Modelling Vague Places with Knowledge from the Web

CHRISTOPHER B. JONES*[1], ROSS S. PURVES[2], PAUL D. CLOUGH[3], HIDEO JOHO[4]

[1]School of Computer Science, Cardiff University
[2]Department of Geography, University of Zurich, Switzerland
[3]Department of Information Studies, University of Sheffield, UK
[4]Department of Computing Science, University of Glasgow, UK

* Corresponding author (c.b.jones@cs.cf.ac.uk ; +44 (0)29 2087 4796)

Modelling Vague Places with Knowledge from the Web

Abstract

Place names are often used to describe and to enquire about geographical information. It is common for users to employ vernacular names that have vague spatial extent and which do not correspond to the official and administrative place name terminology recorded within typical gazetteers. There is a need therefore to enrich gazetteers with knowledge of such vague places and hence improve the quality of place name-based information retrieval. Here we describe a method for modelling vague places using knowledge harvested from web pages. It is found that vague place names are frequently accompanied in text by the names of more precise co-located places that lie within the extent of the target vague place. Density surface modelling of the frequency of co-occurrence of such names provides an effective method of representing the inherent uncertainty of the extent of the vague place while also enabling approximate crisp boundaries to be derived from contours if required. The method is evaluated using both precise and vague places. The use of the resulting approximate boundaries is demonstrated using an experimental geographical search engine.

Keywords: Geographical information retrieval, gazetteers, vagueness, geo-parsing, surface modelling

1 Introduction

Place names play an essential role in communicating geographically-specific information. We use place names in our everyday language when describing the location of places and when giving navigational instructions, and they occur frequently in text documents when it is required to provide geographical context. Despite their importance in communicating geographical information, little attention has been paid to representing knowledge of place names in the context of geographical information systems (GIS). Specification of location by users of GIS has tended to be dependent upon graphical interaction with maps and the explicit use of map coordinates. The prevalence of geographical information on the web and the need for more intuitive methods of referring to geography has led to an increasing demand for automated understanding of place name terminology (Hill 2006, Jones *et al.* 2001, Schlieder 2001). The use of place names is also an aspect of the broader requirement for GIS user interfaces to interpret vague concepts (Kuhn 2001) and understand natural language (Wang 1994).

There are several types of public information systems in which place names are now used as the main method of specifying location. These include transport timetables, routing systems for motorists, map-based web sites and web search engines. Applications that provide information access via place names usually employ gazetteer-based resources to recognise their presence and to resolve ambiguities in names for which there are multiple occurrences. Simple digital gazetteers store the text of the name itself and map coordinates, typically in the form of a

representative point, and some geographical hierarchy data such as the county and state, or nation associated with the name. More detail is found in gazetteers such as that of the Alexandria Digital Library (Hill *et al.* 1999), which makes it possible to encode additional geometric and feature attributes, as well as relations between the place and other places. The Getty Thesaurus of Geographic Names (Harpring 1997) is notable for storing alternative historical names of places and recording geopolitical and some topographic hierarchies. Gazetteers and geographical thesauri can be regarded as types of geographical ontology which are now recognised as a necessary component of systems for geographical information retrieval (Schlieder *et al.* 2001, Egenhofer 2002, Jones *et al.* 2003, Larson 2003, Vogele *et al.* 2003).

The content of such resources is often derived from standard map series produced by national mapping agencies. As such they reflect an official or administrative view of geographic space dominated by administrative subdivisions. However, when people employ place names in natural language, many of the names used refer to places that do not coincide exactly with places of the same name referred to in conventional gazetteers, or indeed are not referenced at all. Thus, there are many vague or vernacular place names, such as the English Midlands, the South of France, the Rocky Mountains, and the Midwest, for which no official boundary exists. Equally, there are places names that have been adopted for administrative purposes, but for which the administrative boundary differs from many people's perception of the extent of the place.

There is a need therefore to acquire knowledge of the common perception of the extent of vague places so that they can be used intelligently in geographically-focused information systems. Even though any such extents will necessarily be approximate, their presence will facilitate access to

map based and text based information. In this paper, we describe and evaluate a method that uses knowledge acquired from the Web to model the extent of and generate approximate boundaries for vague places. The method exploits the fact that when a vague place is mentioned in a text document it is often accompanied by references to other more precise places that lie within the extent of the vague place. Density surface modelling methods can then be used to identify regions corresponding to the most frequently co-occurring places. Evaluation of the method on both precise and vague regions shows promising results, though these are subject to variation in the quality of geo-parsing and to the spatial distribution of the references on the web to co-located places. Its application is demonstrated here using a geographical web search engine. Some of the ideas on which this paper is based were introduced in Purves *et al.* (2005) and in Arampatzis *et al.* (2006). The current paper explores the density surface modelling approach in much greater depth, describes alternative web harvesting techniques and provides new experimental evaluation of the methods presented including application to a geographical web search engine.

In the remainder of the paper we review related work on describing and modelling vague places in section 2, before describing in section 3 our method for web harvesting and subsequent spatial modelling of vague places based on the frequency of occurrence of co-located places. This includes a summary of the geo-parsing and geo-coding methods required to retrieve the co-located places. In section 4 we describe experiments to evaluate the methods using both precise target places and inherently vague places. The use of resulting boundaries is illustrated in the context of a geo-search engine (section 5) before discussing various aspects of the approach and presenting some future research directions in section 6.

2 Related Work

Previously the web has been used as a source of knowledge for applications such as question-answering (Kwok *et al.* 2001, Clarke *et al.* 2001, Radev *et al.* 2002) and studying language usage (Kilgarrif and Grefenstette 2003). The web also provides a rich source for geographic knowledge as many web pages contain references to places on earth and users often search with locations as part of their query (Zhang *et al.* 2006). However, using the web as a geographic knowledge source for deriving region boundaries is not well researched. Most work so far on vague regions has concentrated on two areas: (1) how vague places can be described, for example by human subjects, and (2) how boundaries of vague places can be modelled or derived from empirical data.

2.1 Describing Vague Places

One approach to eliciting knowledge of the extent of a vague place is to ask human subjects to draw the boundary (Aitken and Prosser 1990). For example, in an investigation of people's perception of the vague place "downtown Santa Barbara" Montello *et al.* (2003) asked pedestrians to draw the location on a map in various ways, first assuming that there was a precise boundary and then making a distinction between a 100% certain boundary and a 50% certain boundary. They were also asked to place a point at a location that they regarded as the core of the region.

