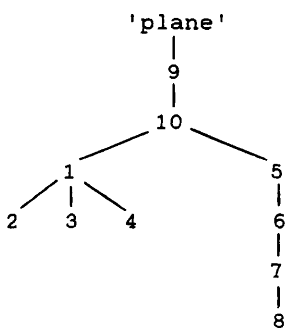
The boundaries of the primitive regions formed by the links.



level	type of boundary
0	
1	hole
2	enclosing
3	hole
4	enclosing

The rooted tree representing the hierarchy of boundaries.

B) With a Map Frame



level
0
1
2
3
4
5
6

The revised tree. Boundaries 9 and 10 are the hole and enclosing boundary formed from the links of the map frame, which does not touch any of the other boundaries.

requiring an exhaustive spatial search. One of the authors (PW) has devised and implemented an algorithm which attempts to minimise the computation required (see Wade _et al_, 1986).

Once a PR assumes an identity, it becomes the basic building block for area modelling and forms the pivot between the geometry and the geography.

Geography

It is necessary to distinguish between types (or categories) and instances of such types of spatial phenomena. Area _entities_ are categories of phenomena. Specific instances of categories are regarded as _objects_, which may be named. At the bottom-most level, there may be a one-to-one mapping between objects and PRs in the simplest case. Where there are disjoint parts, there is a one-to-many relationship between objects and PRs. DAM allows a variable number of objects to be associated with each PR (i.e. a many-to-one mapping). Thus it copes with both hierarchic objects (e.g. administrative hierarchies) as well as objects which overlap at any one level (e.g. broadcasting areas). Most other models cannot cope with the latter case.

DAM is not a universal model of topographic, let alone geographic, phenomena. It does, however, provide a framework for the flexible input of geographic information in a variety of formats. Since phenomena under consideration and the input format are both variable, further development of DAM requires an interface to a rule-processing capability. Ad hoc processes, based on a given rule-set, must otherwise be provided for integrating the geography with the geometry and for data validation and automatic editing.

DAM adopts a 'human' view of the geometry of area entities in that, given the links, it is possible to extract a unique identity for each PR, establish adjacency relationships and also extract the information about the nesting of PRs. Since the geometric and geographic relationships are each processed separately and then related via the PR, DAM also provides a capability for cross-checking the geometric and geographic information (see application below).

Comparison of DAM with precursors

HDS is similar to DAM in that it too regards DPBs as forming a rooted tree consisting of enclosing boundaries and holes although Edwards _et al_ (1977) interpret them as interior and exterior regions respectively. However, DPBs in HDS are boundaries of objects, rather than of PRs. Also, as those authors demonstrate (p 17), the boundaries that are formed need not be unique. This has the unfortunate effect of possibly resulting in more than one hierarchy of boundaries; their algorithm for forming the hierarchy is thus unduly complex and inefficient.

If HDS is mistakenly viewed as an extension of POLYVRT, then GIRAS, ARC/INFO, and DAM may also be wrongly perceived as extensions of GEOGRAF. GIRAS uses objects to form its geometric hierarchy. It therefore does not retain the concept of GEOGRAFs LCGU in its purity. ARC/INFO uses the equivalent of PRs for connecting ARC with INFO but does not include the concept of directed boundaries. The LCGUs in GEOGRAF are simply connected - thus there is a one-to-one correspondence between boundary and LCGU. The PR in DAM may consist of a set of boundaries with distinct references. In the absence of objects, these separate references are mathematically resolved to yield a unique reference to each PR.

DAM provides a synthesis of its precursors and identifies the essential functions of the unit line, unit boundary and unit area. Within a specific pragmatic model, the functions of these various parts may be replicated or transferred to other units for efficient computer handling or for improving the user interface.

It is possible to deduce and use any pragmatic model, which is consistent with DAM, to specify data formats for various tasks, e.g. data input, output, or transfer. As a corollary, DAM can be used to evaluate whether a complete and coherent description of area entities can be derived from a proposed pragmatic model. DAM was in fact constructed precisely for that purpose as described in the next section.

ONE APPLICATION OF DAM

The 1:625,000 database was established by the Ordnance Survey (OS) for purposes of experimentation by themselves and others. The general aim was to provide positive evidence towards the design of the 1:50,000 database (Haywood, 1984). The structure of the database was recognised as a crucial factor influencing not only the usefulness of a small-scale database but also its feasibility. The structure has implications for the cost of initial data capture and subsequent maintenance of the database.

