Published paper
Bennett, B., Mallenby, D. and Third, A. (2008) *An Ontology for Grounding Vague Geographic Terms.* In: Eschenbach, C. and Grüninger, M., (eds.) Formal Ontology in Information Systems - Proceedings of the Fifth International Conference. FOIS-08, Oct 31st - Nov 3rd 2008, Saarbrücken, Germany. Frontiers in Artificial Intelligence and Applications (183). IOS Press , pp. 280-293.
http://dx.doi.org/10.3233/978-1-58603-923-3-280

An Ontology for Grounding Vague Geographic Terms

Brandon BENNETT, David MALLENBYand Allan THIRD
School of Computing,
University of Leeds, UK
e-mail: {davidm,brandon,thirda}@comp.leeds.ac.uk

Abstract. Many geographic terms, such as “river” and “lake”, are vague, with no clear boundaries of application. In particular, the spatial extent of such features is often vaguely carved out of a continuously varying observable domain. We present a means of defining vague terms using *standpoint semantics*, a refinement of the philosophical idea of supervaluation semantics. Such definitions can be grounded in actual data by geometric analysis and segmentation of the data set. The issues raised by this process with regard to the nature of boundaries and domains of logical quantification are discussed. We describe a prototype implementation of a system capable of segmenting attributed polygon data into geographically significant regions and evaluating queries involving vague geographic feature terms.

Keywords. Vagueness, Geographic Entities, Query Answering

1. Introduction

In recent years increasing attention has been paid to the ontology of geographic entities. A major motivation for this has been the recognition that the implementation of computational Geographic Information Systems (GIS) which can support functionality for sophisticated data manipulation, querying and display requires robust and detailed specification of the semantics of geographic entities and relationships. A second, more philosophical, motivation for attention to this domain is that it presents a concrete manifestation of many ontological subtleties. For instance issues of individuation, identity and vagueness arise in abundance, when one tries to give precise specifications of the meanings implicit in geographic terminology [1,2,3,4].

Our concerns in this paper will relate to both these motivations. On the one hand, we will examine the particular ontological issues associated with interpretation of vague geographic feature terms (especially hydrological terms such as ‘lake’ and ‘river) and will outline how the general semantic framework of *standpoint semantics* can be applied to provide a framework within which such vagueness can be represented explicitly. We shall also see that when deployed in conjunction with a geometry-based theory of feature segmentation, this semantics gives an account of how vague features are individuated with respect to the material structure of the world. On the other hand, we shall also be very much concerned with the implementation of certain GIS functionality for which a coherent theory of vagueness and its relation to individuation is a necessary pre-requisite.

We look specifically at the problem of interpreting logical queries involving vague predicates with respect to a geographic dataset. We shall assume that such data takes a typical form consisting of a set of 2-dimensional polygons, each of which is associated with one or more labels describing the type of region that the polygon represents. This is a simplification of geographic data in general, which will often include other types of information such as point or line entities, altitudes, additional cartographic entities such as icons or textual strings and meta-annotations relating to the provenance or accuracy of data items. Moreover, the data will not normally consist simply of a set of entities but a complex data structure supporting indexing and various kinds of computational manipulation of data elements. Nevertheless, labelled 2-dimensional polygons form the core of most real geographic information systems.

The structure of the rest of the paper is as follows. In section 2 we present an overview of the basic theory of standpoint semantics, which is a refinement of supervaluation semantics. Section 3 considers the ontological principles that govern the ways in which one can divide up the geographic realm into distinct regions corresponding to geographic features. In section 4 we consider the implementation of a geographic query interpretation system and see that severe difficulties arise regarding finding an appropriate computationally tractable domain of quantification. We shall see that finding a solution to this problem requires a theory of individuation (such as was developed in section 3). Section 5 then looks in detail at the implementation of our prototype system, which provides a limited proof of concept of our theoretical analysis. Finally, concluding remarks and discussion of future work are given in section 6.

2. Standpoint Semantics

In making an assertion or a coherent series of assertions, one is taking a *standpoint* regarding the applicability of linguistic expressions to describing the world. Such a standpoint depends partly on one's beliefs about the world. This epistemic component will *not* be considered in the current paper: we shall assume for present purposes that one has correct knowledge of the world — albeit at a certain level of granularity (which in the context of geographic information is likely to be rather coarse). The other main ingredient of a standpoint, which we *will* be concerned with here, is that it involves a linguistic judgement about the criteria of applicability of words to a particular situation. This is especially so when some of the words involved are vague. For instance, one might take the standpoint that a certain body of water should be described as a 'lake', whereas another smaller water-body should be described as a 'pond'.

The notion of 'standpoint' is central to our analysis of vagueness. Vagueness is sometimes discussed in terms of different people having conflicting opinions about the use of a term. This is somewhat misleading since even a person thinking privately may be aware that an attribution is not clear cut. Hence a person may change their standpoint. Moreover this is not necessarily because they think they were mistaken. It can just be that they come to the view that a different standpoint might be more useful for communication purposes. Different standpoints may be appropriate in different circumstances. The core of standpoint semantics does not explain why a person may hold a particular standpoint or the reasons for differences or changes of standpoint, although a more elaborate theory dealing with these issues could be built upon the basic formalism.

