
Road Extraction and Topographic Data Validation Using Area Topology

D. A. VARLEY AND M. VISVALINGAM

Cartographic Information Systems Research Group, Department of Computer Science, University of Hull, Hull, HU6 7RX, UK

1. INTRODUCTION

The Cartographic Information Systems Research Group (CISRG) of the University of Hull is researching the automatic recognition of spatial objects based on their spatial descriptions alone. The feasibility study on road extraction, reported in this paper, formed a part of an SERC CASE project (January 1990–January 1993); the collaborating body was the Ordnance Survey of Great Britain (OS). The project involved the recognition of objects implied in large-scale topographic maps. We define object recognition as the process of identification of objects from their forms alone, i.e. without recourse to semantic labels. Unlike recognition, extraction uses all available information, including semantic labels manually associated with line segments. The principal aim of this feasibility study on road extraction was to assess whether road recognition might be impeded by the structure and content of an experimental topographic database, designed and created by the OS. The emphasis in this study was on acquiring a good understanding of large-scale topographic data and of the data processing problems. Although some attention was paid to the efficiency of processing, this was not the primary concern at this stage. Efficient processing of geo-spatial data is a major topic of research which has to take into account a wider range of considerations (Frank, 1991).

This paper makes a number of contributions. Firstly, it identifies novel procedures for road extraction. Roads would normally be extracted by finding the regions within which road centre lines or road names occur. This research describes a more efficient and robust method. Roads can now be extracted by using only the topological information, normally held in memory, using concepts embodied within the Disassociative Area Model (Kirby *et al.*, 1989). There is no need to access the vertices on disc which define the precise boundaries of regions. Secondly, the study also suggests how the extraction of roads can be trivialised by minor changes to the data specification. Thirdly, the paper suggests how software for re-casting the data into Disassociative Area Model (DAM) format can be used for detecting violations of the specified data structure. Fourthly, it suggests the type of semantic checks which can be carried out to ensure that relevant aspects of the data are consistent and to identify errors. These checks indicate that the data specification does not expedite road recognition since the boundaries of extracted roads include extraneous features. Thus, road recognition cannot rely

on simplistic rules about roads but must accommodate the presence of anomalies. Fifthly, the study demonstrated how some of these anomalies may be automatically detected. Finally, the study has provided a catalogue of roads which can be used to derive empirical rules for recognition and to assess automatically the success of alternative procedures for recognition.

The paper is organized as follows. In the next section, the paper sketches a brief background which indicates the rationale for this work and some underpinning concepts. It then briefly describes the input data and its recasting into the data structures associated with the DAM (Kirby *et al.*, 1989). The procedures which were adopted for verifying that the data conformed to the anticipated structure are then outlined and potential consequences of some violations are described. Next, the implications of the point-in-polygon approach to road extraction are considered. The paper then outlines a much simpler novel process for extracting roads. It then describes the procedures which were adopted for checking the consistency of semantic labelling. Next, it discusses the extent to which our findings should influence the database model and concludes by summarizing the main findings.

2. BACKGROUND

The late 1980s has witnessed a growth in Geographical Information Systems (GIS). In Britain, the Chorley Report (DoE, 1987) has led to the formation of the UK Association for Geographic Information (AGI) and GIS Specialist Groups within many learned and professional societies, including the British Computer Society. The Chorley Report refers to the diversity of users and uses of geo-referenced data which give rise to different types of GIS (Visvalingam, 1990). In Britain, Land and Property Information Systems (LIS) use large-scale (1:1250 and 1:2500) topographic data as a base reference against which they record, manage and inter-relate their own data. The OS has been responsible for creating the topographic base maps used by numerous bodies, such as Her Majesty's Land Registry, the utilities and Local Authorities. The OS is renowned internationally for its pioneering activities in digital mapping since the early 1970s, and its Research and Development Division continues to experiment with alternative database designs to meet changing market needs.

At the application level there is growing interest in object-oriented databases. However, the design of the

class structure of objects is application dependent. For example, vehicle routing, highway maintenance, cutting of grass verges and other activities which relate to roads will tend to develop their own object structures. Consequently, data vendors tend to supply only partially structured data and it is the users' responsibility to extract the necessary object descriptions. The vendor may add value to data by making explicit, through a combination of automated and manual processing, some objects implied by the basic topographic data. For example, the OS itself defines some objects, such as vegetation, for its own map production. However, the automatic recognition and labelling of objects encoded in digital topographic maps remains an outstanding problem.

The OS vector topographic data consist of point, line and text features. A semantic label, called a feature code, is associated with each feature to link in relevant semantic information and the cartographic symbolism for rendering. Features are partial descriptions of objects. An object may be described by one or more features and a feature may form a part of several higher-level features and objects. Object recognition seeks to add value to data by grouping features so that they form higher-level entities, both features and objects, and by relating currently free-standing, or else manually input, information with these derived entities.

In the past, data suppliers have supplied feature-coded data in unstructured, so-called spaghetti, form. In spaghetti form, lines are broken when the feature code changes but may otherwise cross each other and themselves. This simple structure is sufficient for semi-automated map production but it does not facilitate the chaining of the polygons which define areal objects. For this, spaghetti vectors must be structured into a link-and-node form. In this structure, lines are not allowed to cross themselves or other features. Where such intersections occur, lines are split forming segments, called links by the OS. All links begin and end at nodes, which establish the connectivity of the links. The process of link-and-node structuring is usually semi-automated. The output of this process can contain a few errors such as overshoots, which need to be trimmed, and is thus subject to a cleaning process. The OSBASE data we received was link-and-node structured and cleaned.

The feasibility study on road extraction had to achieve the following tasks. It had to:

- Organize the link-and-node data into a form which explicitly recorded, and topologically related, the regions of space implied by them.
- Develop efficient and robust procedures for identifying and labelling roads.
- Validate the data.

3. FROM LINE TO AREA TOPOLOGY

3.1. Input data

The OS has derived a number of prototype topographic databases for experimental purposes. The feasibility

study was based on a variant of the OSBASE dataset, which has now been superseded. Many of the properties of this experimental version, described below, had to be uncovered from detailed OS internal reports, personal discussions with R&D staff of OS and by exploration of the database. OSBASE data for Birmingham, consisting of 12 sheets, were received in July 1990. Topologic processing revealed errors, some of which were manually fixed by us. We then received a further 16 sheets of OSBASE data for Canterbury, which was believed to be correctly edge-matched and of higher quality, in February 1992. One of the objectives of the study was to use spatial data models to validate the data.

OSBASE, as the name suggests, was designed as a base layer against which GIS applications could register their own data. It records point, line and text features on published (i.e. hard copy) 1:1250 base maps. OSBASE also includes explicit descriptions of some areal objects, such as roofed areas, water features, vegetation and slopes to satisfy internal requirements for automating area fill in map production, a task which had previously been completed manually. This implies that (i) the geometric details of these features have been rigorously checked to ensure that the polygons describing these objects are properly closed and (ii) that representative points called area seeds have been digitised to include additional feature codes for these areal objects. Roads and land parcels are only implied and not explicitly defined within OSBASE.