Some place names are directly associated with topographic features, prime examples being those of mountains, valleys, ridges and passes. On most maps place names are associated with a location either purely by means of the placement of the text, which may stretch across the relevant feature, or by associating the name with a point or line, such as for mountain peaks and

valleys. Fisher *et al.* (2004) show how the extent of topographic features can be delineated automatically by analysing terrain models. Although many locations were classified as more than one type of feature, depending upon the scale of the analysis, some locations had a dominant classification. A very strong correlation was found between the point references of mountain peak place names on maps and the automatic categorisation of the terrain as either dominantly a peak or a ridge. These methods appear to have the potential to be applied systematically to delineation of vague topographic features provided that map names can be associated with locations on terrain models.

A method that depends heavily upon the text found on maps is described by Lam *et al.* (2002) for the purpose of delineating neighbourhoods in the city of Los Angeles. The extent of the neighbourhoods was represented by circular footprints, centred on points based on the location of labels in a street guide. The city was divided into regions according to the density of neighbourhoods and the radii of the circles allocated to neighbourhoods varied, with smaller circles being used in the denser regions. This approach results in a non-exhaustive partitioning of space, reflecting the fact that there were often regions of space without neighbourhoods. Also the circles often overlapped, reflecting the vagueness and variability in the interpretation of neighbourhoods.

When describing the location of a vague place it is common to explain its location relative to other named places. In his discussion of the definition of the American Southwest, Byrkit (1992) quotes Lawrence Clark Powell as saying that the Southwest included "the lands lying west of the Pecos, north of the [Mexican] Border, south of the Mesa Verde and the Grand Canyon, and east

of the mountains which wall off Southern California and make it a land in itself". Here Powell has used external places of reference. On another occasion Powell refers to Albuquerque as being at the core. Provided that some places internal and external to the vague place can be identified then it will be possible to construct an approximate boundary that lies somewhere between the two sets of places. Because there will often be disagreements as to what is inside and what is outside, it is possible to envisage compiling multiple boundary interpretations that could be input to a vague region modelling method.

Empirical approaches based on interviews with human subjects are a powerful means of exploring how vague places are conceptualized by subjects. However, such experiments are time consuming to conduct and analyse and the question arises as to whether published texts might provide an alternative data source. The most readily available source of text for automated analysis is the web. Web harvesting techniques to identify texts mentioning vague place names in association with other precise place names provide a potential approach that has been subject to an initial set of experiments by Arampatzis *et al.* (2006). Their method depends upon employing web queries that include so-called trigger phrases that may reveal the relevant knowledge. For example, it is possible to search for the phrase “Midwest cities such as” and then retrieve place names that may follow the phrase, on the assumption that they are regarded as being inside the named place. We pursue related web harvesting methods in this paper (see Section 3 onwards) and show that alternative query methods may be more effective.

2.2 Modelling the Boundaries of Vague Places with Empirical Data

By definition a vague place cannot be expected to have a single precise boundary. However, for the purposes of information retrieval, it may be highly desirable to approximate a vague place by

a sharp boundary which can then form the basis of subsequent ordering or ranking of associated information content. In this section we consider some methods that might be employed to generate such approximate sharp boundaries from different forms of acquired knowledge of the extent of vague places.

Boundary drawing methods, such as those referred to earlier, generate multiple sharp boundaries, but a method is required to decide on a single representative boundary. Montello et al. (2003) propose a "frequentist" probabilistic method for modelling the vague region, whereby for each location in the region of interest the number of randomly chosen human subjects who consider that location to be in the vague region is recorded as a proportion of the total number of subjects. These values are obtained by overlaying the binary maps obtained from the boundaries that were drawn. Having created such a probability surface it would then be possible to use it to generate an isoline boundary corresponding to a chosen level of probability of inclusion. Montello *et al.* also indicate how a fuzzy model could be derived by asking respondents to specify the location of boundaries with given levels of confidence.

The supervaluation method of representing vague regions (Kulik 2001) is based on the assumption that there exist precise ("sharp") interpretations of the boundary of a vague region. Thus a vague region can be defined by a set of sharp regions, which are admissible interpretations of the extent of the region. They lie between a definite (inner) core region and a maximal region, the hull, beyond which is definitely external to the vague region. There is a minimum of two interpretations, which provide the extent of the core and the extent of the hull. If there are only two such interpretations then the model is equivalent to the egg yolk model (Cohn

and Gotts 1996). The degree of vague region membership of a point lying between the core and the hull can be quantified in terms of the number of sharp regions that contain it. Kulik points out that this allows for a conversion to fuzzy set membership. A single boundary could be generated at a chosen alpha cut level. Alternatively a probabilistic interpretation could also be placed on the data, similar to that of Montello *et al.*, depending upon how the boundaries were acquired.

A qualitative approach to representing vague places has been presented by Vogele *et al.* (2003). Vague places are described in terms of the topological relations to neighbouring places, using the relations of containment, equivalence and overlap with existing regions such as administrative areas. Places defined in this way have an upper and lower approximation. The lower approximation consists of the related regions that are definitely inside or equivalent to the imprecise place, while the upper approximation consists of these definite regions plus overlapping regions. Assuming data exists for the boundaries of the related existing regions then a boundary could be created from either the lower or upper approximation. It may be noted that rough sets can also be used to model upper and lower approximations of the extent of imprecise regions as described by Worboys (1998).

In the event of knowledge of a vague place being available as points classified as either inside or outside, a simple interpolation procedure may be employed to generate a precise approximation of the boundary. In Alani *et al.* (2001) a Voronoi diagram is created from the internal and the adjacent external points and the cells then categorised according to whether they represent internal or external points. The Voronoi cell edges that lie between these two sets constitute an approximate boundary. A similar application of Voronoi diagrams to generate approximate

boundaries between imprecise regions, for purposes of digitising was described by Gold *et al.* (1996). Arampatzis *et al.* (2006) have adopted related techniques, using Delaunay triangulations, which assume that there may be error in the categorisation of points as internal and external. Their methods modify the boundary between the two sets to eliminate isolated points that can be regarded as outliers or misclassified points. The techniques applied are based entirely on geometric criteria and do not take account of actual measures of likelihood of a point being inside the target region (though the authors point out that the methods could be adapted to use such information).

If a set of candidate points for membership of an imprecise region exist and they are accompanied by measures of probability of inclusion within the region, it is possible to derive a surface from these points through interpolation, where peaks in the surface correspond to a high probability of membership of the imprecise region. Sharp boundaries can then be generated if required by selecting a surface value that serves as a threshold. The resulting isoline on the surface then separates values above the threshold from those below. Furthermore, it is possible to make multiple slices of such a surface, corresponding to multiple sharp regions as described above. This density surface modelling approach was adopted by Purves *et al.* (2005) and is followed in the present work. Given an irregularly distributed set of points, a regular grid of points can be generated by a process of interpolation. Points may also be attributed with values that measure their importance in some way or the number of times the place is referenced. A distance weighted interpolator may then be applied, whereby for each interpolated point the nearest neighbouring sample points are summed to create a weighted average. The reason for employing the approach in the current work is that the modelling method is able to exploit

statistical evidence for inclusion of points within a candidate vague region and in doing so create a model that reflects the variation in confidence of this inclusion. The resulting density surface therefore stands in its own right as a representation of the vague region while also facilitating generation of crisp approximations of the boundary of the region at different levels of confidence if they are required.