In taking a standpoint, one is making somewhat arbitrary choices relating to the limits of applicability of natural language terminology. But a key feature of the theory is that all assertions made in the context of a given standpoint must be mutually consistent in their use of terminology. Hence, if I take a standpoint in which I consider Tom to be tall, then if Jim is greater in height than Tom then (under the assumption that height is the only attribute relevant to tallness) I must also agree with the claim that Jim is tall.

Our *standpoint semantics* is both a refinement and an extension of the *supervaluation* theory of vagueness that has received considerable attention in the philosophical literature (originating with [5]). Supervaluation semantics enables a vague language to be logically interpreted by a set of possible precise interpretations (*precisifications*). This provides a very general framework within which vagueness can be analysed within a formal representation. Here we do not have space to give a full account of supervaluation semantics. Detailed expositions can be found in the philosophical literature (e.g. [6]).

By itself, supervaluation semantics simply models vagueness in terms of an abstract set of possible interpretations, but gives no analysis of the particular modes of semantic variability that occur in the meanings of natural language vocabulary. A key idea of standpoint semantics is that the range of possible precisifications of a vague language can be described by a (finite) number of relevant *parameters* relating to objectively observable properties; and the limitations on applicability of vocabulary according to a particular standpoint can be modelled by a set of *threshold values*, that are assigned to these parameters. To take a simple example, if the language contains a predicate Tall (as applicable to humans), then a relevant observable is 'height'. And to determine a precisification of Tall we would have to assign a particular threshold value to a parameter, which could be called tall_human_min_height.[1] In general a predicate can be dependent on threshold valuations of several different parameters (e.g. Lake might depend on both its area and some parameter constraining its shape.) Thus, rather than trying to identify a single measure by which the applicability of a predicate may be judged, we allow multiple vague criteria to be considered independently.

In the current paper (as in several previous papers on this topic [4,8,9]) we shall assume that standpoints can be given a model theoretic semantics by associating each standpoint with a threshold valuation. In so far as standpoints may be identified with an aspect of a cognitive state, this idea is perhaps simplistic. It is implausible that an agent would ever be committed to any completely precise value for a threshold demarcating the range of applicability of a vague predicate. Cognitive standpoints are more plausibly associated with constraints on a range of possible threshold values (e.g. if I call someone tall then my claim implies an upper bound on what I consider to be a suitable threshold for tallness — the threshold cannot be higher than the height of that person) rather than exact valuations of thresholds.[2] But in the context of cartographic displays, we may more plausibly propose that any useful depiction of geographic entities corresponding to geographic terms should be determined by application of precise criteria associated

[1] Vague adjectives tend to be context sensitive in that an appropriate threshold value depends on the category of things to which the adjective is applied. This is an important aspect of the semantics of vague terminology but is a side issue in relation to our main concerns in the current paper. Here we shall assume that vague properties are applied uniformly over the set of things to which they can be applied. To make this explicit we could always use separate properties such as Tall-Human and Tall-Giraffe, although we won't actually need to do this for present purposes. A formal treatment of category dependent vague adjectives is given in [7].

[2] This elaboration of the status of standpoints in relation to thresholds is being developed in a separate strand of research.

with the term, and that such criteria require a definite value to be associated with every threshold parameter.

A *threshold valuation* appropriate for specifying a standpoint in relation to the domain of hydrographic geography might be represented by:

$$V = [\mathsf{pond\_vs\_lake\_area\_threshold} = 200m^2,\ \mathsf{river\_min\_linearity\_ratio} = 3,\ ...]$$

Here one parameter determines a cut-off between ponds and lakes in terms of their surface area and another fixes a parameter indicating a linearity[3] requirement used to characterise rivers.

3. Geographic Entities and their Boundaries

As noted by Smith and Mark in [3], the geographic domain is distinctive in that typical geographic objects are attached to the world and are not easily demarcated in the way that physically detached objects such as organisms and artifacts can be. Thus the individuation of geographic features is ontologically problematic. Previously, Smith [10,11] had drawn attention to a distinction between of *bona fide* and *fiat* boundaries:

> *Fiat boundaries are boundaries which owe their existence to acts of human decision or fiat, to laws or political decrees, or to related human cognitive phenomena. Fiat boundaries are ontologically dependent upon human fiat. Bona fide boundaries are all other boundaries.* [11]

A paradigm case of a fiat boundary is the border of a country whose location does not depend on any physical boundary in the world.[4] In [3] it is argued that, in so far as they may be said to exist at all, the boundaries of mountains must be fiat because they rely on human judgement for their demarcation. Whilst we have no objection to this use of terminology, we believe that there is a significant difference between the national border and mountain cases. Although any particular demarcation of a border around a mountain is certainly dependent on human judgement, the range of reasonable boundaries is also to a large extent determined by the lie of the land.