The feasibility study was based on some information within the OSBASE layer and a separate road centre line layer, which is described later. It did not use all the layers within OSBASE since it was quite apparent that roads could be extracted using the links in the OSBASE and road centre line layers. Point features, area seeds and text features, which were superfluous to this study, were excluded. Not all line features were included. For example, some overhead features such as electricity lines, recorded in a separate layer within OSBASE, were excluded. Other overhead features, such as overhead roads and walkways, which form a part of the topographic base, were retained.

The input to the feasibility study therefore consisted of a subset of the links which make up some OSBASE features; an example of the OSBASE input for one sheet is shown in Figure 7(a). This subset includes two types of information on roads, i.e. links feature-coded as ROAD_METALLING links and road names. The reasons for not using road names are discussed in Varley and Visvalingam (1993). Pertinent facts on ROAD_METALLING links and the road centre line network are provided below.

3.1.1. ROAD_METALLING links

Only one feature-code is associated with each link. Since a link may form a part of several objects, feature-codes are assigned priorities. In Figure 1, a subset of feature

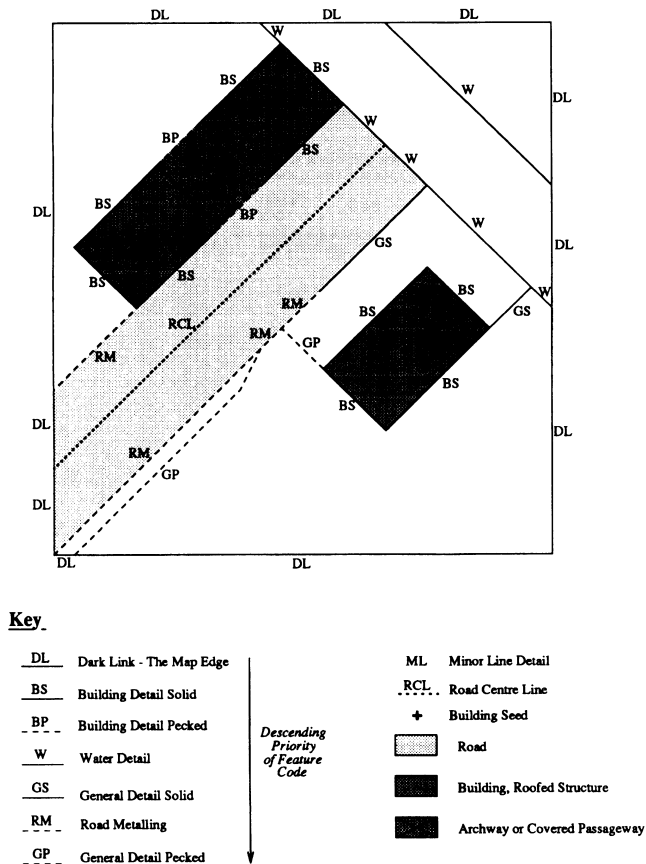


FIGURE 1. Some feature codes associated with road links.

codes, which are associated with road boundaries, are listed in descending order of priority in the data specification. Feature codes associated with road boundaries do not have the highest priority. As a result, only some of the links forming the edges of road polygons are feature-coded as ROAD_METALLING links. Figure 1 shows how a road may be defined by a set of dark links, building, general and water detail in addition to ROAD_METALLING links.

3.1.2. Road centre line network

The networks of road centre lines for the areas covered by the underlying OSBASE maps are recorded in a separate layer. This information was manually digitised as a free-standing layer for vehicle routing applications. It is geometrically connected to the underlying OSBASE data at the map edges and at other places which are irrelevant to this study. The connection at the map edges enables the separate parts to be stitched together into a continuous network.

3.2. The DAM

Kirby *et al.* (1989) described the advantages of disassociating and independently processing the spatial and aspatial descriptions of topographic objects. They described the concepts underpinning their DAM for recording area topology and outlined procedures for deriving such

topological information from link-and-node structured spatial data. They discussed the flexibility that this offers for modelling geographic phenomena by linking aspatial with spatial descriptions at a later date. This facilitates the cross-checking of the consistency of semantic coding. They used the DAM model to locate the few residual errors remaining in the 1:625 000 experimental database of the hierarchy of administrative areas in parts of England and Wales. Visvalingam and Sekouris (1989) also demonstrated the utility of DAM for locating geometric and semantic errors in an experimental 1:50 000 topographic database. However, the methodology for validating topographic data is still underdeveloped.

The creation of digital topographic maps is an involved process with a series of automated and manual stages. Given the size of the databases and the cost of rigorous checking, it cannot be assumed that digital maps are error-free nor that they meet the users' needs with respect to their structure and content. The feasibility study was therefore undertaken to expose the types of data-related problems facing the automatic recognition of roads based only on their forms and juxtaposition *vis-à-vis* other features.

The first decision which had to be taken was to determine whether the OSBASE and centre line links should be treated as separate or an integrated set. Varley and Visvalingam (1993) explain why road centre line links were not integrated with OSBASE data but were used separately.

The data structures in Figure 2 articulate the DAM for representing area topology. Since the input data conforms to the link-and-node model it can be cast into the data structures shown for links, nodes and coordinates. Software developed by Wade (Kirby *et al.*, 1989) was used to add further fields to the links and node structures and to compile the boundaries structure. These structures are only indicative of the type of information required since such information may be derived and represented in other ways. For example, Kirby *et al.* (1987) and Visvalingam and Sekouris (1989) considered how these same entities may be represented and managed using a relational database model.

Wade's software chains the OSBASE links into a set of polygonal boundaries, which are then structured to form a geometric hierarchy consisting of enclosing boundaries alternating with holes. Figure 3 illustrates this structure. A brief description of the underpinning concepts and data structures are provided here; a fuller description is provided in Kirby *et al.* (1989).

In Digital Cartography, link-and-node connectivity is encoded using the link, node and coordinate records represented in Figure 2. One link record is used for each link. The coordinates of the two end points of the link are stored in the node records and any remaining internal points are stored separately as a contiguous block in a coordinates file. The link record points to the start and end nodes, to the start and end of the block

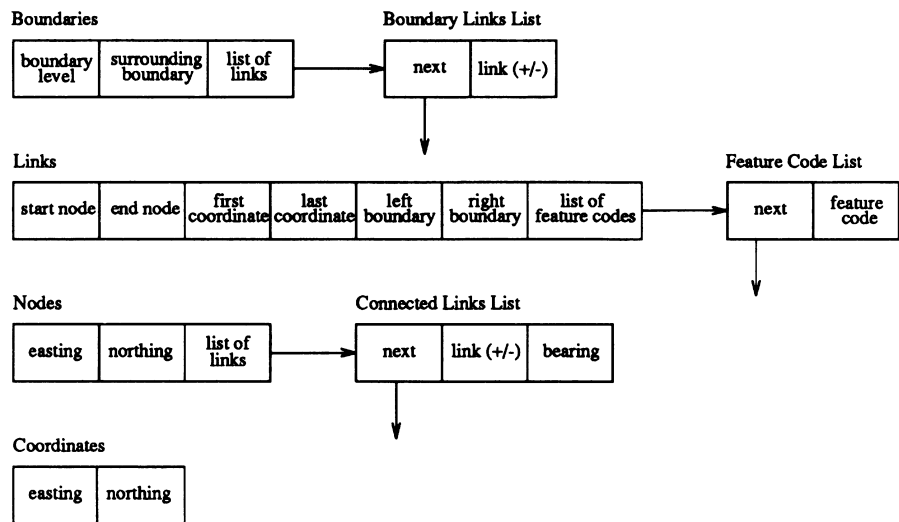


FIGURE 2. Data structures used for road extraction.

