INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2001, VOL. 15, NO. 4, 287–306

Research Article

Voronoi-based region approximation for geographical information retrieval with gazetteers

HARITH ALANI

IAM, Department of ECS, University of Southampton, Highfield, Southampton, SO17 1BJ, England, UK; e-mail: ha@ecs.soton.ac.uk

CHRISTOPHER B. JONES

Department of Computer Science, Cardiff University, Cardiff CF24 3XF, Wales, UK; e-mail: c.b.jones@cs.cf.ac.uk

and DOUGLAS TUDHOPE

School of Computing, University of Glamorgan, Glamorgan, CF37 1DL, Wales, UK; e-mail: dstudhope@glam.ac.uk

(*Received 14 March 2000; accepted 10 September 2000*)

Abstract. Gazeteers and geographical thesauri can be regarded as parsimonious spatial models that associate geographical location with place names and encode some semantic relations between the names. They are of particular value in processing information retrieval requests in which the user employs place names to specify geographical context. Typically the geometric locational data in a gazetteer are confined to a simple footprint in the form of a centroid or a minimum bounding rectangle, both of which can be used to link to a map but are of limited value in determining spatial relationships. Here we describe a Voronoi diagram method for generating approximate regional extents from sets of centroids that are respectively inside and external to a region. The resulting approximations provide measures of areal extent and can be used to assist in answering geographical queries by evaluating spatial relationships such as distance, direction and common boundary length. Preliminary experimental evaluations of the method have been performed in the context of a semantic modelling system that combines the centroid data with hierarchical and adjacency relations between the associated place names.

1. Introduction

There are many contexts in which geographical location may be an important dimension for purposes of information retrieval. It is not always the case however that the conventional facilities of geographical information systems are required or necessary in order to meet the users' needs (Larson 1996, Jones *et al.* 1996, Moss *et al.* 1998). In particular it may not be necessary to produce a digital map or to employ *detailed* spatial data. Examples of types of enquiry where this may be the case (though is not necessarily so) include those that require lists of retail outlets, hotel accommodation, archaeological sites, tourist attractions or industrial sites that

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810110038942

lie inside or in the vicinity of some named place. In such enquiries the user may want to see data or documents that relate to the specified places, while from a spatial viewpoint the needs may be confined to that of having found a reasonable, perhaps approximate match with the specified place.

To service these enquiries it may be sufficient to employ a sparse, or parsimonious, representation of geographic space that combines a rich set of place name data with only limited locational data that provides an approximation of the spatial extent, or *footprint*. Data of this sort are typically referred to as gazetteers or geographical thesauri (Hill *et al.* 1999, Harpring 1997b), while the combination of the data and some associated functionality may be referred to as a geographical name server (GNS). GNS of this kind may lend themselves to limited bandwidth situations, such as in many WWW applications. When used in the context of query processing, such data enable a user-specified name to be matched with place names attached to the stored data. As with conventional information retrieval applications, the underlying information may have been indexed using terms different from those employed in a user's query. For example, a user may refer to a more specific place name, employ an archaic alternative, or simply a nearby place name. The matching process could be precise, or it could be imprecise, whereby the geographical name server was employed to find place names that were similar in the sense of being alternative versions for the same place, or of being spatially related. Determination of spatial closeness could be performed, for example, by traversing a geographical names hierarchy, to find containing places, or by using the limited locational data to find places that were within some threshold distance of the specified place. Following standard procedures for information retrieval it may be appropriate to rank the results of approximate matches in terms of their estimated closeness to the target place.

The use of gazetteers and GNS to process geographical enquiries in this way is at present in its infancy, but it can be expected to grow considerably as users of search engines raise their expectations of the level of geographical intelligence with regard to place name terminology. Research challenges in this area include those of determining the most appropriate data that should be stored in a gazetteer, the provision of support for user specification of spatial relationships in their query, and the development of semantic and spatial closeness measures that accurately reflect a user's perception of the similarity of a found place to one that they specify. A recent initiative in gazetteer design is the proposal of a content standard in association with the Alexandria Digital Library (Hill 1996, Hill *et al.* 1999). It may be regarded as pointing the way forward for future gazetteer design as well as helping to raise the profile of the potential of gazetteers in geographical information retrieval. The proposed gazetteer, which adopts a metadata approach, is very flexible with regard to the types of data that may be stored. Notably, it leaves open the possibility of various types of relationship between place names, which could be spatial as well as administrative. An important issue highlighted by Hill, and one which affects the potential of the gazetteer to assist in geographical query processing, is that of the form of the spatial footprint that should be stored with a place name.

A spatial footprint may be used to determine, or at least estimate, spatial relationships between places. These could include distance, direction, and some topological relationships such as containment and overlap. Some existing gazetteers and geographical thesauri, such as Geographic Names Information System (GNIS) (US Geological Survey 1998) and the Thesaurus of Geographic Names (TGN)

(Harpring 1997a, 1997b), only store single points for footprints and hence are restricted in their potential for estimating spatial relationships. The Alexandria Digital Library (ADL) proposals allow for other forms of footprint, ranging from a minimum bounding rectangle, at the least accurate, to what might be a relatively precise polygonal boundary. The determination of a footprint is in any event not always simple, as some place names refer to inherently imprecise regions, for which there may be no 'official' discrete boundary. In such cases, examples of which would be the 'Rockies' in America, and the 'Midlands' in the UK, it may be desirable to provide a footprint based on a common understanding of the extent of the region and the places within it.

Here we present a method for estimating spatial footprints from the locations of point sites that are known to lie inside a region and of point sites that are known to be outside. It is called the Dynamic Spatial Approximation Method (DSAM) and is based on the Voronoi diagram of the point sites. DSAM is intended to be used in combination with gazetteers and geographical thesauri that contain the co-ordinates of a centroid associated with each settlement or other named site. It is assumed that the place name data are structured hierarchically and that the spatial relationship of *meets* is encoded between the referenced regions. Thus settlements and other point-referenced sites should refer to their parent regions, of which there could be several, while regional place names refer both to their parents and to their connected neighbours. Regions for which DSAM estimates a spatial footprint may be precise administrative areas or imprecise topographic or cultural areas, and they may be current or historical. Support of historical regions necessitates the presence of temporal data associated with the place names, but this is now widely regarded as an important element of gazetteers (Copp 1997, Moss *et al.* 1998, Chappell 1999, Hill *et al.* 1999, Beard *et al.* 1997a), as indeed it is of any metadata.

In the remainder of the paper, we describe in §2 the OASIS system that we have developed for modelling place names and associated semantic data. This system is used to provide the data needed to implement DSAM. The DSAM method is described in detail in §3. Sections 4.1 and 4.2 present results of evaluating the quality of region approximation with regard to area and to visual appearance respectively. In §4.3 we present the results of some preliminary experiments to evaluate the reliability of the method for determining selected topological, directional and proximity relations. DSAM evaluation and future work are covered in §5. Conclusions are presented in §6.

2. The OASIS system

2.1. *OASIS overview*

The DSAM method was developed as an approximate spatial region representation and inference method for the OASIS (Ontologically Augmented Spatial Information System) research project. OASIS is a prototype hypermedia information system, developed to explore the potential of knowledge organisation systems in searching cultural heritage data collections. An important consideration is the provision of functionality to support queries in space and time and to do so in a way that would be global in spatial coverage. To this end, OASIS combines thematic data with a geographical thesaurus which, while containing only sparse geometric data, is relatively rich in encoding qualitative relationships between named places. Several spatial and thematic similarity measures are employed in the OASIS system to find similar items and places by imprecisely matching query terms.