In order to understand this distinction more clearly, it will be instructive first to consider another kind of boundary, which we call an *implicit geometrical boundary*. Such

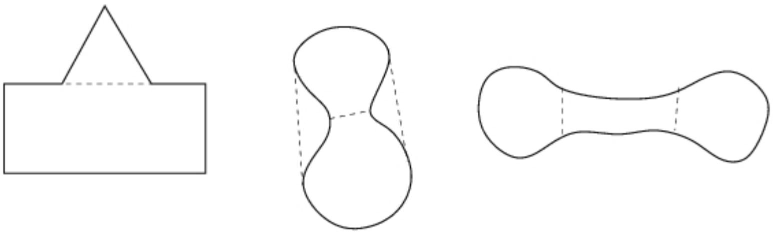

Figure 1. Implicit geometric boundaries.

[3] Note that we use the term ‘linearity’ to refer to elongation of form rather than straightness. Thus we would describe a river as linear, even though it may bend and wiggle. A geometric characterisation of linearity of this form has been presented in previous papers [8].

[4] Of course particular national boundaries may be aligned to physical boundaries such as the banks of rivers but this is a contingent circumstance.

a boundary does not lie upon an actual discontinuity in the fabric of the world but follows a line that is determined by the spatial configuration of other boundaries, which may be either bona fide or fiat (or a combination of both). Such boundaries are depicted in Figure 1. On the left we see a region within which there is an implicit boundary between a rectangular portion and a triangular projection. The middle region involves a 'neck' flanked by concavities, and these features also imply certain geometric boundaries.

In the region on the right, implicit boundaries are not so clear cut. In describing the region one may be inclined to mention two bulbous parts joined by an elongated section. This suggests the existence of implied boundaries between these three portions. These are examples of *vague boundaries* whose course is hinted at, but not completely determined, by the geometric form of a concrete boundary.

This analysis suggests a four-fold classification of kinds of boundary:

- *Bona fide* boundaries between matter or terrain types.
- *Fiat* boundaries imposed on the world by conscious agents
- *Implicit Geometrical* boundaries determined geometrically in relation to bona fide and/or fiat boundaries.
- *Vague* boundaries, which can be made precise in relation to some standpoint taken on an appropriate precisification of vague properties or relations. The resulting precise properties/relations will then determine a geometrical boundary (which will be demarcated in relation to bona fide and/or fiat boundaries).

The latter two types could be regarded as special cases of *bona fide* or *fiat* boundaries. However, it is not completely clear to which camp they should be assigned. Whether implicit geometric boundaries are considered *bona fide* or *fiat* depends upon whether one takes a Platonist or constructivist view of the existence of geometrical entities. It may be argued that vague boundaries must involve an element of human judgement and hence must be *fiat*. However, if one takes a Platonist view of implied geometric boundaries, then vague boundaries also have a *bona fide* underpinning.

Meta-terminological confusion notwithstanding, it is clear that many kinds of natural geographic feature have vague boundaries and that the demarcation of these is determined by a combination of physical properties of the world and human judgement. We believe that the way that this occurs can be explained by standpoint semantics.

This is well illustrated by consideration of the division of a water system into lakes and rivers. As described in [8,9], such a segmentation can be achieved by specifying geometric predicates that can identify linear/elongated *stretches* of a water system (as represented by polygons) and distinguish these from from expansive (lake-like) regions of the system. Indeed these have been implemented in prototype GIS software (GEOLOG). A feature of these predicates is that they depend on a small number[5] of parameters, for which specific values must be chosen to obtain a segmentation into lakes and rivers. This parameterised variability of geometry-based predicates can be directly described within the framework of standpoint semantics. Each choice of parameters given to the computational segmentation procedure corresponds to a standpoint taken with respect to the interpretation of the terms 'river' and 'lake'.

Of course more factors are relevant to the meanings of these natural language terms; so this shape-based characterisation is only part of a full explanation of the usage of

[5] In our simplest implementation there is just one such parameter, but better results have been obtained by adding a second parameter.

hydrographic terms. For instance, water flow is such an essential part of our concept of river that it might appear that no satisfactory characterisation of rivers could omit this aspect. But, GIS and other cartographic data rarely includes flow information (such information is hard to obtain and to depict); and yet, it seems that humans usually have little difficulty in identifying rivers represented in a 2-dimensional map display. One explanation for this is that, although flow is an important criterion in its own right, the dynamic behaviour of water distributed over an uneven but approximately horizontal surface is closely correlated (due to physical laws) with the geometry of the projection of the water system onto the horizontal plane. Thus, given our knowledge of the way the world works, we can infer a lot about flow just from a 2-dimensional representation of a water system.

Having said that, we would in future like to incorporate flow into our hydrographic ontology and believe that can be done within the general framework that we propose. A simple approach would be to take a field of flow vectors (this would have to be interpolated from some set of data points) and segment the water system according to a threshold on flow magnitude, so that we would obtain polygons labelled as either flowing or (comparatively) still. We could then define types of hydrographic feature in terms of a combination of both shape-based and flow-based characteristics. (We could also investigate correlations between the two types of characteristic.)