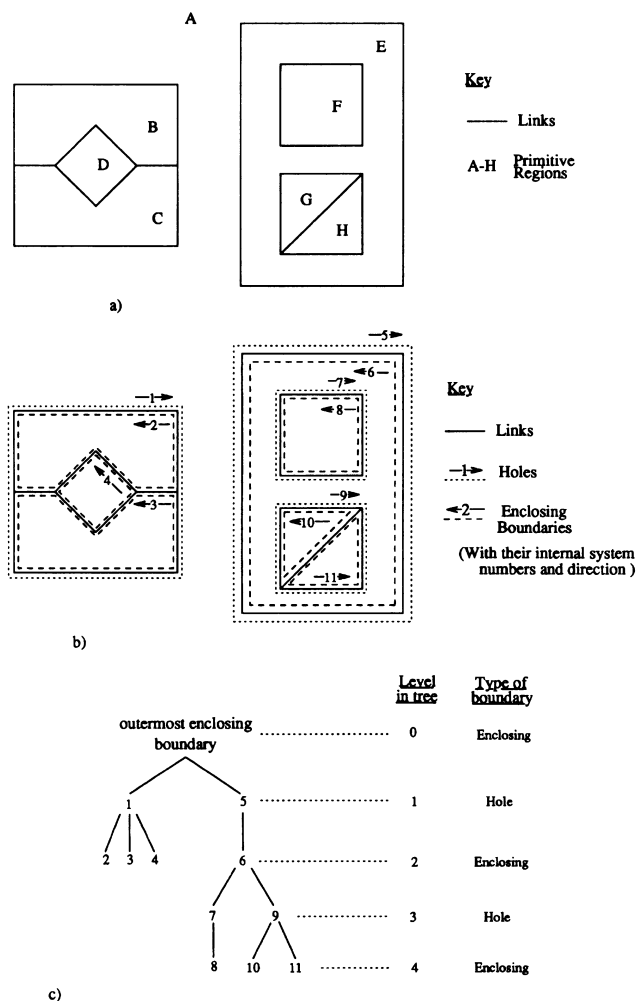


FIGURE 3. Derivation of the containment hierarchy of boundaries.

of coordinates and, to a list of feature codes. Initially, this list consists of only one feature code.

There is one node record for each location where one or more links start and end. With each node is kept a list of the record numbers of the links which meet there.

The record number is signed to indicate whether the link starts (positive) or ends (negative) at the node. The bearing at which the link joins the node is also held and the link list is sorted on this field so that links occur in a clockwise order around the node.

The DAM, first reported in Wade *et al.* (1986), adds a boundary record to model the area topology. Each boundary is a closed loop, with direction (see Figure 3b). Boundaries are sub-classified into enclosing boundaries and holes. Each enclosing boundary records the outer extent of a primitive region; the inner boundaries of the latter are termed holes. Each boundary forms an extent of one, and only one, primitive region and each primitive region is bounded by one enclosing boundary and zero or more holes. There is one exception; the outermost primitive region has no enclosing boundary but just one or more holes (in Figure 3b, the outermost primitive region is described by holes 1 and 5).

The boundary record is used to hold the details of each boundary. The containment relationship between a complete set of boundaries may be viewed as forming a hierarchy represented by a rooted tree (Figure 3c). Each boundary is enclosed spatially by every boundary which precedes it in the tree but no other. The level of each boundary in the tree corresponds to the number of boundaries which surround it. The outermost enclosing boundary at the root of the tree is a nominal reference to the part of the plane surface which surrounds all the other boundaries. The derivation of such a tree fully resolves the containment relationships between the boundaries and thus the spatial extent of the primitive regions since the holes within each primitive region immediately follow the enclosing boundary of that primitive region in the tree. The boundary record notes the level of each boundary and its surrounding boundary. The link record keeps a record of the boundaries to the left and right of each link.

Once a primitive region assumes an identity, it becomes the basic building block for modelling areal

objects and it forms the connection between the geometry and the geography. Within GIS, there is often a many-to-many relationship between primitive regions and areal objects. However, in this feasibility study, the primitive regions need only be labelled initially as CANDIDATE_ROADS and eventually as ROADS, ROAD_NEIGHBOURS or UNCLASSIFIED.

3.3. Verification of input data

To recapitulate, the extraction of the area topology includes the following stages:

- Checking of input data.
- Chaining the links into polygons.
- Forming the geometric hierarchy of polygons as shown in Figure 3.

The checks which were performed were described in Visvalingam *et al.* (1987). When the OSBASE data was processed by this suite of software, it failed to form the hierarchy of polygons for one of the Canterbury sheets because it could not find the level 1 hole, i.e. the map edge polygon. The precise location of the error proved to be a time-consuming task. The problem is characterized in Figure 4. In Figure 4(a), links a–d form part of the map edge. Link e is a part of a ROAD_METALLING link. Part of e (e') is co-incident with a part of c, i.e. c' (see Figure 4b). The link-and-node structure does not allow links to overlap and coincide in this way. If a node was inserted at E, this could result in duplicate occurrences of e', which was only 30 cm on the ground. This is smaller than the survey tolerance. Stage 1 was thus enhanced to detect such cases by including a check which flag links with identical bearings at a node. This revealed a number of instances of duplicate links and dangling lines without free nodes which were only about 2 cm on the ground on average. This suggests that these and other errors, some with configurations similar to those in Figure 4, are software generated. In the event only the situation

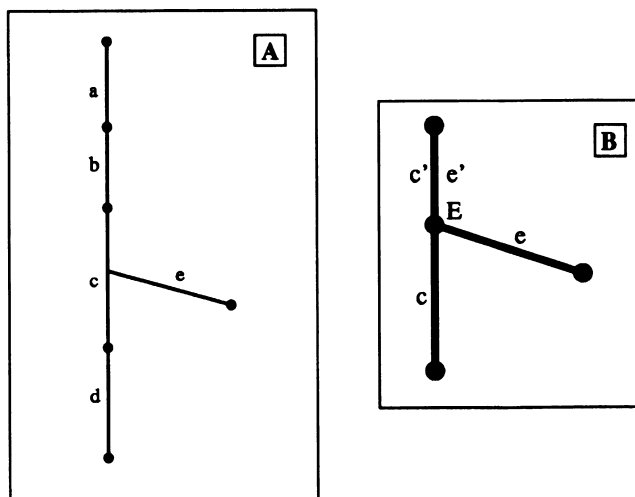


FIGURE 4. An example of a link-and-node violation.

described in Figure 4 proved to be critical and was patched manually. Since cleaning is time consuming, we ignored the other errors which could prove to be critical to other tasks. Although OSBASE does not guarantee that road polygons are properly formed, there were relatively few cases of geometric inconsistencies, some of which are described later.