3 Procedures for Acquiring and Modelling Place Name Knowledge from the Web

Web pages often contain place names in order to provide geographic context. Some place names refer to places that have well defined boundaries, such as a county, while others refer to vague regions, such as the South of France and the Midwest that have vague boundaries. There are also variations in the granularity of the places referred to in that a document that mentions a county or an extensive vague region may also mention smaller places that lie within these larger regions. It has been hypothesised that place names that occur frequently in association with the name of a more extensive region can be expected to lie in the vicinity of, and often inside, the latter region (Purves *et al.* 2005, Arampatzis 2006). This introduces the possibility of modelling vague places in terms of their expected contained places. To test whether frequently co-occurring place names in web pages can be used to estimate the extent of a specified region, we have experimented using both precise and vague places as the target (more extensive) regions. By experimenting with precise target places we can evaluate the efficacy of the method, before moving on to consider how the method can be extended to vague target regions. To determine the approximate boundary of either a precise or vague region from web search, we perform the following:

1) Search the web for pages containing a reference to the target region;

2) Extract all place names from the highest ranked 100 results;
3) Assign spatial co-ordinates to extracted place names;
4) Create a geometric model of the region and extract an approximate boundary if required.

Simple application of this method is based on an assumption that the geographic coverage of the web is complete and unbiased, in the sense that there are actually references on the web to all known vague regions and to all associated places and that the spatial distribution of these references is random. In practice, there are weaknesses in the assumption. Thus there will tend to be many more references on the web to places with higher populations and in particular to places that are popular, or of interest, for some reason, such as tourism. It is also the case that some regions may have relatively few contained settlements, with most of their contained places being other topographic features. As a consequence it is necessary to give careful consideration to the form of the web queries, an issue that is discussed further in section 3.1.

3.1 Searching the Web

To gather information about candidate contained places of an imprecise region, queries containing a reference to a target region were submitted to the Google search engine. The goal of searching is to find web pages which are both rich in geographical content and focused on the target region. Documents which do not fulfil these requirements will either contain few place names, or locations which are unlikely to fall within the target region.

Google is known to use an algorithm called PageRank (Brin and Page 1998) to rank web pages. PageRank is designed to analyse the link structure of web pages, where a hyperlink from a page to another page is seen as a vote. Those pages that contain query words and have more votes from other pages are ranked higher than those with fewer votes. The algorithm also considers the number of incoming links of the voters so that a vote from a page that has more votes is weighted higher than a vote from a less voted page. It is generally believed that the pages with many votes tend to be popular and authoritative for a given topic. Thus it enables us to decrease the level of noise when extracting geographic references. It should be noted that while there are search engines that allow a paid inclusion to their collection, to the best of our knowledge this does not apply to Google except for the sponsored links shown in the distinct sections of search result pages.

There are several forms of web query that can be expected to return associations between a place and its contained places including:

- **Region only:** a query containing a reference to the target region only, e.g. "the Rocky Mountains";
- **Region and concept:** a query containing a reference to the target region and associated concept to select certain types of pages, e.g. "hotels in the Cotswolds" tends to select directory-style pages;
- **Region and pattern:** a query containing a reference to the target region that includes or implies a spatial relationship, e.g. "*in the South of France" and "Midwest towns such as *" (this approach based on lexical patterns was used in Arampatzis *et al.* 2006).

During initial experiments, queries based on region and concept appeared to find the most geographically-rich pages, e.g. directory listings. These pages also included more fine-grained place names such as villages and postcodes. Although the first and last types of query tended to find pages with place names more likely to be related to the target region (e.g. contact-us pages), these approaches also generated far fewer locations. Importantly, the approach of interpolating a surface from point data is relatively insensitive to false positives obtained through the second query type, as will be shown later.

All queries were submitted to Google and were of the form "~hotels [target region]" where the '~' symbol is a synonym operator which will expand the query automatically to search for synonyms of hotels such as "inn" and "accommodation". Searches were restricted to UK pages only and, having identified a set of web documents, we removed all document markup to leave only formatted plain ASCII text for further processing. This helped to reduce the number of false hits, e.g. names within HTML tags.

3.2 Extracting Place Names (Geo-parsing)

Given a set of web pages, Named Entity Recognition (NER) methods were used to detect the presence of place names and other geographical references (e.g. postcodes). This step is called geo-parsing and in our experiments, we used ANNIE, the default Information Extraction (IE) system that comes with GATE (General Architecture for Text Engineering) (Cowie and Lehnert 1996, Cunningham *et al.* 2002). ANNIE was used to perform NER using both internal and external evidence (McDonald 1996) in the form of gazetteers and proper name lists and context rules to disambiguate between named entities, also called *referent class ambiguity* (Smith and Mann 2003). For example, if we found the sequence "<Forename> <Location>" where Location

and Forename exist respectively in a gazetteer and list of proper names, we would assume that <Location> in this case refers to a surname and is not being used in a geographical context.

The standard GATE gazetteer lists were enhanced in the present work by using two main sources of UK data for lookup: (1) the SABE (Seamless Administrative Boundaries of Europe) dataset and (2) the Ordnance Survey 1:50,000 scale gazetteer. These two datasets contain a total of around 270,000 locations of which about 10% are ambiguous, i.e. not unique entries. In addition to the gazetteer lists, we identified postcodes which could also be used to provide valuable spatial information. The UK and European data gazetteer data were supplemented with the Getty Thesaurus of Geographic Names (TGN), a hierarchical geographical thesaurus of over 1 million names that provides global geographical knowledge. Note that all of these resources contain geographic coordinates for the listed places.

Effective geo-parsing is a challenging task and the methods employed in the present work have scope for improvement. The emphasis here has been upon an initial evaluation of the potential for using these methods for modelling vague places. Improved geo-parsing would then result in improved results for co-locating place names.

3.3 Assigning Coordinates (Grounding)

Once a place name has been identified it is "grounded", i.e. a map coordinate is allocated to it using the place name resources described in section 3.2. To assist with the disambiguation of places where the same name is used in different countries, we used the Getty Thesaurus of Names (TGN), in addition to the UK place name resources. The reason for this is that if we only

have resources for the UK and we encounter the location "Lancaster", it would be incorrectly grounded if the name actually referred to "Lancaster" in "Pennsylvania". Having world knowledge enables us to ignore this location rather than incorrectly assign it to the UK.