In many cases there is ambiguity with regard to which objective properties are relevant to a particular natural language term (e.g. is salinity relevant to lake-hood). Such controversy may be modelled by allowing standpoints to vary not-only in respect of threshold parameter values but also in the assignment of definitions to terms. Although this is clearly an important issue, it will not be considered in the present paper.

3.1. Land Cover Types and their Extensions

As well as by referring to geographic features, the geographic domain is very often described in terms of its terrain or land cover. A region may be wooded, ice covered, rocky etc.. In some cases the boundaries of such regions may be clearly *bona fide*, whereas in others, especially where there is a transitional region between terrain types (e.g. jungle ↔ scrub-land ↔ desert), the boundary may be vague. In either case there is certainly a physical basis to land cover demarcations; and in the case where the boundary is vague, the range of reasonable demarcations can be modelled within standpoint semantics in terms of thresholds on appropriate parameters relating to properties of the Earth's surface.

However, apart from such vagueness, there is another characteristic of land cover that is potentially problematic for computational manipulation of geographic data. Land cover types are *downward inherited*, meaning that, if a region is covered by a given type of terrain, then all sub-regions are also covered by this terrain type.[6,7] It is also clear that, if we have a set of regions all covered by the same terrain type, then the mereological sum of these regions is also covered by that type. Both these conditions are entailed by

[6] This kind of inheritance of properties among spatial regions is discussed in detail in [12].

[7] In fact downward inheritance will not normally apply beyond a certain fineness of granularity, but for present purposes we shall ignore this complication and assume that we do not have to worry about fine grained dissections of the world.

the following equivalence, which applies to properties that may be said to be manifest homogeneously over extended regions of space:[8]

TT-hom) $\mathsf{HasTerrainType}(r, t) \leftrightarrow \forall r'[\mathsf{P}(r', r) \rightarrow \mathsf{HasTerrainType}(r', t)]$

With regard to computational manipulation of land cover information this homogeneity property has both positive and negative implications. On the negative side it suggests that if a GIS ontology includes land cover terms that can be predicated of arbitrary regions of geographic space, then the set of regions that can instantiate such predicates, must include arbitrary sub-regions of its base polygons. For example, if the ontology includes a predicate $\mathsf{Water}(r)$, meaning that r is completely covered with water then this will be satisfied by arbitrary dissections (and unions) of those data polygons labelled with the 'water' attribute.[9]

But on the positive side it is clear that one would never want to actually exhibit all water-covered polygons. Once we give the total extent of a given terrain type we can simply exhibit this, and the fact that all its sub-regions also have that type is implicit. It is obvious to a GIS user that an extended region of blue represents water and moreover that every sub-region of the blue area is also wet. (By contrast it is also obvious that, where regions corresponding to countries are indicated on a map, their sub-regions are not themselves countries.) Hence, although a geo-ontology must certainly take account of the downward inheritance of land cover types, it seems that it should be possible to do this without requiring an explicit representation of arbitrary subdivisions of the Earth's surface.

4. Handling Geographic Data: Queries, Definitions and Domains of Quantification

In order to construct an ontology-based GIS capable of answering queries expressed in terms of formally defined geographic concepts and evaluated with respect to geographic data represented by labelled polygons, the following rather challenging problems must be addressed:

P1) The ontology must define all terms in a way that enables their extensions to be somehow computable from the spatial properties and attributes of polygon data.

P2) The formalism must enable the characterisation of features with vague boundaries.

P3) The implementation must be able to deal with regions with implicit geometrical boundaries that are determined by but not explicitly present in the base polygons, without explicitly modelling potentially infinite geometrical dissections of space.

P4) The implementation must be able to take account of the fact that predicates relating to spatially homogeneous properties (such as terrain types) are downwardly inherited (without explicitly modelling arbitrary dissections of space).

P5) An effective method of ontology-based geographic query evaluation must be implemented.

[8] In natural language, such properties are typically associated with mass nouns.

[9] The situation here can be contrasted with the case of a non-downward-inherited feature type predicate such as $\mathsf{Lake}(r)$. In this case, even if we consider geographic space to include arbitrary polygons, only a finite number of these could satisfy this predicate. Hence, it is plausible that instances of $\mathsf{Lake}(r)$ can be obtained by some finitary computation over the base water polygons. Indeed, we have implemented such a computation.

4.1. Spatial Regions and Relations

In order to address **P1**, we need a method of determining the spatial relations that hold between two regions. We use the Region Connection Calculus (RCC) [13], which allows us to express topological relations between regions and to use these to define features involving complex configurations of spatial parts.

However, the standard models of RCC are infinite domains — typically, the sets of all regular closed (or regular open) subsets of Cartesian space (either two or three dimensional). Relating such models to actual data is problematic, because in a computational implementation, one can only refer explicitly to a finite set of entities. Real spatial data usually consists of finite sets of polygons, but the domain of quantification in the standard RCC would include not only these polygons but also all possible ways of carving these up into further polygons.