4. ROAD EXTRACTION USING POINT-IN-POLYGON TESTS

Road centre lines are manual abstractions of the connectivity of the road network. Since centre lines are always located within road polygons, the extraction of road networks may be conceived as essentially one of finding those polygons which contain the centre lines. This idea may be articulated in different ways. In this section, we consider the implications of using the point-in-polygon approach for finding the containing polygon. In the next section this is contrasted with the use of topological clues.

A full description of the process of road extraction, including a discussion of special cases and algorithms is provided in Varley and Visvalingam (1993); only a brief summary is provided here.

In the previous section we described how the OSBASE links were re-structured to make explicit the primitive regions they describe. Road extraction now consists of the following steps illustrated in Figure 5.

- Select only those primitive regions which were likely to be roads for further consideration. These include at least one ROAD_METALLING link. Only the enclosing boundary of these regions need to be considered.
- Select suitable seeds, i.e. points on the road centre line network, for point-in-polygon checks
- Label primitive regions which contain these seeds as those forming road networks

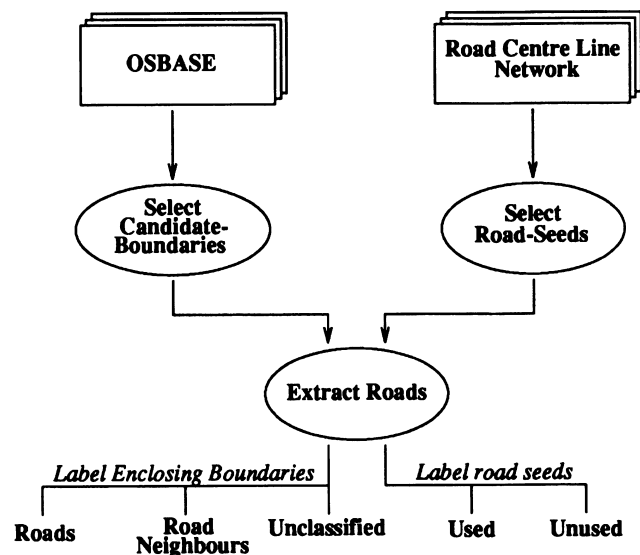


FIGURE 5. Point-in-polygon approach to road extraction.

- Label the neighbours. The neighbour is a primitive region which shares one or more ROAD_METALLING links with the identified road. If these are discounted, then a neighbour which has no other ROAD_METALLING link is unlikely to be another road. The status of a neighbour with additional ROAD_METALLING links cannot be determined.

At the end of this labelling process, candidate_boundaries will be labelled as ROAD, ROAD_NEIGHBOUR, or as UNCLASSIFIED. Seeds would be labelled as USED or UNUSED.

4.1. Discussion of results

The process of road extraction is essentially one of classification. The success of this process may be evaluated by identifying and explaining errors of omission and commission. All cases of the following types were automatically identified and manually inspected:

- Road seeds which remain UNUSED. There was only one instance of this type. It occurs because its enclosing_boundary, which forms a genuine road, does not contain any ROAD_METALLING links. The probability of this occurring in an extensive road network is very small. These errors of omission tend to occur where truncation by the map edge leaves a small section of road as shown in Figure 6(a). As a result of building and water detail having higher priority in the feature coding hierarchy than ROAD_METALLING, none of the links in this road boundary have the ROAD_METALLING feature code. The polygon was thus not identified as a candidate.

There are two possible solutions to this problem. One is to highlight these cases for manual checking and labelling. The other is to achieve a more complete solution. The latter solution is possible but this means that all enclosing_boundaries which touch the map edge, and not just those with ROAD_METALLING links, are considered as candidate_boundaries. This would mean that some 6170 instead of 1669, boundaries (i.e. 14.0% instead of 3.8%) would become candidates.

- Candidate_boundaries with ROAD_METALLING links which remain UNCLASSIFIED. These occur in two situations:
 1. When only one side of a road skirts the map edge resulting in a missing seed (see Figure 8). Without a road centre line, both the road and the adjoining regions will remain UNCLASSIFIED and will have to be manually checked and labelled as either ROAD or ROAD_NEIGHBOUR.
 2. When a road network is segmented by overhead features (see Figure 6b). Here selection of a single seed is insufficient as the overhead feature truncates the road into two polygons, only one of which contains the sample seed.

An analysis of these errors indicate the following:

- The algorithm for road extraction was too simplistic.
- The data specification does not expedite road extraction.

The problems related to data are considered first since they guide the problem-solving strategy. As pointed out earlier, OSBASE is a cartographic model of data; these cases show how only the uppermost features are depicted in maps. Road centre lines, on the other hand, record connected networks. Thus, where overhead features occur it is not possible to use just one seed from the centre line network.

Again there are two solutions. One is to use automatic validation procedures to highlight UNCLASSIFIED enclosing_boundaries for manual labelling as before. Another somewhat involved procedure is to consider all the points in the road centre line network (see Varley and Visvalingam, 1993). It is insufficient to consider only the nodes of this network. (This would also eliminate another type of error. There were two instances when road centre lines occurred within what appeared to be neighbours. In one, the centre line extended into and terminated within a roundabout. This is a residual error due to a change in data specification. The other case is typified in Figure 6(c). Here link A crosses a ROAD_METALLING peck and extends into an access road. Other similar access roads have not been included in the road centre line network; this suggests that the digitisation of link A may have been an error.)

Unfortunately, the inclusion of all points in the road centre line network will lead to errors of commission since points on the road centre line can and do fall within other regions which are not roads. For example, in Figure 6(d), road centre line points can fall in region A and C, which cannot be ruled out as candidates until region B is correctly identified as a road.

In addition to these errors of omission, the data specification introduces some anomalies. This is because ROAD_METALLING links are sometimes omitted, for example, where a road ends and a drive to a block of garages begins. Since neither roads nor appended features are area-filled in OS maps, OS does not need to separate roads from other features. Cases labelled A-C in Figure 7(b) illustrate how the road can extend as a result into pavements, driveways, car parks and other objects. Such merging of regions for a road with one or more of its neighbours can impede road recognition since it can distort the metrics, such as shape, which have been used by others to guide the process of object recognition. Thus, procedures need to be developed to detect automatically the inclusion of extraneous regions.