2.2. *OASIS data*

The two main types of data used in this project are thematic and spatial. Thematic data includes data on museum objects and artefacts, where and when they were found or made, the material they are made of, etc. This data was taken mainly from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS), which contains a large amount of information on archaeological sites and historical buildings and monuments in Scotland (Murray 1997). This was combined with the *Art and Architecture Thesaurus* (AAT) which is a vocabulary resource developed by the J. Paul Getty Trust. It contains over 120 000 structured terms for the description of art, architecture, and material culture (Petersen and Barnett 1994). The classification schema of the OASIS system was linked to the FORTH (Foundation for Research and Technology—Hellas) implementation of the AAT which provides the necessary thematic descriptors for the RCAHMS data, such as 'bronze', 'disk', 'building', 'castle', etc.

The spatial data in the OASIS system includes information on hierarchical and adjacency relations between named places, place types, and co-ordinates. Much of this information was derived from the *Bartholomew's* (Harper Collins 2000) digital map data for Scotland, which includes place names and co-ordinates of several types of objects, such as towns, villages, airports, hills, golf courses, etc. Some data, such as place types and alternative names, was also taken from the *Getty Thesaurus for Geographic Names* database, which is a structured vocabulary of place names (Harpring 1997a). The TGN places are represented hierarchically according to the current political and physical world. A place name can be vernacular, English, or a historical name.

2.3. *OASIS classification schema*

OASIS has a rich classification schema that enables the storage of different versions of place names (e.g. current and historical names, names in different languages), place types (e.g. City, Building, Port, Hill), latitude and longitude co-ordinates, and topological relationships (e.g. overlaps, meets, part of) (figure 1). The schema is implemented using the Semantic Index System (SIS), an objected-oriented

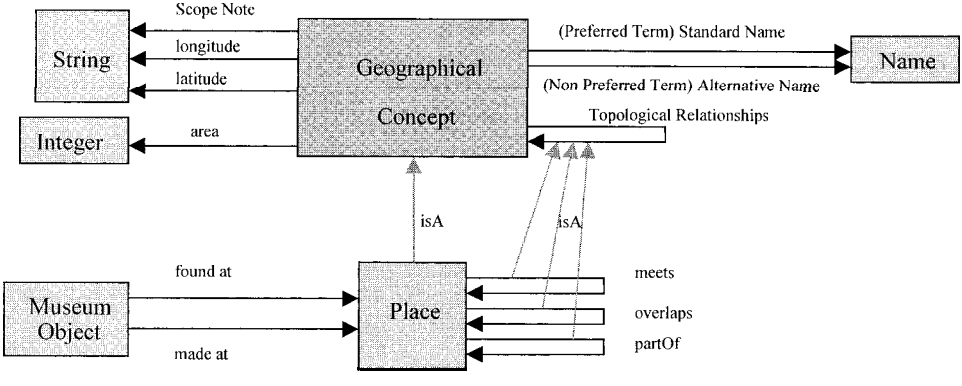


Figure 1. The classification schema of 'Place' in the OASIS system.

database system (Constantopolous and Doerr 1993, Doerr 1997) which is also used to store the data. The SIS has a meta modelling capability and an application interface for querying the schema.

The meta level classification of *Place* is shown in figure 1. Classes are represented by rectangular boxes linked to other classes via relationships represented by arrow lines. Relationships can also be instantiated or subclassed from other relationships. For example, the relationships *meets*, *overlaps*, and *partOf* are subclasses of the relationship *Topological Relationships*. *Standard Name* and *Alternative Name* are instances of the relationships *Preferred Term* and *Non Preferred Term* respectively (shown between brackets in figure 1). *Place* inherits all other relationships such as *longitude*, *latitude*, *area*, etc., from its superclass *Geographical Concept*.

Figure 2 shows the classification diagram of the City of Edinburgh. The *City of Edinburgh* is an instance of *Geopolitical Place*. The places that are part of the City of Edinburgh are represented in figure 2 by a dummy class called *MANYpartOf*, which can be opened in a separate window listing all the places that are administratively part of the region of the City of Edinburgh. The *City of Edinburgh* is also linked by *meets* relationships to the regions that it shares a boundary with. Although the *meets* relationships are symmetric, they were entered in both directions to increase the speed of search and retrieval. The information stored in the OASIS database can be accessed using a set of functions through which it is possible to find all the information related to a given place, or find all the places with specific relationships. For example to find all the places that are part of Edinburgh, the system would return a set of all the places that are linked with a *partOf* relationship pointing to Edinburgh. All relationships can be associated with dates. For example a *partOf* relationship can be linked to a certain date to indicate when the place became part of the other. This makes it possible for example to find all places that used to be part of Edinburgh during a certain period of time. This type of information is necessary in generating fuzzy and historical boundaries, as will be explained in §3.

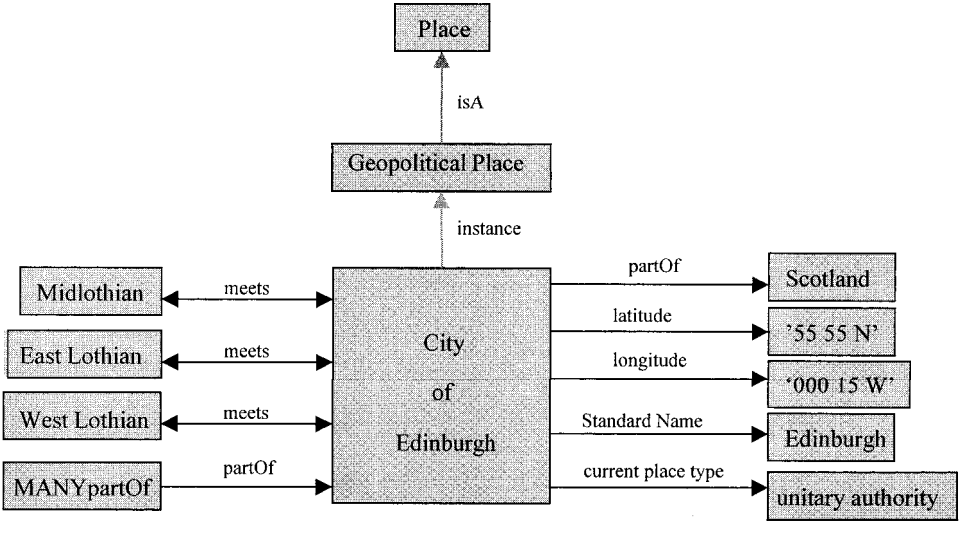


Figure 2. The classification of the City of Edinburgh.

3. The DSAM method

The Dynamic Spatial Approximation Method (DSAM) generates approximations of spatial regions based on Voronoi diagrams (figure 3). These diagrams are built from the co-ordinates of places within these regions, and of places in regions that are known to surround them. DSAM cannot be applied in its current form to mainly empty regions, or to a region that is connected to an empty region. An empty region is one that has no places to be part of it, either due to lack of data, or due to its topographical type (e.g. lake, sea, etc.). Several approaches are proposed later in this paper to widen the scope of the method's implementation.