Our approach to solving this problem is to find a way of working with a finite set of regions, which is adequate to characterise the domain in so far as is relevant to any given spatial query. As discussed in [14], the full set of regions contains many regions we are not interested in, such as tiny regions or obscure shapes with convoluted boundaries, thus we would prefer to work only with the set of regions that we can derive useful or interesting features from. For example, if we are interested only in inland water features, we are only interested in segmenting up the inland water regions, and it may be sufficient to represent the land as a single polygon. We thus choose to restrict our domain of regions to polygons, as previously proposed in [15,16]. To expand upon this, our domain consists of polygons which are initially generated from the data, with further polygons derived from this polygonal information through predicates using standpoints. In [8], we showed how the calculation of the RCC relations between a set of polygons can be performed efficiently.

A problem that arises with such an approach is the generation of this domain. Ideally we would generate all possible polygons to begin with, but this would be too large a set to work with when answering queries. Instead, we start with an initial set of polygons designed to represent the basic separation of *matter types* [17], thus each initial polygon is filled by some specified matter type. These polygons may be further segmented during the query interpretation process.

Such further segmentation will normally arise from shape related or metrical predicates being used in a query (or occurring in the definition of a predicate used in a query). Moreover, since shape and measurement predicates will often be vague, these can correspond to different geometrical conditions, and thus different ways of carving up the initial polygons, according to the standpoint relative to which the query is evaluated.

4.2. Demarcating Vague Regions

Our approach to demarcating vague regions is of course based upon standpoint semantics. This has been explained above and also in several previous papers [8,9] and some further details will be given below in describing our prototype implementation. Here we just give a brief overview. Our procedure first determines a medial axis skeletonisation of the region occupied by a given land cover type. This is then used to segment the region into linear and expansive sub-regions based on threshold values of certain parameters determined by a given standpoint. Vague regions corresponding to different types of geographic feature can then be specified definitionally, in terms of the distribution of land cover types over topological configurations of the regions in this segmentation and over regions derived by further geometrical dissection of these regions.

4.3. Controlled Quantification over Geometrically Derived Regions

We now turn to problem **P3**. One method of constructing an ontology that is computationally tractable over a concrete domain, is to constrain quantification in such a way that all entities (in our case spatial regions) that are relevant to the evaluation of a given formula are either present in an initial finite set of entities, or are members of further finite sets that can be effectively computed from the initial entity set. We now sketch a relatively limited modification of first order logic in which this can be achieved.

Let Base be the finite set of entities (e.g. polygons) present in our data-set. Restricting quantification to range just over entities in Base is clearly tractable, so we can certainly allow quantification of the form:

QB) $(\forall x \in \mathsf{Base})[\phi(x)]$

Many domains have a natural Boolean structure which may be useful for defining properties and relations. Thus in the spatial domain we are often concerned with sums, intersections and complements of regions. Let Base^* be the elements of a Boolean Algebra over Base. We may then allow quantification of the form:

QB*) $(\forall x \in \mathsf{Base}^*)[\phi(x)]$

If Base is finite then so is Base^*. So quantification can still be evaluated by iterating over the domain. But unfortunately Base^* will be exponentially larger than Base, so it would almost certainly be impractical to do this in a real application. However, there is another way of extending the domain of quantification, which is both more controllable and more flexible.

Let $\Gamma(t_1, \ldots, t_m;\, x_1, \ldots, x_n)$ be a relation, such that given any m-tuple of ground terms $\langle t_1, \ldots, t_m \rangle$, one can effectively compute the set of all n-tuples $\langle x_1, \ldots, x_n \rangle$, such that $\Gamma(t_1, \ldots, t_m;\, x_1, \ldots, x_n)$ holds. We may call $\langle t_1, \ldots, t_m \rangle$ an input tuple and $\langle x_1, \ldots, x_n \rangle$ an output tuple. The condition on Γ means that for any given finite set of input tuples there is a finite set of output tuples such that some pair of input and output tuples satisfies Γ. For example, Γ might be a spatial relation $\mathsf{BisectNS}(r; r_1, r_2)$ which is true when r_1 and r_2 are the two parts of r obtained by splitting it into northern and southern parts across the mid-line of its extent in the north-south dimension. Another example is $\mathsf{Concavity}(r, r')$, where given an input polygon r there are a finite number of polygons r' corresponding to concavities of r (i.e. maximal connected regions that are parts of the convex hull of r but do not overlap r).

We shall call relations of this kind *effective generator relations*. They are simply logical representations of a certain kind of algorithm that could be implemented in computer software — and indeed much of the functionality of a GIS depends on algorithms of this kind. Given an effective generator relation Γ, we can now define the following form of controlled quantification:

QEGR) $(\forall x_1, \ldots, x_n : \Gamma(t_1, \ldots, t_m;\, x_1, \ldots, x_n))[\, \phi(x_1, \ldots, x_n)\,]$

Here, the variables $t_1, \ldots, t_m$ must be already bound to wider scope quantifiers, which can be either quantifications over Base or over domains specified by other effective generators. Hence, the range of each variable is restricted either to Base or to a set of entities that can be computed from Base by applying algorithms corresponding to a series of effective generator relations.