There are three main ambiguities in geo-references: (1*) referent* ambiguity - the same name is used for more than one location, (2*) reference* ambiguity - the same location can have more than one name and (3) *referent class ambiguity* – place names can be used in non-geographic contexts such as organisation or person names (Smith and Mann 2003). The simplest method for resolving referent ambiguity is to assign ambiguous places a default position. This can be decided by, for example, the most commonly occurring place (Smith and Mann 2003), by population of the place name (Rauch *et al.* 2003) or by semi-automatic extraction from the Web (Li 2003). The method we use to ground locations is based on matches between place names within the local context of a location and the associated hierarchy provided by the geographic resource, for example. World > Europe > United Kingdom > England > Lancashire. If matches between the local contexts are not found, then place names are assigned a default sense (a coordinate). In these experiments the default sense corresponds to the "largest" location that has the given name, based on feature types and hierarchy depth as provided by the gazetteers. UK postcode data are particularly useful as they are effectively unambiguous and thereby introduce less error.

3.4 Spatial Modelling and Boundary Generation

We use spatial density estimation methods to represent the distribution of the co-occurring point referenced places found in the geo-parsing and grounding stages. These methods are relatively robust to false positives, so long as the data points are randomly distributed in space.

Density estimation techniques allow us to interpolate a continuous density surface from a point data set with numerical attributes. The resulting surface is usually represented as a raster. The most basic approach to density estimation applies so called naïve methods, where density is calculated by summing all the point values within a circle and dividing by the area of the circle:

$$\tilde{\rho}_q = \frac{\sum p_i \in C(q, r)}{\partial r^2} \quad (1)$$

where

ρ_q is the density at some location $\boldsymbol{q}$;

$C(\boldsymbol{q},r)$ is a circle centered on $\boldsymbol{q}$ with a radius r; and

$\boldsymbol{p_i}$ are values at points contained within the circle $C(\boldsymbol{q},r)$.

Kernel density estimation (KDE) methods add a weight to the values at points $\boldsymbol{p_i}$ according to their distance from $\boldsymbol{q}$ which smooths the influence of points with distance so that points nearer to $\boldsymbol{q}$ have a greater influence on the density value (O'Sullivan and Unwin 2003). This weighting often, and in this paper, takes the form of a quadratic kernel function with a value of zero at radius r. Calculating density surfaces requires the selection of the surface resolution (i.e. grid cell size) and the kernel radius. Resolution of the surface must be sufficient to resolve the boundaries of the region and will vary according to region size and kernel radius. Kernel radius should ideally be small enough to represent local variation within the region at a scale commensurate with the size of the region and large enough to capture multiple point locations within the kernel radius. In general, kernel radii are arrived at experimentally (O'Sullivan and Unwin 2003), and in

our case values of between a half, and an order of magnitude less than, the maximum estimated diameter of the regions under investigation was used. Adaptive kernel density estimation (Brunsdon 1995) varies the kernel radius as a function of the original point distribution, but this was not done in this paper as the results obtained were considered to be reasonable with the methods applied.

Experiments were carried out to assess the effectiveness of density surfaces using several different point attributes, including term frequency (the total number of occurrences of a place name in the retrieved documents) and document frequency (the number of documents a place name occurs in). Peaks in the resulting surface correspond to clusters of places that occur frequently in association with the target place name. Ideally there will be a single major peak corresponding to the vague region, but, as is shown in the experimental results and explained later, additional spurious peaks may also appear.

If sharp boundaries are required to delineate the extent of the major peak(s) in the surface, then they can be obtained by selecting threshold values for membership of a region and retrieving the resulting contour-bounded regions. Threshold point densities were selected interactively for precise regions. The initial density was set to a value of one point-referenced place location per grid cell, and where this resulted in multiple non-homogenous surfaces the threshold was progressively halved until a single, dominant region was identified. This selection of a threshold value removes vagueness from the representation, and simply considers locations to be inside or outside of the vague region. This is a common weakness shared with many other techniques that attempt to define an inherently vague object using some form of threshold. Vague geographical

objects are Sorites susceptible (Fisher 2000) in that there is no single threshold value that genuinely distinguishes the object from the non-object. In practice however, as indicated previously, an approximate precisification can be very useful for purposes of information retrieval.

The key elements of producing approximate polygon boundaries from candidate datasets of associated point place locations can thus be described as follows:

1. Select appropriate kernel size and surface resolution;
2. Generate a density surface using KDE;
3. Identify relevant regions based on a threshold point density.

4 Experiments

Here we present experiments to derive the boundaries of named places using knowledge derived from the web. In order to refine and evaluate the potential of the method it was applied initially to precisely defined places, namely four UK counties of Leicestershire, Hertfordshire, Surrey and Devon. The same methods were then applied to find approximate boundaries of three imprecise UK places, the Cotswolds, Mid Wales and the Highlands of Scotland. Purves et al. (2005) describe the evaluation of imprecise regions involving human participants in a study of the Mittelland in Switzerland, showing a strong correlation between the human-generated boundaries and the boundary generated by the methods described in this paper. In this paper we assess the vague regions by visual inspection.

4.1 Comparison of Query Methods

Before presenting results for particular target regions we present the results of an experiment to determine the most effective form of web query, given the options explained in section 3.1, using the English Midlands as the target region. A total of 1,700 unique candidate locations were extracted from results generated using the following queries:

(Q1) "the Midlands" [Region Only],

(Q2) "hotels" and "the Midlands" [Region & Concept],

(Q3) "places to visit" and "the Midlands" [Region & Concept] and

(Q4) 41 lexical patterns [Region & Pattern].

One of the authors manually checked each location for containment within the Midlands giving a binary score to indicate membership. Note we did not analyse the web pages themselves but only the ranked lists of retrieved places and therefore assume that extracted geo-coded places are used in correct geographic sense (e.g. Rugby the place and not the sport) and ambiguous places all refer to the same place. Table 1 summarises the numbers of extracted locations that lay correctly within the imprecise region for each query. Column 2 shows the numbers of unique place names found for the given query type, column 3 ("Correct") shows the numbers of place names (locations) judged to lie inside the Midlands, while the remaining columns show the numbers of places judged to lie within the Midlands that were found in the top 10 and top 100 ranked extracted place names, using different ranking methods. The column headings TF, DF and F4 refer to the statistics that were used to rank the retrieved place names. TF (term frequency) refers to the number of times that a given place name occurs within all the retrieved web documents. DF (document frequency) refers to the number of documents in which a place name occurs. F4 (Robertson and Spark-Jones 1976) takes account of the frequency (CF) of a place name in the entire collection of documents resulting in higher F4 values when a place name has a lower CF,

i.e. if a place name occurs very frequently irrespective of the query then it should have less significance.