As explained in §2.3, the places that are part of a specific spatial region can easily be found from the OASIS system via *partOf* relationships and their co-ordinates can be retrieved. Each region is linked with *meets* relationships to its surrounding regions. Hence, the co-ordinates of the places within the surrounding regions can also be retrieved. These co-ordinates are then used to construct the Voronoi diagrams that represent the boundary approximation.

Boundary approximations generated by DSAM are used to infer spatial relations between regions in the absence of digitised boundaries. Tests presented in this paper show that the inconsistency between DSAM approximations and the original boundaries can be quite low (figure 4). From the approximations, it is possible to calculate area measures, directional relationships (e.g. 10% north, 90% north-east), topological relationships of overlap, lengths of shared boundaries, and point-to-region, and region-to-region Euclidean distances.

DSAM has the capability of handling change and generating fuzzy and historical boundaries from the co-ordinates of the places that are known to be, or have been, within these boundaries. For example to get the administrative boundaries of a region at a certain date, all that is required are the places known to be part of that region at that time. The OASIS system can store and retrieve this type of information efficienctly as demonstrated in §2.3. To take a hypothetical example, assume that a user is interested in the area of South Lanarkshire, that was affected by a flooding at a specific date. The DSAM approximation for South Lanarkshire can be generated as usual, then the set of Voronoi polygons of the places that were affected by the flooding can be retrieved. The area and boundary length of the selected zone can be measured from the area and boundaries of the Voronoi polygons in the retrieved set (figure 5).

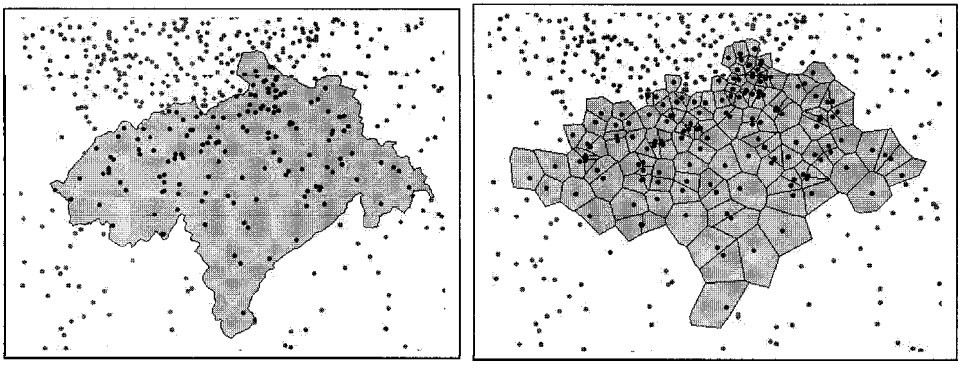


Figure 3. Midlothian region : (*a*) Actual boundaries and points, (*b*) DSAM approximation.

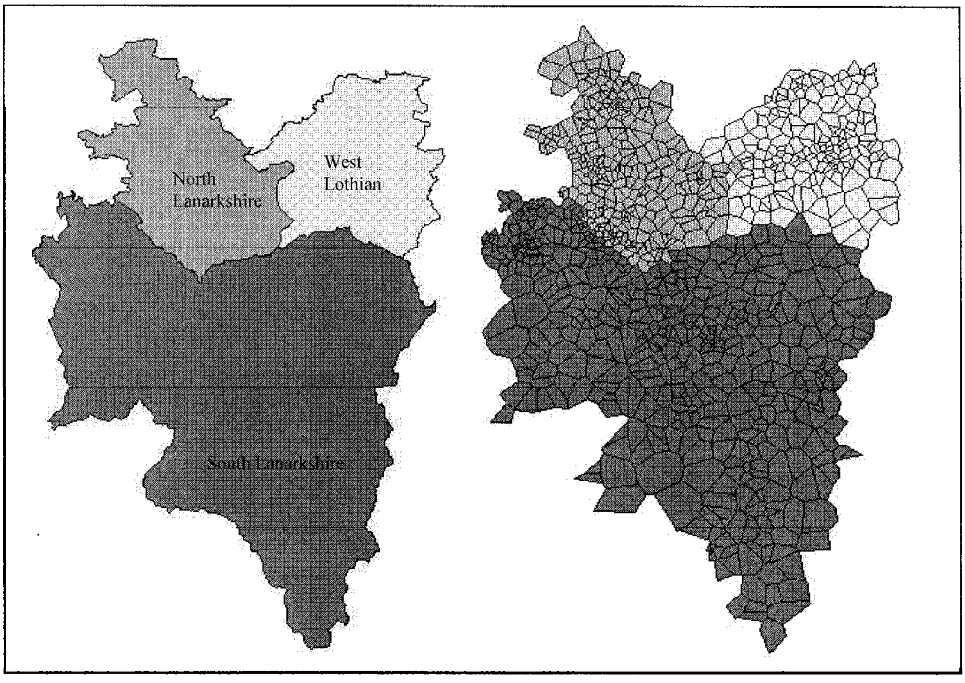


Figure 4. Three regions in Scotland. (*a*) The original administrative boundaries, (*b*) the DSAM approximation.

Having a large number of accurate place co-ordinates is important as the quality of DSAM approximations increases with the number of place co-ordinates available.

3.1. *Voronoi diagrams*

Voronoi diagrams are fundamental types of data structures in computational geometry (Aurenhammer 1991). They have important value in many fields, such as computer science, mathematics, physics, natural science, etc. The Voronoi diagram of a set of points is the partition of space around them into cells, where each cell represents the area of space that is closer to the associated point than to any other point (figure 6). The Voronoi polygon of a point is created from the perpendicular bisectors of the segments linking that point to those points closest to it.

Several versions of Voronoi algorithms have been developed for various types of spatial query (Okabe *et al.* 1992). Voronoi diagrams are valuable in defining solutions to spatial adjacency and nearest-neighbour problems (Sedgewick 1988, Aurenhammer 1991, Gold 1989, Gold 1991). Points are considered to be adjacent to each other if they share a Voronoi polygonal boundary. Having constructed a Voronoi diagram, the nearest data point to an arbitrary location may be found by determining the Voronoi polygon that contains the location.

Insertion and deletion of points to Voronoi diagrams can be implemented at run time, without the need to reconstruct the whole diagram. This assures that updated data can be available at all times, and opens new possibilities for users to experiment with the data and formulate 'what-if' type of queries (Gold 1992).