Semantically, **QEGR** is interpreted as equivalent to:

- $(\forall x_1, \ldots, x_n)[\,\Gamma(t_1, \ldots, t_m;\, x_1, \ldots, x_n) \rightarrow \phi(x_1, \ldots, x_n)\,]$

4.4. Spatially Homogeneous Properties and Downward Inheritance

So far we have not implemented any mechanism for handling downward inheritance. Instead we have circumvented the issue by limiting our predicates to those satisfied either by maximal components of uniform land cover, or by regions derived from these by particular geometrical decompositions. For instance, we define 'linear stretches' of water which are geometrically dissected (relative to a given standpoint) from the total region of water. In the future we would like to handle spatially homogeneous properties by representing their logical relationship to base polygons.

4.5. Query Evaluation

We express a query by means of the notations $\mathsf{?} : \phi$ representing a test as to whether ϕ is true in relation to a given data-set and $\mathsf{?}(x) : \phi(x)$, which means: return a list of all entities e_i such that $\phi(e_i)$ is true as determined by interpreting the symbols of ϕ in relation to the data-set. More generally, $\mathsf{?}(x_1, \ldots, x_n) : \phi(x_1, \ldots, x_n)$ would return a list of n-tuples of entities satisfying the given predicate. In our context, the entities returned will normally be polygons. Query variables cannot occur within any of the quantifiers of our representation, however they can be identified with values of these variables by the use of an equality predicate.

Since queries will be interpreted in relation to actual geographic data, it is natural to use a *model-based* approach to query evaluation.[10] General purpose model building systems, such as MACE [19], allow consistency checking of arbitrary first order formulae, by checking all possible assignments to predicates. But in our case we have a single interpretation of basic predicates that can be derived directly from the geographic dataset. Thus, we can compute sets of all tuples satisfying the predicates that occur in a query and then evaluate the query formula over this model.

Boolean connectives can be evaluated in an obvious way, but the treatment of quantifiers is somewhat more complex. Since quantification is restricted to range over either base polygons or derived polygons generated by the **QEGR** quantifiers, this means that the domain of regions that must be considered is finite. By examining the structure of nested **QEGR** quantifiers occurring in a query, we can determine sequences of spatial function applications which, when applied to the base polygons, will generate all polygons that are relevant to that query. Once these polygons have been computed, quantifiers can be evaluated over this extended domain. Our current prototype does not explicitly include the **QEGR** quantification syntax, but implements a simplified version of this

[10] Model-based reasoning has been applied in various areas of AI. For instance, a similar approach to ours has been used in Natural Language Processing [18].

mechanism. It is geared towards evaluating queries containing a limited range of predicates and generates domains of polygons that are sufficient to deal with these. This will be described in the next section.

5. Implementation within a Prototype GIS

We now give some details of our GIS prototype which we call GEOLOG. The system is implemented in Prolog and operates on several hydrographic data-sets covering estuarine river systems in the UK. The system implements geometric shape decomposition algorithms based on a number of parameters. These are linked to an explicit representation of shape predicates using a first order formalisation in which the parameters attached to predicates are interpreted according to standpoint semantics. First order queries can be evaluated and their instantiations depicted on a cartographic display.

5.1. Shaped-Based Properties and Segmentation

Since queries may contain RCC relations describing topological relations between regions, a database of RCC relations over all stored polygons is maintained. This requires a considerable amount of storage but means that these relations do not have to be recomputed whenever a new query is executed, which greatly speeds up query answering times. As described in [20,8] segmentation of regions into linear and expansive parts is computed using a *medial-axis* approach which is supported by use of the VRONI software package [21]. The idea is to measure width variation along the medial axis. Given a medial axis point p of region r which is distance d from the edge of r, we compute the maximum and minimum distances, max, min, to the edge of r of all medial axis points within distance d of p. The value l=$\mathsf{max}/\mathsf{min}$ gives a useful measure of the width variation at p. $l = 1$ means the width is constant, and a value of 1.2, for example, means that there is a 20% width variation in a section of the medial axis centred at p along a length equal to the width at p. Using this value as a standpoint parameter, the predicate $\mathsf{Stretch}(r)$, corresponding to the vague concept of a 'linear stretch of water' is defined. This is a maximal connected water region all of whose medial axis points have a value of l less than a given threshold.

5.2. Query Evaluation

In developing an effective implementaton, we wanted to minimise both the number of polygons stored in the system and the time required to construct polygons by geometrical computation. This led us to an approach of 'just in time', incremental expansion of the domain. The basic idea is that that when presented with a query, GEOLOG ensures that all polygons relevant to its interpretation are generated before evaluating the query as a whole. But it then stores the generated polygons as they are likely to be required again for subsequent queries.

The initial dataset consists simply of a partition of the geographic space comprising polygons labelled with the basic land cover types: *land*, *sea* and (fresh) *water*. Queries relating to the base polygons themselves can be answered straightforwardly, although they are of little interest as they do not take any account of the semantics of geographic features. However, a number of higher level geometric and hydrographic predicates are also available for use in queries. Each of these predicates is associated with an algorithm for expanding the domain of polygons by geometrical computations, to include additional polygons corresponding to all their possible instances. When a query containing one or more of these non-basic predicates is entered, the domain is first expanded according to the associated algorithms (in general this must be done recursively until a fixed point is reached), and the newly generated polygons are labelled with appropriate attributes. Once this procedure has has been carried out, the dataset contains polygons corresponding to all possible instances of predicates occurring in the query. Quantifiers can now be evaluated by iterating over polygons in this expanded dataset.