The most successful query types were Q2 and Q3, of the form is "Region + Concept", which returned pages with the most extracted locations and generated the most correct region members. The F4 ranking method provides the most correct locations in the top 10 and 100 compared to ranking with term and document frequency. This indicates that including collection statistics can help to increase the rank position of correct locations.

Table 2 summarises the set of documents collected using the Web search and from column 2 the table shows: the number of documents (web pages) found, the average document length in words, the average proportion of words which are place names (locations), the average number of place names per document and the average percentages of all the locations in a document that were judged to be correct in the sense that they are inside the Midlands. Note that the latter percentages do count multiple occurences of the same name. Query 2 tends to find pages which are a) geographically rich, in that the highest proportion of words are locations and they have the largest number of locations found; b) more fine-grained in that there are for example villages and towns, not just cities; and c) have a more precise geographical *extent*, in that the focus of the *entire* Web page is the vague region. The baseline approach (query 1) retrieves, on average, pages containing the fewest locations. This is typically because personal or institutional home pages are returned which are geographically sparse. The least successful query is 4 which tends to find geographically sparse pages including personal pages, home pages and discussion lists.

4.2 Evaluation with Precise Regions

We now present results based on locations retrieved from web pages resulting from queries of the form "Region + Concept" using English counties as the target region and "hotels" as the concept. The first set of tests measures the number of associated place location points found within the borders of the target administrative regions. Table 3 shows counts for unique associated points for each region and also gives point counts taking repeated points into account (i.e. multiple references to the same place). Between 30% and 50% of unique points retrieved in a web search were found to lie within the target region. Bearing in mind that possible locations are distributed over the whole of the UK, this result suggests that the density of associated points will be much higher within the regions being queried than over the whole of the UK. Figure 1 shows the raw point data for Devon illustrating candidate points lying within and outside the administrative borders of Devon.

This result provides support therefore to the hypothesis that place names that are within a specified region occur in web documents much more frequently in association with the region name than do other names of places that are external to the region.

Surfaces were generated for these regions using a kernel radius of the order of 25km and resolutions of the order of 1km, commensurate with the relatively small size of English counties (for example Devon has a bounding box approximately 75 x 100km). In order to generate boundaries for the regions, a threshold value had to be selected for each surface. By choosing a number of values it is possible to generate polygons representing the regions as a multiple set of sharp regions. For the precise regions modelled here, two threshold values were chosen, 0.25 and 0.5 points per square kilometre. Table 4 shows the area of the administrative units correctly classified, with a threshold of 0.25 points per square kilometre. In every case almost the whole

administrative unit is correctly classified – however, as is shown in Figure 2 this is at a cost of an overestimation of the total area of between 70% and 40%. However, large areas of this over estimation are the result of falsely classified locations which lie completely outside the region. Such outliers could be easily removed, and are discussed more in the section on further work. A further difficulty is shown in the case of Surrey and Hertfordshire, which both lie adjacent to London. Very many British web documents contain references to locations in London and, where the region being derived is also adjacent to London, these will have the effect of *smearing* the region over London. An approach to this problem is also discussed in further work.

4.3 Evaluation with Vague Regions

The methods described for precise regions were applied to the three vague UK regions of the Highlands (of Scotland), the Cotswolds and Mid-Wales.

In Figure 3 we present results which explore the sensitivity of the derived surface to different point attribute values for the Highlands of Scotland, assuming the vague region for this term rather than the precisely defined unitary authority of the same name. In this case, since the initial region is considerably larger than the precise regions previously discussed, a larger kernel with a radius of 50km was used to identify the surfaces. The surfaces shown in Figure 3 are all thresholded with a value of 10% of the maximum surface density.

Figure 3a shows the result of interpolating the surface based only on the density of distribution of the point locations, i.e. locations per unit area. Thus the attribute value for each location is set to

1 and the peak of the surface corresponds to the area where most place names occur. Figure 3b is based on interpolating with the term frequency used as the point attribute value, whilst Figure 3c illustrates the result of using document frequency of place names as the point value. Figure 3b also displays a spurious peak in the surface in the southerly Scottish county of Fife. This is the result of a grounding error, whereby the surname of Cameron, which often occurs with references to the Highlands, is also present in Fife (an example of referent class ambiguity). Apart from this, the highest parts of the surfaces correspond well with the authors' understanding of the location of the Highlands. The highest densities in all cases correspond with the city of Inverness which is a popular tourist centre for the Highlands. It is clear from these results that, unsurprisingly, the use of term frequency as a point attribute can significantly bias results through a falsely assigned place name. The use of points without asociated attribute values (i.e. set to 1) or using document frequency as a value gives broadly similar results, and the following regions are derived through the use of document frequency interpolation.

The results for the Cotswolds are displayed in Figure 4 which shows the extent in 2D for a surface based on a threshold value of 0.125 points per square kilometre, and in 3D for a threshold of 0.5 points per square kilometre. The Cotswolds have been described in Wikipedia as running "through six counties, particularly Gloucestershire, Oxfordshire and southern Warwickshire" being bounded by Oxford in the east. In the figure the large central region corresponds well with this description, as indicated by the administrative boundaries of these counties on the figure. Furthermore, a number of features of the Cotswolds can be detected - in particular the ridge in the surface which could be described as the heart of the Cotswolds running from west to east on the 3D figure. The smaller regions in the 2D figure are clearly beyond any coarse estimate of the

location of the Cotswolds and arise due to the presence of wrongly grounded ambiguous place names as well as a prevalence of locations found in and around London.

The results for Mid Wales are presented in Figure 5. The highest part of the surface corresponds well with the heart of Mid Wales. There is a separate peak to the southwest of Wales which is located in Pembrokeshire and is not part of Mid Wales. This is probably due to the common co-occurrence of references to Pembrokeshire in tourist web pages about Mid Wales. The other peak on the west coast of Wales may be regarded as part of Mid Wales. The trough in the surface between the highest peak and this region is due to the lower density of named places in that area and suggests that, in this case, an approach focused on queries based on co-occurrence of terms with "hotel" is likely to fail, since this region is characterised by a landscape where relatively few hotels are found.

5. Application to Geographical Information Retrieval

The prime motivation of the techniques described in this paper is to generate representations of vague regions that can be stored in digital gazetteers. It will then be possible to process queries that name such vague places by associating them with quantitative geographic regions.

Search engines such as SPIRIT (Jones *et al.* 2004; Purves *et al.* 2007) or Google Local depend upon the use of gazetteers to recognise place names within queries and they would therefore be able to recognise the very large number of vague or vernacular place names that users commonly employ when searching for geographically-referenced information. For example, Figure 6 shows the results of a query sent to the SPIRIT search engine for "hotels in Cotswolds", where the

Cotswolds is represented within the SPIRIT gazetteer as the core region shown in Figure 4. The figure shows how the titles of retrieved relevant documents are listed in geographically and thematically relevance ranked order, along with map symbolisation of their corresponding locations. The approximated boundary of the Cotswolds can be used to identify places that lie inside and in the vicinity of that boundary and to rank the results geographically with respect to distance from the boundary.