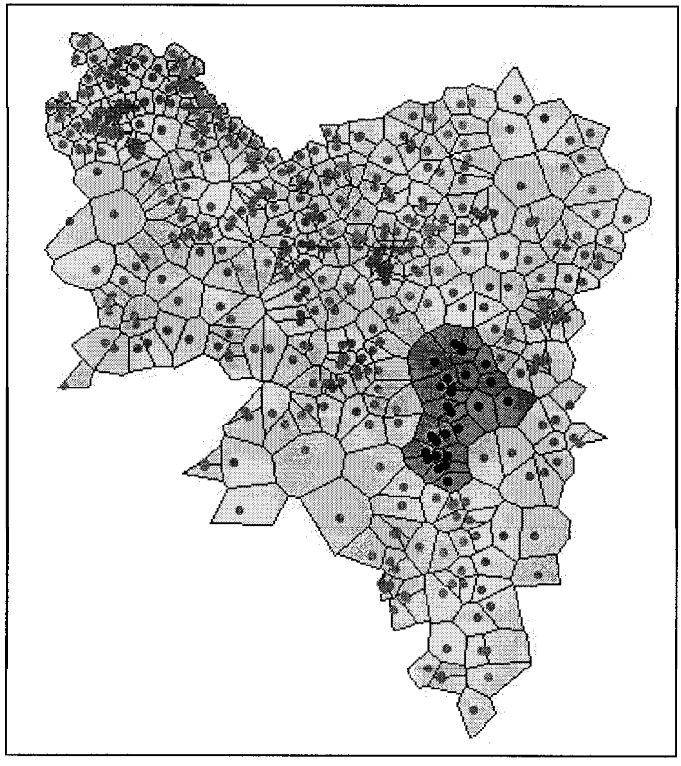


Figure 5. DSAM approximation of fuzzy boundaries.

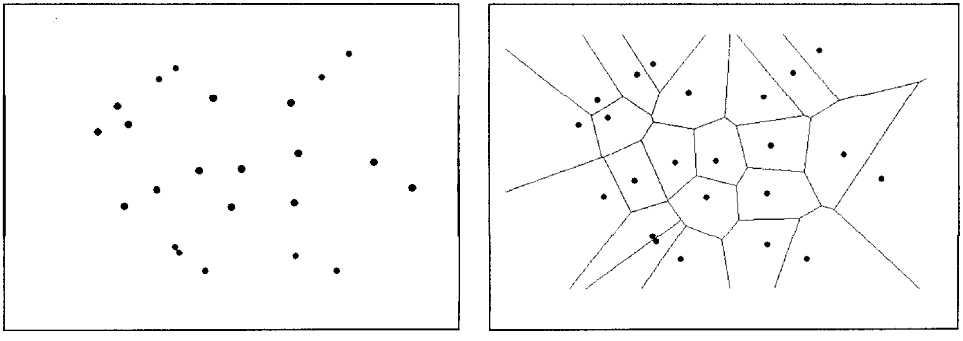


Figure 6. Voronoi diagrams for a set of points.

The Voronoi diagrams used in this paper were generated using Arc/Info's Thiessen algorithm. The research also made use of a program called Qhull (1997) which computes Voronoi diagrams, convex hulls, Delaunay triangulations, and other geometric algorithms. Areas of Voronoi polygons and their edge lengths can be calculated from the co-ordinates of the Voronoi nodes produced by Qhull. It is also possible to find the two place-points that a specific polygonal edge falls between, and vice versa. This information is used by DSAM to identify the boundary between two regions from the polygonal edges that lie between pairs of points, one from each region. Areas of regions can be approximated from the total area of Voronoi polygons

of the points that are part of that region. Other spatial information can be derived from Voronoi diagrams as will be seen later in this paper.

3.2. *Queries*

DSAM can be used to answer a variety of spatial queries, and to rank results.

1. Shape representation. '*Show me the region of Falkirk*'.
2. Shared boundary lengths between regions known to meet. '*How long is the boundary between Midlothian and West Lothian?*'. DSAM can calculate the lengths of shared boundaries and rank the regions accordingly.
3. Area of overlap between overlapping regions. '*Which hills overlap the Lothian area?*'. DSAM can rank the answer set according to the degree of overlap.
4. Directional relationships. '*What administrative regions surround Aberdeenshire from the south?*'. '*Are there any Roman sites to the west of Midlothian?*'. DSAM can calculate and rank directional relationships with regard to areal extent.
5. Generating current, historical, and imprecise boundaries. '*Show me the boundary of the Lothian region as it used to be in 1992*', '*Give me information on bronze lamps found within the 17th century borders of Edinburgh*', '*Show me the boundary of the Middle East*'.
6. Euclidean distance calculation between point and area referenced objects. '*How far is Edinburgh Castle from the Scottish Borders?*'.
7. Nearest neighbour queries. '*Which is the nearest museum?*'. DSAM is based on Voronoi diagrams, which in turn can be used to answer nearest neighbour queries.

4. Error of approximation

As mentioned earlier, to assure good quality DSAM approximations, it is necessary to have a uniformly distributed large amount of point-location co-ordinates. In the experiments that follow, the point-locational data were taken from the Bartholomew's digital map data sets for the UK and include various point-referenced sites in addition to settlements. Quality of approximation can be measured in different ways that can be categorized as follows:

1. *Total areal error*. This is the common way to measure approximation quality, which is based on comparing the area of approximation with the area of the original region.
2. *Visual error*. This is related to how different the approximated shape is from the original region's shape. This is important when representing objects graphically.
3. *Quality of spatial relationships*. This concerns the extent to which the spatial relationships between the approximated regions are consistent with the spatial relationships between the original regions.

4.1. *Total areal error*

The areal approximation error can be calculated from the difference in area measure between the approximation and its original region. As explained in the following section, the DSAM method, as does any boundary generalization method, produces negative and positive false areas. However, for areal change calculations only total areas will be compared disregarding whether the area is a positive or negative false area (figure 7). Accuracy of total area is important for queries of the sort '*which is the largest region in Scotland*', or simply '*what is the size of*

Aberdeenshire?'. To process such queries it is important that the total area of approximation is as close as possible to the area of the original region. The areal error can be calculated as follows:

$$Areal\ Error = \frac{|A_p - A_0|}{A_0} 100\% \quad (1)$$

where A_p is the total area of approximation, and A_0 is the area of the original object. Areal error of 0% means that the area of approximation equals the original area. If areal error is more than 100%, this means that the approximated area is more than double the original one.

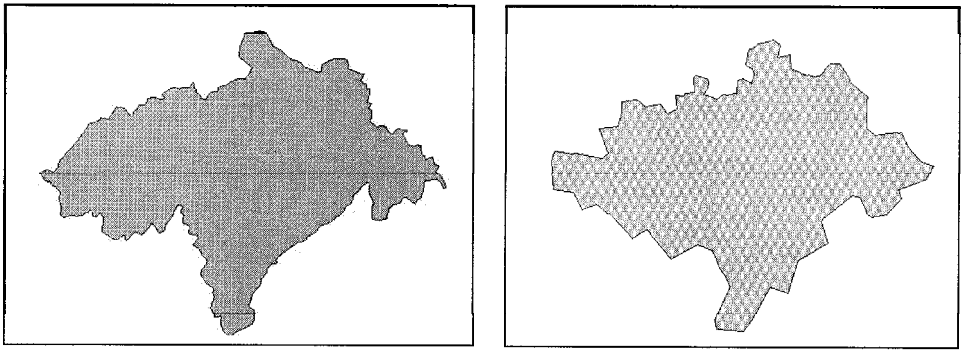
4.1.1. *Experiment with areal error calculations*

Table 1 shows the areal error calculations for five regions in Scotland using equation (1). It can be seen that the average areal error was around 2%, with almost an exact match in some cases.

4.2. *Visual error*

The visual error of approximations becomes an important issue when providing graphical displays of spatial regions. The user might like to see the location of a certain place with respect to its region. For some applications it might be sufficient and acceptable to use these approximations to represent some of the query results graphically.

The DSAM method for approximation of the boundaries of spatial regions



(a) (b)

Figure 7. Midlothian region in Scotland. (*a*) Actual administrative boundary, (*b*) DSAM approximated boundary.