For instance, if one enters a query $\mathsf{Stretch}(x)$ GEOLOG would perform a linearity segmentation relative to a given standpoint, so that the required linear and expansive polygons are generated. We can now answer queries involving reference to stretches or to any concepts that have been defined in terms of linear and/or expansive polygons. A user of the system has direct access to the threshold assignment defining the standpoint and can modify the thresholds. When this is done the system must recompute the segmentation, and this in turn will lead to different polygons being returned from queries that depend on the segmentation.

5.3. Results of Querying for Stretches and Rivers

Results of executing the query $\mathsf{Stretch}(x)$ with different input datasets and linearity parameter thresholds are shown in Figure 2. The images on the left correspond to a threshold of 1.2, whereas those on the right are for a threshold of 1.4. Thus, the interpretation on the right takes a more liberal view of what can count as linear than the one on the left.

As is clear from Figure 2, the artificial concept of ‘linear water stretch’ does not correspond directly to the natural concept of ‘river’. Typically a river will consist of many such stretches, interspersed with more expansive areas of water, corresponding to bulges in the watercourse. We experimented with a range of threshold parameters governing how loosely or strictly the predicate ‘linear’ is interpreted; but found that there is no threshold that yields a natural interpretation of ‘river’. If we use a loose definition that allows bulges to be classified as parts of a stretch, we find that very expansive, lake-like water regions become incorporated into stretches. But if we tighten the linearity threshold to rule out obvious lakes, then rivers must be consist of fragmented sequences of stretches.

In order to circumvent this problem, we propose that a river should indeed be modelled as a sequence of stretches interspersed by bulges. To make this precise we have introduced a further artificial concept of *interstretch*. This is a water region that is expansive but such that all its parts are ‘close’ to a water stretch, where closeness is defined by a second threshold applied to a suitable geometric measure. This enables us to incorporate small bulges into rivers without needing to unduly weaken our general criteria for identifying linear water features. As described in [8], this has been found to interpret the concept river in a very plausible way.[11]

The introduction of interstretches might at first sight appear to be an *ad hoc* hack. However, we believe that a plausible general explanation can be given as to why this seems to work. In classifying a vague feature, we suggest that one is looking for criteria

[11] Further complications arise from the branching structure of water systems. These have been only partially solved and are a topic of ongoing work.

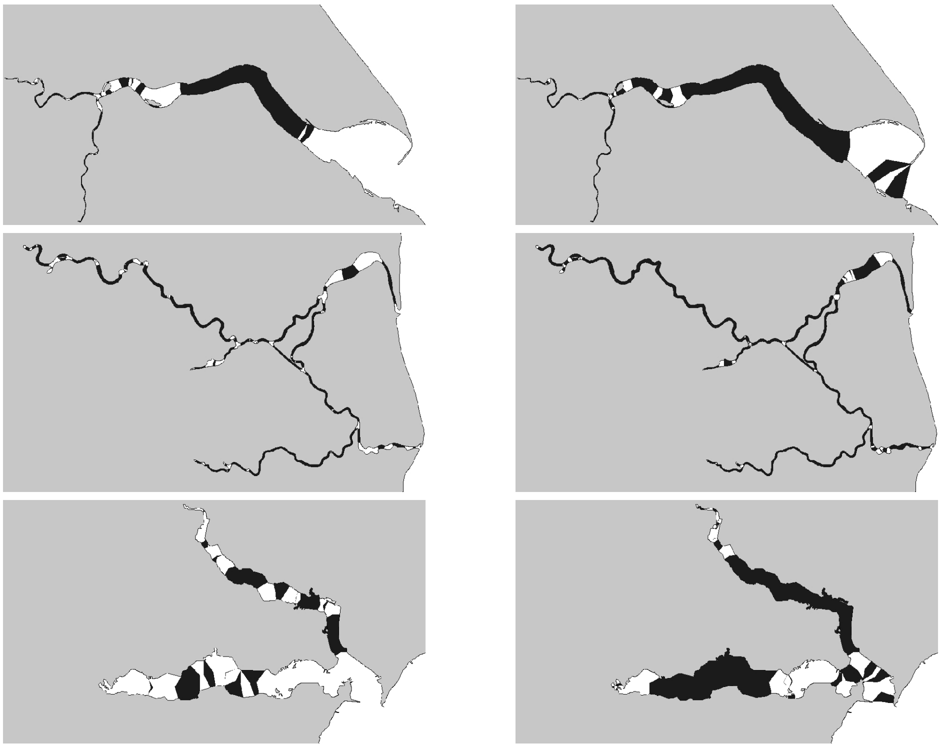

Figure 2. A comparison of marking ‘linear water stretches’ relative to different The top images are of the Humber Estuary, the middle images are of the Norfolk Broads at Great Yarmouth and Lowestoft. The bottom images are of the Stour And Orwell Estuary.

that are satisfied globally by a region but is also prepared to allow exceptions in regard to small parts of the region that deviate from these criteria. For instance, to classify a surface as approximately planar, one is looking for a global approximation to a plane but will accept small areas where the surface departs considerably from planarity, which are regarded as insignificant bumps on the surface. We thus plan to apply a similar approach to classifying other types of geographic feature.