5. ROAD EXTRACTION USING TOPOLOGICAL CLUES

The point-in-polygon approach to road extraction is modelled on the direct visual perception of roads as

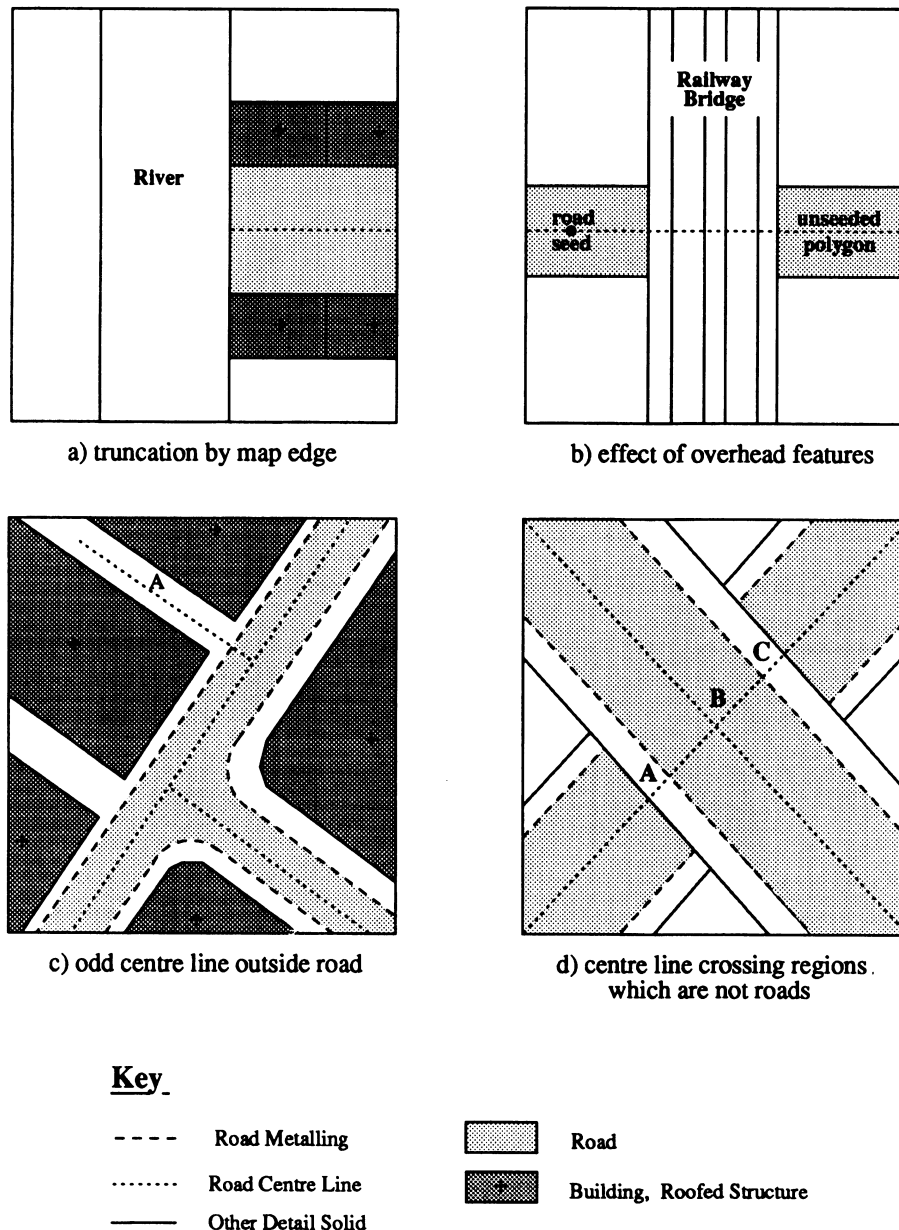


FIGURE 6. Some special cases not resolved by the point-in-polygon approach.

regions traversed by the road centre-line network. It is inefficient because it involves the retrieval from disc of locational information (coordinates). Spatial data models, such as DAM, are intellectual constructs devised to provide a framework for reasoning. They guide us towards more efficient solutions although, once found, the solutions are not dependent upon the re-structuring of the data into DAM form. Also, these solutions encourage us to visualize the problem in a different way.

The problem of road extraction is essentially one of identifying the region within which the road centre line lies. Centre lines have to be used because, the ROAD_METALLING links do not indicate whether they bound regions to the left or the right of them. This impedes extraction when we have sibling regions as in Figure 9(a). However, since roads form connected net-

works, ROAD_METALLING links in holes provide unambiguous clues. For example, in Figure 9(b), region C which is inside a hole cannot form a road since it does not lead anywhere. Thus, ROAD_METALLING links forming holes normally indicate that the region which contains C, i.e. B in Figure 9(b), must be the road. Unfortunately, not all roads contain holes. Moreover, not all holes in roads contain ROAD_METALLING links.

However, holes at all levels of the geometric hierarchy provide clues for identifying roads by elimination of unlikely candidates. The hole at level 1, i.e. the map border, is equally informative. We noted earlier that the road centre lines connect to the OSBASE layer at the map edge. Since the region outside the map lies to one side of the MAP_EDGE links with centre line nodes,

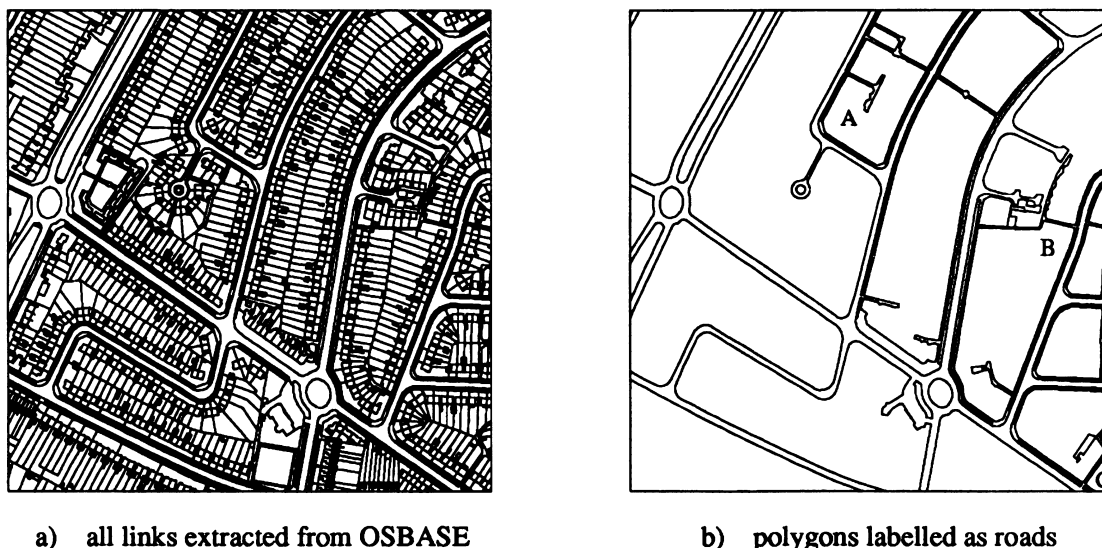


FIGURE 7. Roads which include neighbouring features. (Data: Crown Copyright)