Table 1. Areal change calculations.

Region	Actual area (m^2)	Approximate area (m^2)	Areal error (%)
Midlothian	357 933 800	369 804 965	3.32
City of Edinburgh	262 121 500	263 735 365	0.62
West Lothian	425 225 700	438 630 988	3.20
North Lanarkshire	469 075 300	469 353 531	0.06
South Lanarkshire	1 772 854 808	1 812 363 430	2.23
Average			1.89

normally results in negative and positive false areas. Negative false areas are the original areas that were left out in the approximation, while positive false areas are the ones that were added to the approximation, and did not belong to the original region (figure 8). Here we use the sum of these false areas (*symmetric difference*) as a simple measure of the extent to which the method retains the visual form of the original region.

The following function is used to calculate visual error:

$$Visual\ Error = \frac{A_{pp} + A_{np}}{A_0} 100\% \tag{2}$$

where A_{pp} is the positive approximated false area, A_{np} is the negative approximated false area, and A_0 is the original area. Visual error of 0% is an exact match in shape between the approximation and the original object. If visual error exceeds 100%, this means that the area of symmetric difference exceeds the original area.

4.2.1. *Experiment with visual error calculations*

The visual error was calculated for five regions in Scotland using equation (2), and the results are presented in table 2 below. The average visual error was found to be around 13%.

4.3. *Quality of spatial relationships*

For purposes of information retrieval it may be desirable to combine qualitative relations with a quantitative value that reflects the extent or degree of the relationship. This type of information is important when measuring similarities between spatial objects and ranking query results (Bruns and Egenhofer 1996, Beard and Sharma 1997b). For example, the answer to a query of the form; '*Give me information on woodland overlapping region A*' could be ranked according to the amount of overlapping area, so that woodland with most overlapping area would be placed at the top of the list. Traditional gazetteers cannot usually be used to obtain the necessary data. They are also weak in calculating Euclidean distances, adjacency, and nearest

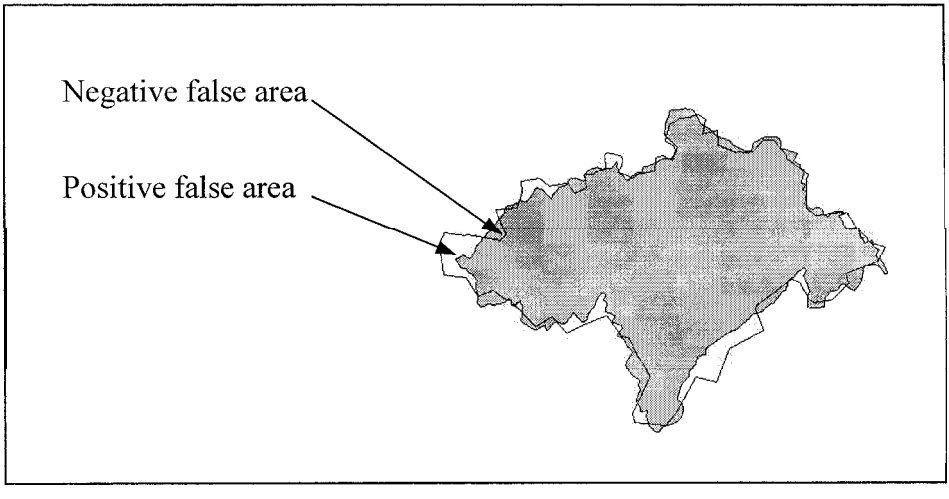


Figure 8. Midlothian. The grey region is the actual boundary, and the black line represents the DSAM approximation.

Table 2. Calculating the visual change.

Region	Actual area (m^2)	Area of symmetric difference (m^2)	Negative false area (m^2)	Visual error (%)
Midlothian	357 933 800	33 184 245	21 313 198	15.23
City of Edinburgh	262 121 500	25 004 927	23 493 275	18.50
West Lothian	425 225 700	31 154 227	17 748 877	11.50
North Lanarkshire	469 075 300	30 033 283	29 755 049	12.75
South Lanarkshire	1 772 854 808	87 229 767	47 720 807	7.61
			Average	13.12

neighbour relationships for regions, as point co-ordinate information maintained by gazetteers is not sufficient to compute such relationships. For example measuring Euclidean distances between regions from their centroid co-ordinates would produce inaccurate results if the user were interested in Euclidean distances between their boundaries.

The three main categories of spatial relationships are topological, directional, and proximity relationships (Pullar and Egenhofer 1988, Egenhofer 1991, Clementini *et al.* 1993). The following sections describe how the DSAM method can be used to derive spatial relationships using simple mathematical procedures, and present some preliminary experiments to test the quality of result.

4.3.1. *Topological relationships*

Topological relationships are related to how objects interconnect and form an important component of spatial information systems.

The six main topological relationships are *inside*, *cover*, *equal*, *meet*, *overlap*, and *disjoint* (Egenhofer 1991). DSAM approximations can be used to evaluate all of these types of relationship but the results can be expected to be more error-prone than when using full boundary representations. Examples of specific types of error are provided in §5. Figure 9 illustrates an example of the result of using DSAM approximation extents to determine overlap and adjacency respectively between pairs of regions. In

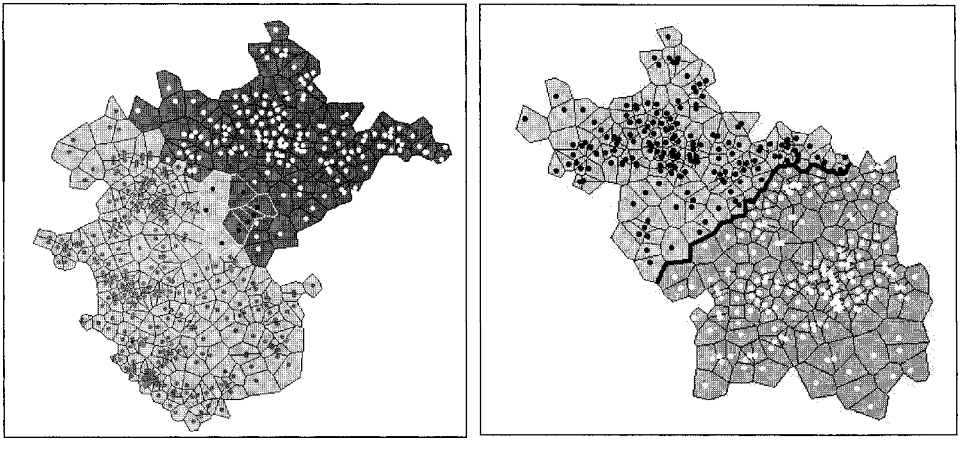


Figure 9. (*a*) Calculating overlapping area between two regions, (*b*) Measuring a shared boundary length.

figure 9(*a*) the overlap is represented by the shared Voronoi polygons, which are bounded by white lines. In the case of adjacency in figure 9(*b*) the boundary is estimated from the Voronoi polygon edges that are common to the two regions.

4.3.1.1. *Experiments to estimate boundaries between regions.* Table 3 shows the results of measuring lengths of shared boundaries between selected regions. The table compares actual boundary lengths with those calculated from DSAM approximations. The shared boundary error can be calculated as follows:

$$Boundary\ Error = \frac{|B_p - B_0|}{B_0} 100\% \tag{3}$$

where B_p is the boundary calculated from DSAM approximations, and B_0 is the original boundary length. Boundary error of 0% is an exact match between the approximated boundary and the original boundary. A boundary error of more than 100% indicates an approximated boundary of more than double the original boundary length.