6. Conclusion

We have described a variety of ontological issues that complicate the issue of defining and individuating geographic regions and features. From theoretical analysis of the semantics of vagueness and of computational manipulation of geometric decompositions of polygonal data, a possible architecture for implementing an ontology-based GIS is taking shape. Our current prototype gives a strong indication that this can lead to a new kind of GIS in which geographic terminology is grounded upon data *via* rigorous definitions rather than *ad hoc* segmentations. However, much work remains to be done, both in terms of specifying a more extensive geographic ontology and also in relation to developing a more flexible and efficient query answering mechanism.

Acknowledgements

We are grateful to the EPSRC for supporting this work *via* grant EP/D002834 and to the Ordnance Survey, for their contribution towards funding a PhD on this topic. Valuable comments from the reviewers are also gratefully acknowledged.

References

[1] Achille C. Varzi. Vagueness in gegraphy. *Philosophy and Geography*, 4:49–65, 2001.

[2] Brandon Bennett. What is a forest? on the vagueness of certain geographic concepts. *Topoi*, 20(2):189–201, 2001.

[3] Barry Smith and David M. Mark. Do mountains exist? towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30(3):411–427, 2003.

[4] Paulo Santos, Brandon Bennett, and Georgios Sakellariou. Supervaluation semantics for an inland water feature ontology. In L.P. Kaelbling and A. Saffiotti, editors, *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 564–569, Edinburgh, 2005.

[5] Kit Fine. Vagueness, truth and logic. *Synthèse*, 30:263–300, 1975.

[6] Timothy Williamson. *Vagueness*. The problems of philosophy. Routledge, London, 1994.

[7] Brandon Bennett. A theory of vague adjectives grounded in relevant observables. In Patrick Doherty, John Mylopoulos, and Christopher A. Welty, editors, *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 36–45. AAAI Press, 2006.

[8] David Mallenby and Brandon Bennett. Applying spatial reasoning to topographical data with a grounded ontology. In F. Fonseca, M. Andrea Rodrígues, and Sergei Levashkin, editors, *GeoSpatial Semantics, proceedings of the second international conference*, number 4853 in Lecture Notes in Computer Science, pages 210–227, Mexico City, November 2007. Springer.

[9] Allan Third, Brandon Bennett, and David Mallenby. Architecture for a grounded ontology of geographic information. In F. Fonseca, M. Andrea Rodrígues, and Sergei Levashkin, editors, *GeoSpatial Semantics, proceedings of the second international conference*, number 4853 in Lecture Notes in Computer Science, pages 36–50, Mexico City, November 2007. Springer.

[10] Barry Smith. On drawing lines on a map. In Andrew Frank and Werner Kuhn, editors, *Spatial Information Theory — Proceedings of COSIT'95*, number 988 in Lecture Notes in Computer Science, pages 475–484, Berlin, 1995. Springer.

[11] Barry Smith. Fiat objects. *Topoi*, 20(2):131–148, 2001.

[12] Alberto Belussi and Matteo Cristani. Mereological inheritance. *Spatial Cognition and Computation*, 2:467–494, 2000.

[13] D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic-based on regions and connection. In *Principles Of Knowledge Representation And Reasoning: Proceedings Of The Third International Conference (Kr 92)*, pages 165–176. Morgan Kaufmann Pub Inc, San Mateo, 1992.

[14] Ian Pratt and Dominik Schoop. A complete axiom system for polygonal mereotopology of the real plane. *Journal of Philosophical Logic*, 27(6):621–661, 1998.

[15] Volker Haarslev, Carsten Lutz, and Ralf Möller. Foundations of spatioterminological reasoning with description logics. In *Principles of Knowledge Representation and Reasoning*, pages 112–123, 1998.

[16] Rolf Grütter and Bettina Bauer-Messmer. Towards spatial reasoning in the semantic web: A hybrid knowledge representation system architecture. In S. Fabrikant and M. Wachowicz, editors, *The European Information Society: Leading the Way With Geo-information*, volume XVII of *Lecture Notes in Geoinformation and Cartography*, page 486, 2007.

[17] Brandon Bennett. Space, time, matter and things. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 105–116, New York, NY, USA, 2001. ACM.

[18] Patrick Blackburn and Johan Bos. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI, 2005.

[19] William McCune. Mace4 reference manual and guide. *CoRR*, cs.SC/0310055, 2003.

[20] David Mallenby. Grounding a geographic ontology on geographic data. In E. Amir, V. Lifschitz, and R. Miller, editors, *Proceedings of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense-07)*. AAAI, 2007.

[21] Martin Held. VRONI: An engineering approach to the reliable and efficient computation of voronoi diagrams of points and line segments. *Computational Geometry: Theory and Applications*, 18(2):95–123, 2001.