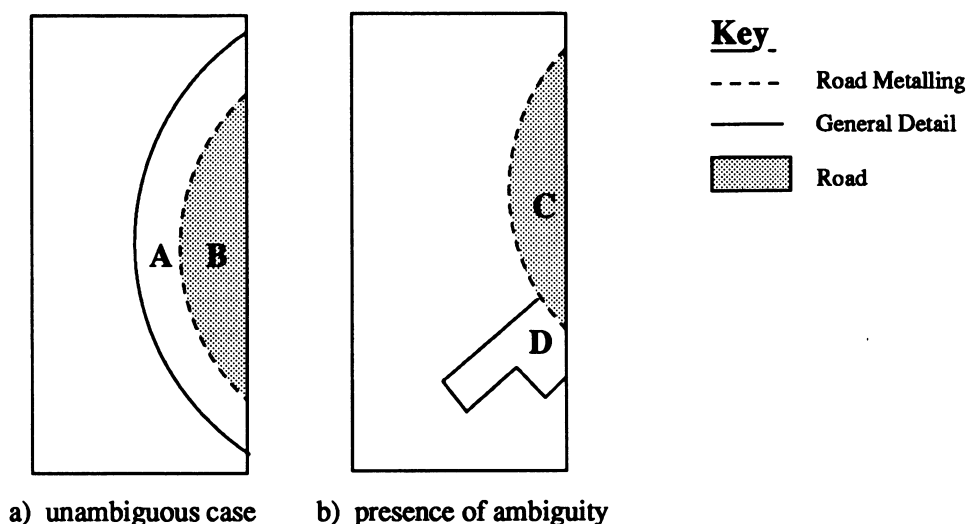


FIGURE 8. Extraction of roads at the map edge with no centre lines.

the level 2 enclosing boundaries on the other side must be roads. We can use this knowledge for labelling the candidate boundaries used in the point-in-polygon test as explained in detail by Varley and Visvalingam (1993). Only a brief description is provided here.

- Label the obvious cases first. These may be extracted by using the road centre line nodes at the map edge to locate the MAP_EDGE links within which they lie. Here the point-in-polygon test is replaced by a range check, involving only the nodes; thus the vertices in the coordinates file on disc need not be accessed. This step will identify all roads with centre lines which intersect the map edge.
- The roads skirting the edge which do not have a centre line can then be labelled by examining the still UNCLASSIFIED level 2 enclosing boundaries, with both MAP_EDGE and ROAD_METALLING links. If any of these have one and only one MAP_EDGE

link, it can be provisionally labelled as a ROAD. In cases, such as that in Figure 8(a), this will correctly identify B as the road. If A were the road, a road centre line will have been included. However, in some cases, as shown in Figure 8(b), it is possible to pick D first and wrongly identify it as a ROAD. Region C will not be labelled as ROAD_NEIGHBOUR since it includes ROAD_METALLING links other than those which form the boundary between C and D. Since two neighbouring polygons will now be labelled as ROADS, it is possible during the review stage to re-label D as ROAD_NEIGHBOUR automatically as it does not have any other ROAD_METALLING links. As these roads are deduced rather than explicitly recorded in the database, their labelling must be manually checked since data conditions can lead to wrong conclusions. There were no problems of misidentification in the sheets we processed.

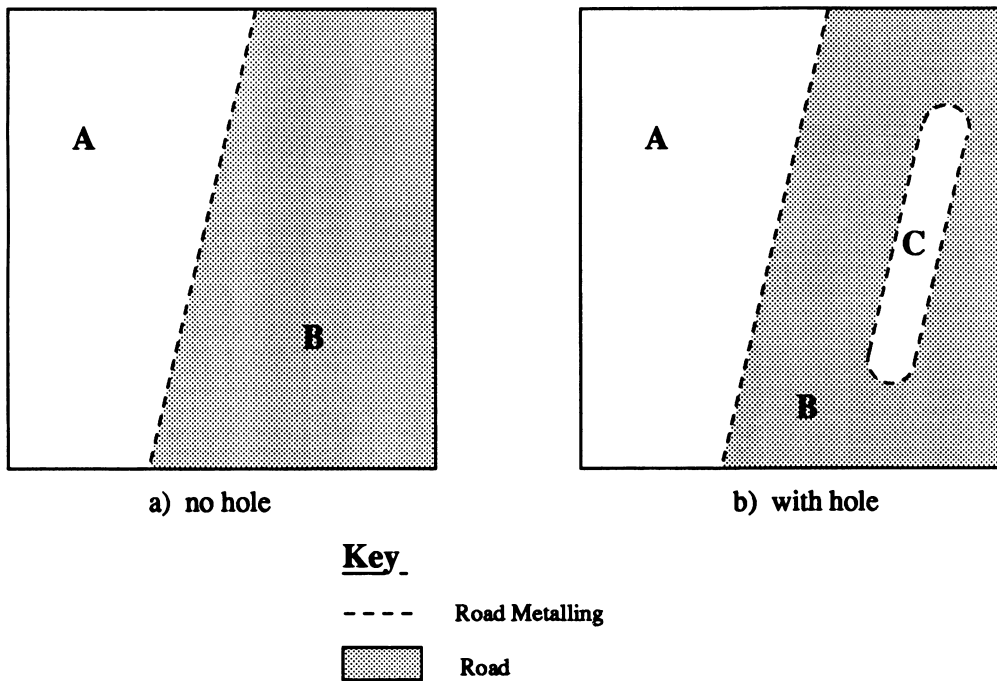


FIGURE 9. Use of a hole to locate the road.

- By now only those candidate boundaries which are detached from the main road network, and others arising out of digitizing or link-and-node structuring errors, will remain UNCLASSIFIED. (In the given data, these form only 2.6% of the candidate boundaries.) Clusters of boundaries can become detached because of the presence of overhead features (see Figure 10a). It is possible to resolve these detached roads manually; this should be a last resort and should really be used for checking purposes only. Alternatively, point-in-polygon checks could be used as before. By now, only a small number of candidate boundaries, which consist of few vertices remain. Thus, these checks are no longer as onerous. Topologic reasoning could be used instead. If we visualize each cluster of detached enclosing boundaries as occurring within a hole (as in Figure 10b), then only the road has at least two sibling candidates with which it shares ROAD_METALLING links. Although this rule resolves all the cases we have encountered, it may not be foolproof.

A change to the data specification would make this process more reliable. Currently, the road centre line layer is not topologically integrated with OSBASE. For example, there are no nodes where the road centre line crosses links recording overhead features. The outermost links recording overhead features can be distinguished since they are feature coded as either BUILDING_PECKED, where a building forms the overhead feature, or GENERAL_SOLID for other features. If these were treated as vertical edges, i.e. like the map edge, and if the centre line network was connected

to OSBASE at these links by nodes, then road extraction would become a trivial process.