Table 3. Measuring lengths of shared boundaries.

Region 1	Region 2	Actual shared boundary length (m)	Approximate shared boundary length (m)	Boundary error (%)
West Lothian	North Lanarkshire	23 958.62	19 843.47	17.18
West Lothian	South Lanarkshire	20 842.75	22 916.00	9.95
West Lothian	Scottish Borders	8366.62	8692.54	3.85
North Lanarkshire	South Lanarkshire	32 403.17	33 696.14	3.99
South Lanarkshire	Scottish Borders	55 201.04	57 375.11	3.94
Falkirk	North Lanarkshire	32 574.32	29 406.23	9.73
Midlothian	City of Edinburgh	29 732.45	34 633.31	16.48
			Average	9.30

The results presented above show that the average error of shared boundary measures is 9.3%.

4.3.2. *Directional relationships*

Directional relationships (Frank 1992, 1996) are widely used in spatial queries, and complement topological relationships in measuring spatial objects' similarity (Goyal and Egenhofer, in press). Most systems use centroid co-ordinates, Minimum Bounding Rectangles (MBRs) or some representative points as models to derive directional relationships (Goyal and Egenhofer, in press). Due to the roughness of these approximations, several alternative approaches have previously been investigated. These approaches focus on partitioning space around the MBR of the main object into 9 tiles (N, NW, NE, S, SW, SE, E, W, and 0) as shown in figure 10 (Papadias and Egenhofer 1997, Goyal and Egenhofer, in press, Papadias and Theodoridis 1997, Theodoridis *et al.* 1998), and measuring how much area of the target object falls into each tile.

A typical example of query results on directional relationships would be of the form; region A is 30% north-east of B, and 70% east of B. Figure 11 shows the DSAM approximations of two such regions; Midlothian and South Lanarkshire.

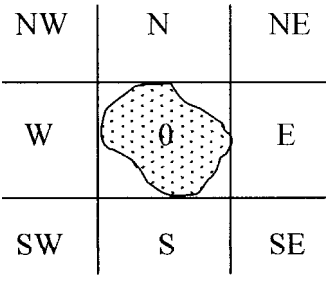


Figure 10. The nine spatial tiles for deriving directional relationships.

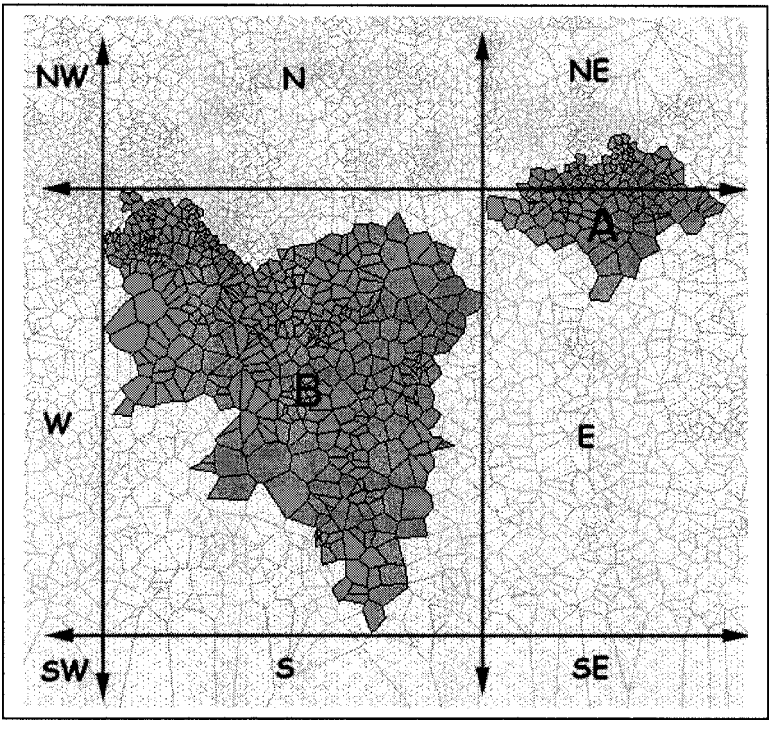


Figure 11. Deriving directional relationships from DSAM approximations.

The actual area of the original object in each tile can be approximated from the sum of the area of the Voronoi polygons in these tiles.

The following section presents an experiment on calculating directional relationships from DSAM approximations.

4.2.2.1. *Experiment with directional relationship calculations.* Directional relationships were calculated between three pairs of regions using their DSAM approximations. The amounts of Voronoi-polygonal areas from these approximations were calculated in each tile, and compared to the actual area measures. The error of results is calculated using equation 1 and presented in table 4. Results showed that the average error of directional relationships driven from DSAM approximations was around 2%.

Table 4. Calculating directional relationships.

Query object	Target object	Direction	Actual area	DSAM area	Directional error (%)
Midlothian	South	North-east	96 022 632	94 523 244	1.56
	Lanarkshire	East	261 911 168	275 281 721	5.10
North	West	North-west	31 864 180	31 057 230	2.53
Lanarkshire	Lothian	West	424 699 900	425 625 738	0.22
		South-west	12 511 233	12 670 561	1.27
South	West	West	95 025 296	92 830 618	2.31
Lanarkshire	Lothian	South-west	700 201 400	715 834 415	2.23
		South	977 543 200	1 003 688 874	2.67
				Average	2.24

4.3.3. *Proximity relationships*

Proximity relationships represent qualitative and quantitative distances between objects. An example of a quantitative query is, '*find all forests within 10 miles from the river*'. A qualitative query could include near or far relationships, '*which castles are near to the river?*'.

As mentioned earlier, Voronoi diagrams can be used to answer nearest neighbours and spatial adjacency queries. The Voronoi polygon for each point represents the area closest to that point. Therefore, nearest neighbour queries can easily be answered by identifying the polygon that the query place falls in. For example to find the nearest castle to a specific location, all that is needed is to identify which castle-Voronoi-polygon the location falls in.

Distance calculations form an important part of spatial information systems (Laurini and Thompson 1994). Euclidean distances between point-referenced objects can easily be calculated from their co-ordinates. However, Euclidean distances between area-referenced objects, or between point and area-referenced objects, are more difficult to calculate. An area-referenced object, such as a region, has a boundary as well as a centroid-point. Measuring Euclidean distances to the centre of the region is not always desired. Consider the query, '*How far is this place from the boundary of region A?*', or '*Show me all castles within 20 miles of region B*'. To process such queries accurately, it is important to measure the distance to the boundaries, rather than to the regions' centres. Euclidean distances between boundaries can be calculated from DSAM approximations. Shortest Euclidean distances between point and area objects can be measured between the point's co-ordinates and the closest polygonal boundary point of the DSAM approximation for the area object. In the same way it is possible to measure the shortest Euclidean distance between two area objects from the distance between their closest pair of polygonal boundary points.

The following section presents an experiment with Euclidean distance calculations between point and area referenced places.