6 Conclusions and Future Work

Place names are used commonly for purposes of information retrieval, yet many place names are vague in the sense that they do not have precisely defined boundaries. It is important therefore that these vague places can be recognised within natural language queries and correctly interpreted with regard to their spatial extent. In this paper we have described and evaluated a set of techniques that model the extent of such regions as a statistical density surface that can be used to generate approximate sharp boundaries if required. The methods are based on retrieval of the locations of places, the names of which are commonly associated with the target region in text documents on the web. An initial evaluation of the techniques was carried out by thresholding the density surface to obtaining boundaries for precise target regions and comparing these boundaries with the known borders for these regions.

The resulting regions showed good agreement with known borders, though in general they are somewhat larger than the known regions' borders. These techniques are similar to the empirical approaches described by Montello *et al.* (2003), but crucially use the web instead of human subjects as a data source. Such an approach allows rapid collection of large datasets of candidate

points for many vague regions, which would in turn facilitate populating gazetteers with boundary information for such regions.

When applied to vague regions, the method gave results that are in agreement with the authors' notion of these regions, and initial experiments with human subject testing by Purves *et al.* (2005) suggest that extents calculated in this way are plausible.

The quality of the results produced is sensitive to the quality of geo-parsing of the retrieved text and the subsequent grounding of detected place names. For some places this process can result in locally low confidence values of the modelled surface even within locations that appear well within the confines of the target region. This will occur when there are few references to places found in that part of space, due for example to low population levels. In future work, the web search methods will address the problem of low population levels and low popularity of locations by formulating queries that refer to topographic features in addition to named populated places. Thus the web queries could use concepts such as hill and river, rather than just "hotels" as used in this study. Furthermore, certain locations, such as a country's capital city are very likely to appear in very many documents, even though they are not relevant to the vague region itself. Several other approaches may improve results here. Place names that appear with similar frequencies in the results of all queries for a specific country might be automatically discarded. Improved identification of only those place names which are potentially relevant to the vague region in the trigger phrase would enable automatic filtering of the spurious locations. Thus it would be possible to identify a prime region resulting from the analysis and automatically remove relatively distant regions (outliers) that had no topological connection to the primary

region. This process could be accompanied by a spatial window that restricted the extent of the surface model.

The methods described to interpolate surfaces are not fully automated in that there is some interactive setting of relevant parameters such as kernel radius, surface resolution and threshold value, albeit based on guidelines developed from the study of precise places. Future work will investigate higher degrees of automation of such parameter setting, which might be based for example on machine learning methods using training data based on expert knowledge of vague places. It should be noted however that the density surface model could be exploited in its own right as a quantified vague model and used as such within geographical information retrieval systems to assess spatial relevance to the given place.

ACKNOWLEDGEMENTS

This research is supported by the EU-IST Project No. IST-2001-35047 (SPIRIT) and the Swiss BBW. Ordnance Survey gazetteer data was provided by Digimap.

REFERENCES

AITKEN, S.C. and R. PROSSER,1990, residents' spatial knowledge of neighbourhood continuity and form. *Geographical Analysis*, 22, 301-25.

ALANI, H., C.B. JONES, and D.S. TUDHOPE, 2001,Voronoi-Based Region Approximation for Geographical Information Retrieval with Gazetteers. *International Journal of Geographical Information Science*. 15(4), 287-306.

ARAMPATZIS A, M., M. VAN KREVELD, M., I. REINBACHER *et al***,** 2006, Web-Based Delineation of Imprecise Regions, *Computers Environment and Urban Systems* 30 (4), 436-459.

BRIN, S. and L. PAGE, 1998, The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference (WWW7)*, Brisbane, Australia, pp. 14-18,.

BRUNSDON, C., 1995, Estimating probability surfaces for geographic point data: an adaptive kernel algorithm. *Computers and Geosciences*, 21 (7) 877-894.

BYRKIT, J.W., 1992, Land, Sky, and People: The Southwest Defined. University of Arizona Press, Tucson. http://digital.library.arizona.edu/jsw/3403/index.html

CLARKE, C.L., A., CORMACK, G.V. and LYNARN, T.R., 2001, Exploiting Redundancy in Question Answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, pp. 358-365.

COHN, A.G. and N.M. GOTTS, 1996, The 'egg-yolk' representation of regions with indeterminate boundaries. Geographic objects with indeterminate boundaries, P.A.Burrough and A.U. Frank, Editors. Taylor and Francis, pp. 171-88.

COWIE, J. and W. LEHNERT, 1996, Information extraction. *Communications of the ACM*. 39(1), 80-91.

CUNNINGHAM, H., et al., 2002, GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of 40th Anniversary Meeting of Association for Computational Linguistics* (ACL'02).

EGENHOFER, M., 2002, Toward the Semantic Geospatial Web. In *Proceedings of ACM-GIS'02*, A. Voisard and S.-C. Chen (eds.), pp. 1-4.

FISHER, P., 2000, Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, 113(1), 7-18.

FISHER, P., J. WOOD and T. CHENG, 2004, Where is Helvellyn? Fuzziness of Multiscale Landscape Morphometry. *Transactions Institute of British Geographers*. 29(1), 106-28.

GOLD, C.M., J. NANTEL and W. YANG, 1996, Outside-in: An alternative approach to forest map digitizing. *International Journal of Geographical Information Science*, 10(3) 291-310.

HARPRING, P., 1997, Proper words in proper places: The Thesaurus of Geographic Names. *MDA Information.* 2(3), 5-12.

HILL, L.L., 2006, *Georeferencing: The Geographic Associations of Information (Digital Libraries & Electronic Publishing)*. MIT Press.

HILL, L.L., J. FREW, and Q. ZHENG, 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., A.I. ABDELMOTY, and G. FU, 2003, Maintaining ontologies for geographical information retrieval on the web. In *On the Move to Meaningful Internet Systems : Proceedings of ODBASE'03*, Lecture Notes in Computer Science 2888, (Berlin: Springer), pp. 934-51

JONES, C.B., H. ALANI, and D.S. TUDHOPE, 2001, Geographical information retrieval with ontologies of place. In *Spatial Information Theory: Foundations of Geographic Information Science, Proceedings of COSIT'01*, D.R. Montello (Ed), Lecture Notes in Computer Science, 2205, (Berlin: Springer), pp. 336-51.

JONES, C.B., A.I. ABDELMOTY, D. FINCH *et al*., 2004, The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proc Third International Conference on Geographic Information Science GIScience*, M. Egenhofer, C. Freksa, H. Miller (eds.), Lecture Notes in Computer Science 3234, (Berlin: Springer), pp125-139.