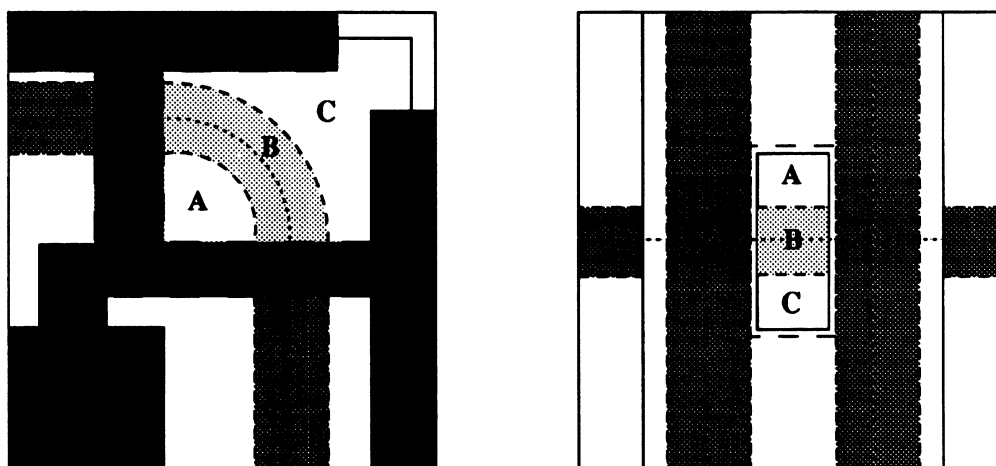
In summary, this process consists of three steps: labelling the obvious cases, then those cases skirting the map edge and, finally, the detached sections. Only the loops of ROAD_METALLING links and other errors, generated during the link-and-node structuring stage, and ambiguous cases as in Figure 6(c), will remain UNCLASSIFIED.

6. LABEL-BASED VALIDATION

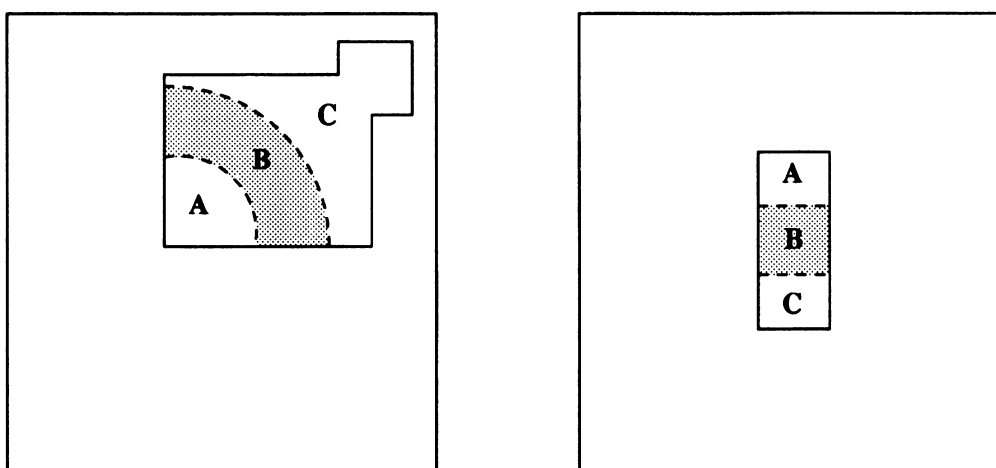
The labels assigned to links and boundaries can be used now to locate inconsistencies in the value-added database and assess whether they are critical or insignificant. These inconsistencies must be ascribed to one or more of the following reasons:

1. Errors in extraction process; these may be due to incorrect interpretations of the data specification or inappropriate logic.
2. Anomalies caused by edge effects and overhead features.
3. Errors in data arising out of contraventions of the data specification.
4. Inconsistencies arising out of ambiguities in the data specification.
5. Difficulties arising out of the nature of the data specification.

Two sets of checks were undertaken. The first set checked the correspondence between the OSBASE and road



a) Examples of detached parts of roads remaining UNCLASSIFIED



b) sibling candidates visualised as occurring within holes

Key

- Road Metalling
- Road Centre Line
- - - General Detail Pecked
- - - Building Detail Pecked
- Other Detail Solid




-  Road (still UNCLASSIFIED)
-  Road (already CLASSIFIED)
-  Building, Roofed Structure

FIGURE 10. Segments of roads detached by overhead features.

centre line layers; the second validated the consistency of OSBASE data.

6.1. Correspondence between OSBASE and road centre line layers

The following checks were made.

- Manually check if all regions labelled ROADS, which contain at least one point on the road centre line, are actually roads. They appear to be.

- Automatically flag all regions labelled as ROADS if they do not contain at least one point in the road centre line network. This check only identified the cases which could be attributed to cause 2. Although the centre line does pass through it, the small detached section of road located in a hole in an overhead motorway (see Figure 10), need not have contained road centre line points. The point-in-polygon approach would miss such cases.
- Automatically flag points in the road centre line

network which lie within regions which have not been classified as roads. This identified some overhead features as expected (cause 5; see Figure 6b). Since the case, sketched in Figure 6(c), is the only one of its kind in 28 sheets, it is likely that it is due to cause 3. There was one further case where the road centre line extended into and terminated in a roundabout. This arose from data digitized under a now superseded digitization specification.

- Automatically flag MAP_EDGE links, bounding a region labelled ROAD, which do not contain a centre line node. There were several violations of this rule and they were all the result of missing LOGICAL links to close off roads, i.e. cause 5). For example, the MAP_EDGE links of feature C in Figure 7(b) will be flagged. These checks indicate that apart from the two cases due to cause 3, manual input of centre lines is very reliable.

6.2. Consistency of the OSBASE layer

- Automatically flag enclosing boundaries with residual ROAD_METALLING links which remain UNCLASSIFIED. In this check, ROAD_METALLING links which already form a part of boundaries labelled as ROAD are discounted. This found a few (37) cases due to digitizing or cleaning errors which formed either a loop in the side of a road or left a small section of ROAD_METALLING link in a neighbouring feature, such as a pavement. These should have been removed in the cleaning stage but they do not impede road recognition.
- Automatically flag links of ROADS which have the same boundary number in both the left/right boundary fields. This implies that a ROAD has become a neighbour of itself. There were several instances of this. They appear to be due to missing LOGICAL links (cause 5).
- Automatically flag ROADS which are neighbours of other ROADS. This does occur and can be attributed to missing LOGICAL links (cause 5).
- Automatically flag boundaries forming roads which include feature codes which are of lower priority than ROAD_METALLING. This check identified a number of occurrences of links coded as GENERAL_PECKED or VEGETATION in road boundaries. They seem to be due to missing LOGICAL links (cause 5) more often than to coding errors. However, there are a few cases where the semantic labelling is incorrect. It would be very difficult to locate such residual errors manually without using a model-based approach, such as that used in this paper.

In general, the node-in-edge method for road extraction is more reliable than the point-in-polygon approach and, even using our still unoptimized implementations, it is

about five times faster. OSBASE did contain some link-and-node structure violations and a few feature coding errors. These are unlikely to impede road recognition in the majority of cases. The inclusion of extraneous regions within roads is more of a problem. However, the study has revealed several ways in which their presence could be detected.