4.3.3.1. *Experiment with Euclidean distance calculation.* The shortest Euclidean distances were measured from six randomly selected towns to the boundary of North Lanarkshire. The distances were first measured to the actual boundary of North Lanarkshire, then compared to the distance values measured to the DSAM approximated boundary (figure 12). The measurement using the DSAM approximated

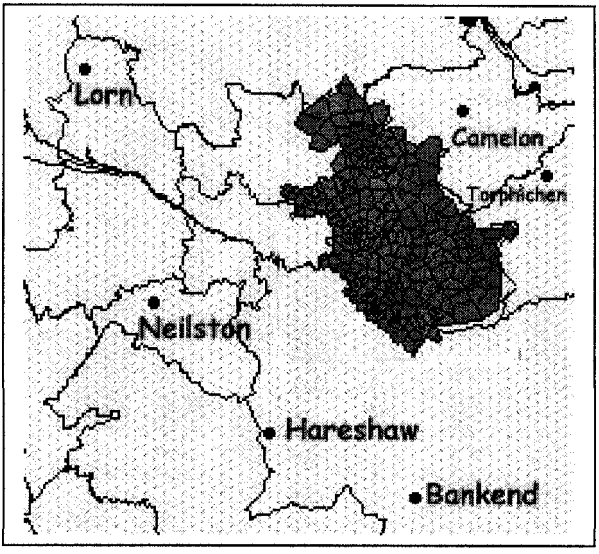


Figure 12. Calculating Euclidean distances to DSAM approximations.

boundary was subject to the simplification of measuring distances from the given points to the nearest node of the DSAM boundary.

Euclidean distance error is calculated using an equation of the same form as equation (3). From the table below, it can be seen that Euclidean distances measured to the DSAM approximated boundary are of good quality, with an average error of around 4%. DSAM quality in measuring Euclidean distances is expected to increase with the increase in distance. This is because the further the point location is from the target region, the less is the effect of the boundary approximation error on the Euclidean distance measure.

5. Evaluation

DSAM approximation quality depends on the amount and quality of data available, in the form of places and co-ordinates. In general, DSAM approximations could produce three types of error:

1. Qualitative errors, when inferring topological relationships.
2. Measurement errors, when calculating area and boundary lengths.
3. Statistical errors, when ranking query results.

Table 5. Calculating Euclidean distances.

Town	Actual distance (m)	DSAM distance	Distance error (%)
Lorn	27 327	26 751	2.11
Camelon	5347	4715	11.82
Torphichen	7008	7307	4.27
Neilston	20 253	20 332	0.39
Hareshaw	18 513	18 168	1.86
Average			4.09

Qualitative errors might occur when inferring topological relationships from DSAM approximations. Some topological relationships such as *partOf* and *meets* were retrieved from the *Bartholomew's* data and stored explicitly in the OASIS database. DSAM relies on *partOf* relationships in constructing Voronoi diagrams for region approximation, and on *meets* relationships to decide whether two regions are connected or disjoint.

If little data is available, or it is of low quality, then inconsistencies may arise between the spatial relationships inferred from DSAM, and those existing between the original regions. Figure 13 gives an example of such incompatibility, where the DSAM approximations for Midlothian and West Lothian share a boundary, while the actual regions are disjoint. This is due to the lack of points in our dataset in the parts of the City of Edinburgh and the Scottish Borders that separate Midlothian from West Lothian. The result of this error falls in two parts. First, the approximations of two very close, but separate regions (West Lothian and Midlothian) became connected and shared a boundary. Secondly, the approximations of two slightly connected regions (Scottish Borders and the City of Edinburgh) became separate. The first part of the problem may be solved by making use of the *meets* relationship to decide whether two regions are connected or not, rather than trying to infer this information from their approximations, which will be a more error-prone and computationally intensive process. The second part of the problem is rather more difficult to solve, due to the loss of information involved. A zero length of shared boundary between two regions linked with a *meets* relationship is a sign of inconsistency, which an integrity checking procedure could detect and request a manual intervention to insert notional points for purposes of DSAM processing.

Calculating area and boundary lengths from DSAM region approximations is subject to measurement errors. Such calculations are affected by the amount of false area involved in the approximations.

Ranking can introduce statistical error, where the ranks of query results may differ slightly from the ranks of the original values. For example in the experiment results presented in table 3, although the approximation quality was good, the ranking was not totally accurate. Consider the first two records. The actual shared boundary length in the first record is longer than in the second record. However, the approximated shared boundary in the second record turned out to be longer than the first one. This occurs where the actual boundary lengths to be approximated are similar. For example, in the first two records, the difference between the actual

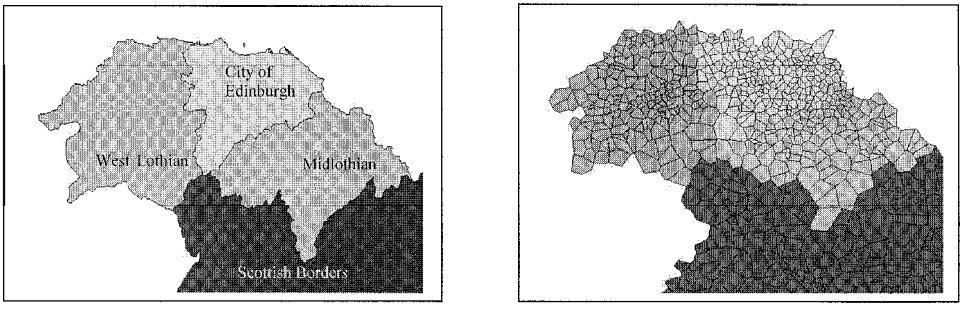


Figure 13. Change in shared boundaries: (*a*) actual boundaries, (*b*) approximated boundaries.

shared boundary lengths was less than 15%. In such cases, a slight negative or positive change in their approximated lengths could change the ranking. However, an overall comparison between the ranks of the approximated and original values in table 3 yields a Spearman coefficient of 0.822 (where 1.0 means an exact match between the two ranks). This indicates that ranking is of good quality overall. Ranking quality should improve even more when dealing with larger sets of results.

At present, DSAM is suitable for approximating inland regions, surrounded by other regions. This is because the approximation is built from the Voronoi diagrams of the co-ordinates of the places within these regions. Therefore, if the region to be approximated is connected to a sea on the one side, for example, then the obvious lack of place co-ordinates in the sea region results in a high degree of inaccuracy in the boundary approximation of that particular side. Incomplete Voronoi polygons and dangling segments normally occur when co-ordinates are available from one or more sides of a region to be approximate. Such cases can be flagged by most Voronoi programs.

6. Conclusions and future work

Gazetteers use sparse spatial databases in which geographical place names are associated with a spatial footprint. Usually the footprint is confined to a centroid or minimum bounding rectangle and does not include boundary data. As a consequence gazetteers have conventionally been of limited use in answering queries on spatial relationships and boundaries. DSAM provides an effective way to approximate the extent of spatial regions and derive their spatial relationships using sets of co-ordinates in association with place name hierarchies and region adjacency data. Traditional spatial information systems often face difficulties in storing and representing historical and imprecise boundaries. DSAM is capable of approximating boundaries of precise and imprecise regions with respect to time.

The DSAM method uses a set of functions based on Voronoi diagrams to estimate spatial relationships between named places represented by their centroid co-ordinates. Tests presented in this paper showed that the DSAM method gives good results for measures of area, shared boundary lengths between regions that are known to meet, area of overlap between overlapping regions, Euclidean distances and directional relationships. In doing so they provide good approximations of current, historical, and imprecise boundaries, and help in answering nearest neighbour queries.