KILGARRIFF, A. and G. GREFENSTETTE, 2003, Web as corpus. *Computational Linguistics*, 29, 1-15.

KUHN, W., 2001, Ontologies in support of activities in geographic space. *International Journal of Geographic Information Science*, 15, 613-31.

KULIK, L., 2001, A geometric theory of vague boundaries based on supervaluation. in Conference on Spatial Information Theory, In *Spatial Information Theory: Foundations of Geographic Information Science; Proceedings of COSIT'01*, D.R. Montello (Ed), Lecture Notes in Computer Science, 2205 , (Berlin: Springer), pp. 44-59.

KWOK, C.C.T., O. ETZIONI and D.S. WELD, 2001, Scaling Question Answering to the Web. In *Proceedings of the 10th International World Wide Web Conference (WWW10)*, 1-5 May, Hong Kong, China. pp. 150-161. ACM, Hong Kong, China.

LAM, C.S., J.P. WILSON, and D.A. HOLMES-WONG, 2002, Building a Neighborhood-Specific Gazetteer for a Digital Archive. Pap0300. ESRI User Conference. Conference, http://gis.esri.com/library/userconf/proc02/pap0300/p.htm

LARSON R.R., 2003, Placing cultural events and documents in space and time. M. Duckham, M.F. Goodchild and M.F. Worboys (eds), *Foundations of Geographic Information Science*, Taylor and Francis, pp. 223-239.

LI, H., 2003, InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In *Proceedings of HLT-NAACL 2003 Workshop on Analysis*.

MCDONALD, D., 1996, Internal and external evidence in the identification and semantic categorisation of proper names, in Corpus Processing for Lexical Acquisition, B. Boguraev and J. Pustejovsky, Editors. MIT Press: Cambridge, MA, pp. 21-39.

MONTELLO, D., *et al*., 2003, Where's downtown?: behavioural methods for determining referents of vague spatial queries. *Spatial Cognition and Computation*. 3(2&3), 185-204.

O' SULLIVAN, D. and D.J. UNWIN, 2003, Geographic Information Analysis. Wiley, London

PURVES, R., P. CLOUGH, and H. JOHO, 2005, Identifying imprecise regions for geographic information retrieval using the web. In Billen, R., Drummond, J., Forrest, D., and João, E. (eds), *Proceedings of the GIS RESEARCH UK 13th Annual Conference*, Glasgow, UK, pp. 313-318.

PURVES, R, P. CLOUGH, C.B. JONES *et al.*, 2007, The Design and Implementation of SPIRIT: a Spatially-Aware Search Engine for Information Retrieval on the Internet, *International Journal of Geographical Information Science*, 21(7), 717 – 745.

RADEV, D.R, K. LIBNER and W. FAN, 2002, Getting answers to natural language questions on the web. *Journal of the American Society for Information Science and Technology*, 53(5), 359-364.

ROBERTSON, S. E. and K. SPARCK-JONES, 1976, Relevance Weighting of Search Terms. *Journal of the American Society For Information Science*, 27, 129-146.

RAUCH, E., *et al.*, 2003, A confidence-based framework for disambiguating geographic terms. In *Proceedings of HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 50-4.

SANDERSON, M., J. KOHLER, 2004, Analyzing geographic queries. In *Proceedings of Workshop on Geographic Information Retrieval SIGIR*, 2004.

SCHLEIDER, C., T. VOGELE, and U. VISSER, 2001, Qualitative spatial representations for information retrieval by gazetteers. In *Proceeding of COSIT'01*, D.R. Montello (Ed), Lecture Notes in Computer Science 2205, (Berlin: Springer), pp. 336-51.

SMITH, D.A. and G.S. MANN. 2003, Bootstrapping toponym classifiers. In Proceedings of HLT-NAACL 2003 Workshop on Analysis of Geographic References, pp. 45-9.

VOGELE, T., C. SCHLIEDER, and U. VISSER, 2003, Intuitive modelling of place names for spatial information retrieval. In *Proceedings of COSIT'03*, W. Kuhn, M.F. Worboys and S. Timpf (Eds) Lecture Notes in Computer Science 2825, (Berlin: Springer), pp. 239-52.

WANG, F., 1994, Towards a natural language user interface. International Journal of Geographical Information Science. 8, 143-62.

WORBOYS, M.F., 1998, Imprecision in finite resolution spatial data. *Geoinformatica*, 2, 57-79.

ZHANG, V.W., B. REY, E. STIPP and R. JONES, 2006, Geomodification in Query Rewriting. *Workshop on Geographical Information Retrieval, ACM SIGIR 2006*, http://www.geo.unizh.ch/~rsp/gir06/papers/individual/zhang_jones.pdf

Table 1 Summary of extracted locations for each type of query, givings the numbers of unique locations (place names), the numbers of locations judged to lie inside the "Midlands" ("Correct") and the numbers of these "correct" locations found in the top 10 and top 100 ranked lists of locations, using different ranking methods (see text for explanation of ranking methods)

			Correct in top 100			Correct in top 10		
Query	Unique	Correct	TF	DF	F4	TF	DF	F4
Q1	291	116	41	50	60	4	5	10
Q2	549	285	62	62	72	7	5	10
Q3	655	308	48	40	48	4	2	7
Q4	205	51	23	31	41	2	3	8

Table 2 Summary of retrieved documents recording, for each query type, the numbers of documents retrieved and the averages of, respectively, words per document, percentage of words that are locations (places), numbers of locations per document and the percentage of those (non-unique) locations judged to lie inside the "Midlands"

Query	Docs	Words/doc	%Locations	Locns/doc	%Correct
Q1	100	630	3.2%	22	51%
Q2	100	1156	6.9%	81	57%
Q3	100	1528	2.8%	42	41%
Q4	47	1637	1.8%	28	35%

Table 3 Evaluation of membership of associated points within administrative units. Values give count of unique locations (and counts multiplied by their frequency).

Region	Membership	
	Points *inside* region	Points *outside* region
Leicestershire	310 (1425)	653 (3189)
Hertfordshire	213 (1253)	592 (2605)
Surrey	225 (1109)	660 (3469)
Devon	358 (3667)	954 (5488)

Table 4 Percentage agreement between areas of derived boundaries and the corresponding administrative boundaries.