7. DISCUSSION

The design of information systems is only partly influenced by academic considerations; it is constrained as much, if not more, by a host of external or inherited constraints. The OS, for example, has to adhere to government directives on increasing its levels of cost recovery, which influence decision-making within the organization (Rhind, 1991, 1993). In addition, a number of government committees of inquiry make more detailed recommendations based on user views. The Chorley Report (DoE, 1987), for example, suggested changes to the data specification to reduce the cost of data and to accelerate its analogue-to-digital conversion. The specifications for digitizing data place some limitations on subsequent structuring and modelling of the data. Historical influences have to be taken into account since the OS began its digitizing programme in the early 1970s. Changes to product models can have some impact on earlier stages with cost implications.

It is therefore not pertinent to make independent far-reaching recommendations without access to all of the relevant managerial information. Even from an academic standpoint, it is now quite widely accepted that systems should be designed to simplify human processes. Such improvements in human-computer interaction within a total system can often incur some additional complexity in computer processing. This background therefore has encouraged us to limit the scope and nature of our recommendations to those which are needed to eliminate ambiguity.

One of the reasons for creating experimental datasets, such as OSBASE, is for in-house assessment of the cost-benefit analysis and marketability of various designs. As the name suggests, OSBASE was designed as a base layer against which GIS applications could record their own data. Although some objects (e.g. buildings, vegetation, water) can be extracted using the semantics attached to links and seeds, others can only be properly defined with extensive editing. During the course of this study, OSBASE was superseded by other experimental prototypes of which we do not have any experience to-date. The road centre line network was created for route planning applications. Neither of the two layers were created for road extraction *per se* and it is therefore not surprising that the data specification is not ideal for road extraction.

A full review of the OSBASE and road centre line layers is outside the remit of this study, with its relatively narrow focus. The paper has pointed out how some

minor changes to the specification could simplify road extraction and make it less of a hit-and-miss process. The problems, and our recommendations, are summarized below.

7.1. Connection of OSBASE and road centre line network

The internal connectivity of the road centre line layer was being reviewed by the OS and is outside the remit of this paper. This layer may be visualized as occurring at a level above the OSBASE map. These two layers are vertically connected at the map edge and at other places which are irrelevant to this study. The network thus forms a disjoint layer even above the overhead features in OSBASE. This is topologically incorrect since the road network actually weaves under and over different overhead objects. To our knowledge, there are no formal systems for modelling 2.5-dimensional topographic data. It is therefore premature to suggest far reaching changes without further research. However, a minor change to the specification could simplify the process of extraction and take out some of the guesswork. The inclusion of nodes where the road network disappears under and re-emerges from beneath overhead regions, would also make it possible to distinguish automatically between genuine cases, where the continuity of the road network results in it crossing other objects, and errors arising from inconsistent digitizing and/or dated specifications.

7.2. The OSBASE layer

The OSBASE specification does not require that the spatial definition of road objects is complete and there are gaps in some boundaries. This problem may be transient since OS is already defining road objects explicitly in more recent experimental prototypes.

The imposition of artificial sheet boundaries can result in road centre lines being excluded and small sections of roads not having ROAD_METALLING links. We have shown how some of these cases may be found in sheet-by-sheet processing. Such ambiguities may be more easily resolved when seamless databases with edge-matched sheets become more widely available. The specification therefore does not need to address this transient problem.

The bulk of the remaining problems can be attributed to the semi-manual pre-processing of data, which is likely to incur some residual errors given the scale of OS operations.

7.3. Summary

This study has shown how roads may be extracted efficiently and reliably despite the presence of inevitable residual errors in link-and-node structured cartographic databases. We hope that the methodology we have developed will help to validate the data against the specification and correct it. It is quite likely that vendors will supply object-based data in future. Even if roads

are established by manual inclusion of area seeds, there will still be a need to validate the database.

8. CONCLUSION AND FUTURE WORK

This paper has demonstrated that the point-in-polygon approach to road extraction is inefficient and unreliable owing to special cases in topographic data. It provides a much simpler and more robust method based on topologic reasoning. The DAM provided a framework which allowed us to reason about road extraction in different terms. Once found, the solutions may be implemented without re-casting link-and-node structured topographic data into the full DAM format.

The paper has also demonstrated how the process of object extraction can assist in the verification of the topology and the validation of the semantic content of data. Now that the roads have been extracted, it is possible to assess the scope for automatic segmentation and naming of roads. This process would not only use the attributes of road centre lines and road text features but also the clues deduced by this feasibility study.

The catalogue of roads versus road neighbours and other objects can be used to guide the process of road recognition and evaluate the success of recognition algorithms. The exercise has also suggested how minor changes in the data specification can trivialise road extraction. Roads also provide contextual information to guide the recognition of other topographic objects of relevance to GIS.

ACKNOWLEDGEMENTS

This research was made possible by the award of a UK Science and Engineering Research Council (SERC) CASE studentship to Dominic Varley. We are grateful to the Ordnance Survey of Great Britain (OS), the collaborating body, for providing access to their digital topographic data and to the required information. We are particularly grateful to John Farrow, formerly of the OS, for his help, encouragement and input by way of discussion. Thanks are due to others at the OS, especially Nigel Venters and Steve Erskine for answering numerous queries relating to data, and Ross Christie for permission to publish this paper. We are indebted to Phil Wade, a past SERC CASE student within the CISRG, for permission to use his software for extracting the area topology.

REFERENCES

- Department of Environment (1987) *Handling Geographic Information*. Report of the Committee of Enquiry chaired by Lord Chorley. HMSO, London.
- Frank, A. U. (1991) Properties of geographic data: requirements for spatial access methods. In Gunther, O. and Schek, H.-J. (eds), *Advances in Spatial Databases, Proc. 2nd Symp.*, pp. 225–234, Zurich.
- Kirby, G. H., Wade, P. and Visvalingam, M. (1987) Storage and retrieval of topographic data using a relational database management system. In Haywood P. (ed.), *Proc. SORSA '87 Symp.*, Durham, pp. 1–28. Ordnance Survey, Southampton.

- Kirby, G. H., Visvalingam, M. and Wade, P. (1989) Recognition and representation of a hierarchy of polygons with holes. *Comp. J.*, **32**, 554–562.
- Rhind, D. W. (1991) The role of the Ordnance Survey of Great Britain. *Cartographic J.*, **28**, 188–199.
- Rhind, D. W. (1993) Policy on the supply and availability of Ordnance Survey information over the next five years. *Mapping Awareness*, **7**(1), 37–41 and **7**(2), 37–40.
- Varley, D. A. and Visvalingam, M. (1993) *Area Topology for Road Extraction and Topographic Data Validation*. Cartographic Information Systems Research Group Discussion Paper 11, University of Hull, Hull, UK.
- Visvalingam, M. (1990) Trends and concerns in digital cartography. *Computer-Aided Design*, **22**, 115–130.
- Visvalingam, M., Kirby, G. H. and Wade, P. (1987) Extraction of a complete description of hierarchically related area objects from feature-coded map details. In Haywood, P. (ed.), *Proc. SORSA '87 Symp.*, Durham, pp. 1–37. Ordnance Survey, Southampton.
- Visvalingam, M. and Sekouris, N. M. (1989) *Management of Digital Map Data Using a Relational Database Model*. End-of-grant Report to OS available as Cartographic Information Systems Research Group Special Issue 3, University of Hull, Hull.
- 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, Hull.