As regards limitations, the DSAM method presented cannot be used reliably for determining topological relationships, due to its sensitivity to the configuration of the source points. While the method is intended to be used in combination with supplied topological relationships, problems can occur when attempting to make measurements of associated boundaries. However there is clearly scope for extending the method to ensure consistency between DSAM-determined relationships and the supplied topological relationships. This could be done by inserting appropriately located dummy points. There is also the problem of unbounded DSAM regions due to the presence of real-world regions that are not represented by centroid data, a prime example being the sea. Possible solutions include the insertion of dummy points in the unrepresented region, or boundary approximation based on the convex hull of the centroids of the represented region. It is important to note that the accuracy of a DSAM region approximation depends upon the density and distribution of the centroid data. Future work will investigate the multiresolution nature of DSAM and its relationship to depth of descent within place name hierarchies.

Acknowledgements

We would like to acknowledge assistance provided to this project by the J. Paul Getty Trust, the FORTH Institute of Computer Science, and the Royal Commission on the Ancient and Historical Monuments of Scotland.

References

Aurenhammer, F., 1991, Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys*, **23**, 345–405.

Beard, K., Smith, T., and Hill, L., 1997a, Meta-information models for georeferenced digital library collections. In *Proceedings of the 2nd IEEE International Metadata Conference* (Silver Spring, MD), http://computer.org/proceedings/meta97/papers/kbeard/kbeard.html

Beard, K., and Sharma, V., 1997b, Multidimensional ranking for data in digital spatial libraries. *International Journal on Digital Libraries*, **1**, 153–160.

Bruns, T. H., and Egenhofer, M. J., 1996, Similarity of Spatial Scenes. In *Proceedings of the 7th International Symposium on Spatial Data Handling*, Delft, The Netherlands (London: Taylor and Francis), pp. 31–42.

Chappell, C., 1999, Changing boundaries: gazetteers, information retrieval and data browsing. ASSIST Quarterly, **23**, 19–21.

Clementini, E., Di Felice, P., and van Oosterom, P., 1993, A small set of formal topological relationships suitable for end-user interaction. In *Advances in Spatial Databases: Proceedings of the 3rd International Symposium, SSD'93* (Singapore), Lecture Notes in Computer Science, **692**, pp. 277–295.

Constantopolous, P., and Doerr, M., 1993, The Semantic Index System—a brief presentation. Institute of Computer Science Technical Report. FORTH-Hellas, Gr-71110 Heraklion, Crete.

Copp, C. J. T., 1997, The JNCC Electronic Dictionary of Administrative Areas. *MDA information*, **2**, 13–19.

Doerr, M., 1997, Reference information acquisition and co-ordination. In *Proceedings of the 60th Annual Meeting of the American Society for Information Science*, 34 (C. Shwartz, M. Rorvig eds), Medford, NJ: Information Today, pp. 295–312.

Egenhofer, M. J., 1991, Point-set topological relations. *International Journal of Geographical Information Science*, **5**, 161–174.

Frank, A. U., 1992, Qualitative spatial reasoning about distances and directions in geographic space. *Journal of Visual Languages and Computing*, **3**, 343–371.

Frank, A. U., 1996, Qualitative spatial reasoning: cardinal directions as an example. *International Journal of Geographical Science*, **10**, 269–290.

Gold, C. M., 1989, Spatial interpolation, spatial adjacency and GIS. In *Three Dimensional Applications in Geographic Information Systems*, edited by J. Raper (London: Taylor and Francis), pp. 21–35.

Gold, C. M., 1991, Problems with Handling Spatial Data—the Voronoi Approach. *CISM Journal*, **45**, 65–80.

Gold, C. M., 1992, Dynamic Spatial Data Structures—the Voronoi Approach. In *Proceedings of the Canadian Conference on GIS* (Ottawa: CIG and ISPRS), pp. 245–255.

Goyal, R. K., and Egenhofer, M. J., in press, Cardinal directions between extended spatial objects. *IEEE Transactions on Knowledge and Data Engineering*.

Harper Collins, 2000, *Bartholomew*. http://www.bartholomewmaps.com.

Harpring, P., 1997a, The limits of the world: theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. In *Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, ICHIM'97*, Archives and Museum Informatics (Paris: Le Louvre), pp. 237–251.

Harpring, P., 1997b, Proper words in proper places: the thesaurus of geographic names. *MDA Information*, **2**, 5–12.

Hill, L. L., 1996, Proposal for developing an Internet-accessible gazetteer, based on a new content standard and open contribution of gazetteer data. http://www.alexandria.ucsb.edu/~lhill/alex-imp/gazprop6.htm

Hill, L. L., Frew, J., and Zheng, Q., 1999, Geographic names. The implementation of a gazetteer in a georeferenced digital library. *Digital Library*, **5**(1). www.dlib.org/dlib/january99/hill/01hill.html

Jones, C. B., Taylor, C., Tudhope, D., and Beynon-Davies, P., 1996, Conceptual, spatial and temporal referencing of multimedia objects. In *Advances in GIS Research II, Proceedings 7th International Symposium on Spatial Data Handling* (London: Taylor and Francis), **2**, pp. 13–26.

Larson, R. R., 1996, Geographic information retrieval and spatial browsing. *GIS and Libraries: 32nd Annual Clinic on Library Applications of Data Processing Conference* (Urban-Champaign: University of Illinois), pp. 81–124.

Laurini, R., and Thompson, D., 1994, *Fundamentals of Spatial Information Systems* (London: Academic Press Limited).

Moss, A., Junbg, E., and Petch, J., 1998, The construction of WWW-based gazetteers using thesaurus techniques. In *Proceedings of the 8th International Symposium on Spatial Data Handling (SDH'98)* (Vancouver, Canada: International Geographical Union), pp. 65–75.

Murray, D., 1997, GIS in RCAHMS. *MDA information*, **2**, 35–38.

Okabe, A., Boots, B., and Sugihara, K., 1992, *Spatial Tessellations—Concepts and Applications of Voronoi Diagrams* (Chichester: Wiley).

Papadias, D., and Egenhofer, M. J., 1997, Algorithms for hierarchical spatial reasoning. *Geoinformatica*, **1**, 251–273.

Papadias, D., and Theodoridis, Y., 1997, Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographic Information Science*, **11**, 111–138.

Petersen, T., and Barnett, P. J., 1994, *Guide to Index and Cataloging with the Art and Architecture Thesaurus* (Oxford: Oxford University Press).

Pullar, D. V., and Egenhofer, M. J., 1988, Towards formal definitions of topological relations among spatial objects. In *Proceedings of the 3rd International Symposium on Spatial Data Handling* (Sydney: IGU), pp. 225–243.

Qhull, 1997, http://www.geom.umn.edu/software/qhull/.

Sedgewick, R., 1988, *Algorithms* (New York: Addison-Wesley).

Theodoridis, Y., Papadias, D., Stefanakis, E., and Sellis, T., 1998, Direction relations and two-dimensional range queries: optimization techniques. *Data and Knowledge Engineering*, **27**, 313–336.

US Geological Survey, 1998, Geographic Names Information System (GNIS), http://mapping.usgs.gov/www/gnis/