We undertook to evaluate whether the OS design for representing the hierarchy of administrative areas (districts, counties, and countries) was adequate in concept, structure, and content for other purposes. The OS scheme uses the feature code of a link to indicate the type of administrative boundary it represents. Since administrative units form a hierarchy, it could be inferred that the boundaries of high-level objects also form a boundary of objects below them in the hierarchy and that the coastline could form all other boundaries.

Each detached part of an administrative unit is indicated by an area-seed, a representative point within the polygon enclosing that part of the object. This polygon is similar to a GEOGRAF polygon. The object's name and feature code, indicating its type, are associated with the seed.

This pragmatic data model is extremely convenient and cost-effective for data capture since it records once, and only once, each explicitly recorded map detail. The implications of this design are as follows. Not all PRs carry area-seeds, even at the bottom-most level of the administrative hierarchy. The sea areas in particular are not explicitly identified and have to be inferred. Furthermore, the spatial subset supplied to us was known to lack area-seeds for some objects (and thus PRs) which were cut by the map edge.

At higher levels, the seed within an object's polygon will only occur within one bottom-level object and one PR. The object hierarchy must therefore be inferred from the fragments of information, using the scattered clues and the nesting rules for the hierarchic link and area-seed feature codes. Also, island objects with land counterparts do not in general carry information on nationality although a county seed may be present. Finally, holes in objects can only be found by spatial searches.

The concept of the PR provided a convenient framework for solving the puzzle. The first stage was the extraction of the full geometric topology, i.e. a DAM model. All known information was then filled in and others, such as land and sea areas, base-level objects and parts of the object hierarchy, were inferred. The partially-formed object hierarchy was then used to fill other information, e.g. the nationality and/or county of islands and seaward extensions of some administrative units.

We were consequently able to identify objects at all levels whose area-seeds were missing, i.e. we have the capability for identifying missing data. The object hierarchy and the rule set were then used to validate the data and we identified the one link, within the data set, whose feature code was wrongly encoded (for details see Visvalingam _et al_, 1985). Finally, we have the capability to output this data in any of the previously reviewed formats.

CONCLUSION

DAM allows the geometry and the geography of areal entities to be decoupled for separate analysis. This disassociation offers flexibility. This, combined with the flexibility offered by the concepts of concurrent and rule processing, promises a means whereby the disparate data requirements of various tasks, people and processes, can be reconciled.

This paper has also described how DAM can be used with relevant rule sets in a post-process to emulate 'human' interpretation of feature-coded area maps by computer.

ACKNOWLEDGEMENTS

This research was sponsored by three organisations. The Science and Engineering Research Council granted the CASE studentship held by P. Wade, with the Market Analysis Division of CACI as the collaborating body. The Ordnance Survey also provided a small grant and access to their data and internal documentation.

REFERENCES

Allder, W.R. and Elassal, A.A. (1984) USGS Digital Cartographic Data Standards. Digital line graphs from 1:24,000-scale maps. U.S. Geological Survey Circular 895-C.

Edwards, R.G., Durfee, R.C. and Coleman, P.R. (1977) Definition of a hierarchical polygonal data structure and the associated conversion of a geographic base file from boundary segment format. An Advanced Study Symposium on Topological Data Structures for Geographic Information Systems, Harvard University, Cambridge, Massachusetts.

ESRI (1985) ARC/INFO Users Manual - Version 3. Environmental Systems Research Institute, California.

Haywood, P.E. (1984) The Ordnance Survey 1:625,000 Database: general pinciples and data structure, Ordnance Survey Internal Report.

Mitchell, W.B., Guptill, S.C., Anderson, K.E., Fegeas, R.G. and Hallam, C.A. (1977) GIRAS: A Geographic Information Retrieval and Analysis System for Handling Land Use and Land Cover Data, U.S. Geological Survey Professional Paper 1059.

Moellering, H. (1984) A working bibliography for digital cartographic data standards. Issues in Digital Cartograpic Data Standards, 5, National Committee for Digital Cartographic Data Standards, Columbus, Ohio.

Peucker, T.K. and Chrisman, N. (1975) Cartographic data structures, The American Cartographer, 2(1), 55 - 69.

Wade, P., Visvalingam, M. and Kirby, G.H. (1986) From line geometry to area topology, Cartographic Information Systems Research Group Discussion Paper 1, University of Hull.

Visvalingam, M., Kirby, G.H. and Wade, P. (1985) Extraction of a Complete Description of Hierarchically Related Area Objects from Feature-coded Map Details, Stage I Report, Ordnance Survey (OS) contract on Computer Handling of OS 1:625 000 Digital Maps. (This document is available as an OS Technical Paper).