County	Area of county classified	Area of county not classified
Devon	98%	2%
Surrey	100%	0%
Leicestershire	99%	1%
Hertfordshire	99.9%	0.1%

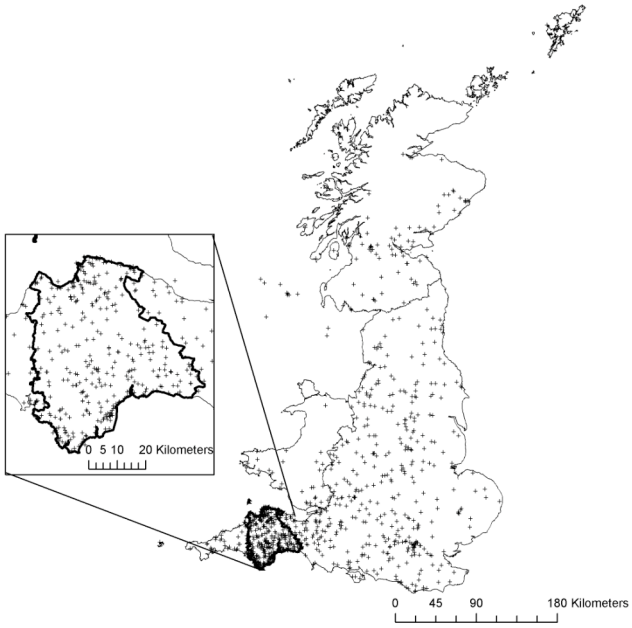


Figure 1 Point set retrieved for top 100 documents for query "~hotels Devon". Boundary of administrative region corresponding to Devon is shown.

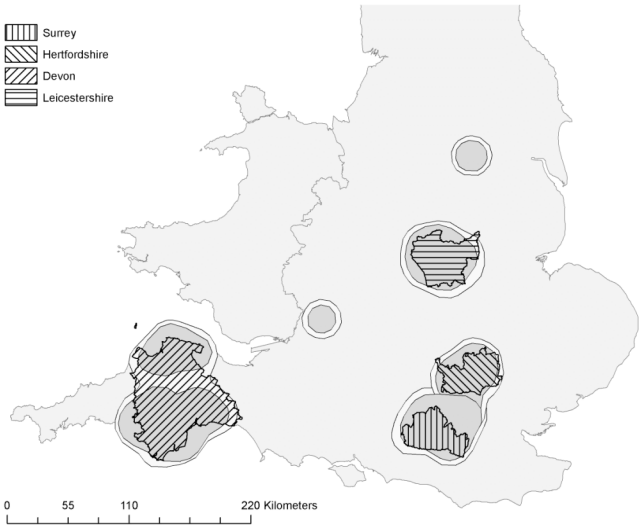
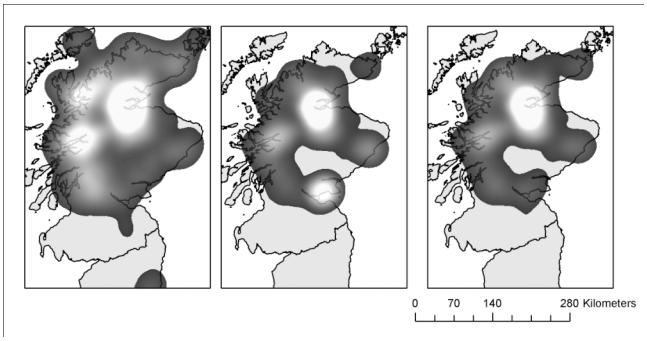


Figure 2 Four precise regions and actual boundaries for two threshold values of derived polygons



(a) (b) (c)

Figure 3 The Highlands using density surfaces generated with a value of 1 for each unique point (a), a value according to term frequency (b) and a value of the document frequency (c)

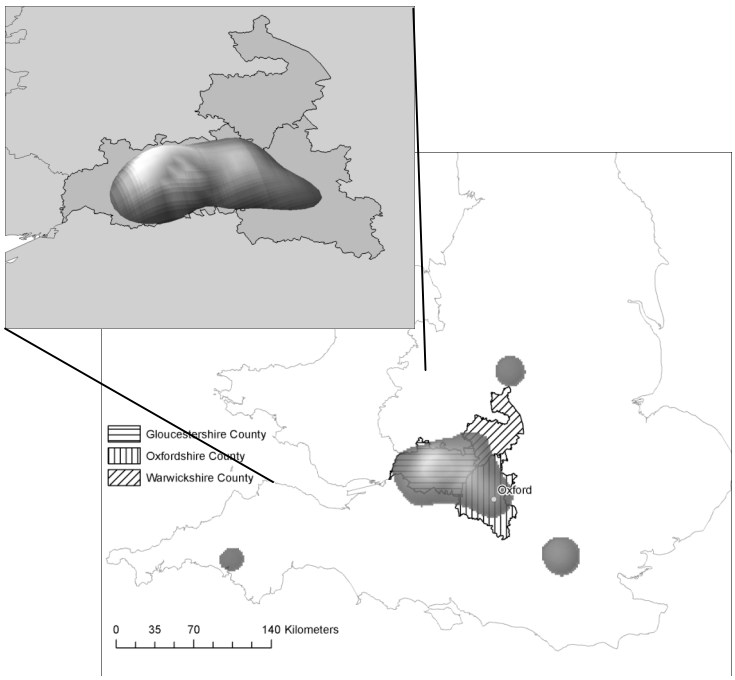


Figure 4 Cotswolds and correspondence with three English counties in 2D and 3D

Figure 5 Surface derived for Mid Wales with a threshold of 0.125

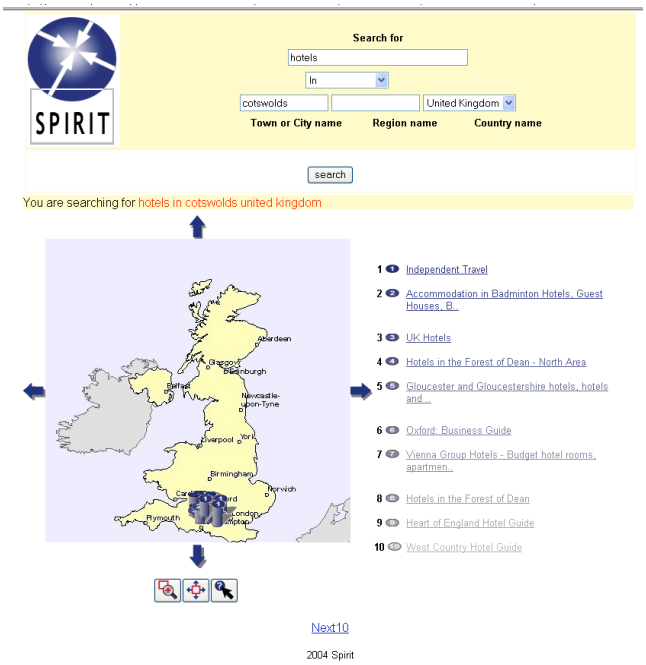


Figure 6 Results of a search using the SPIRIT system for “hotels in Cotswolds”