ABSTRACT

In this paper, we present a method for deriving an approximate spatial boundary for imprecise or vague geographical regions such as the English Midlands. Geographic knowledge is crucial for any geographic information system (GIS) and often provided through gazetteers and geographical thesauri. However, these resources do not contain entries for vague regions even though common in natural language. We model spatial boundaries by deriving approximate polygon boundaries formed using geographic locations extracted from pages retrieved by searching the web. The method is evaluated using both precise and vague geographic knowledge acquired from the Web. Evaluation on both precise and vague geographical regions shows promising results. 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 or at all with the places encoded in conventional gazetteers. There are many vague or imprecise places such as the English Midlands, the South of France, the Rocky Mountains, and the Mid West, for which no politically official boundary exists and there are other places which may be used for administrative names, but for which the administrative boundary interpretation may differ from many people’s perception 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. There are several types of 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 (e.g. Google Local). Applications that provide information access via place names usually make use of gazetteer-based resources to recognize their presence and to resolve ambiguities in names for which there are multiple occurrences. 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 [1,2,9,10,11].

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 or at all with the places encoded in conventional gazetteers. There are many vague or imprecise places such as the English Midlands, the South of France, the Rocky Mountains, and the Mid West, for which no politically official boundary exists and there are other places which may be used for administrative names, but for which the administrative boundary interpretation may differ from many people’s perception 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 [4]. 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 evaluate our method for generating approximate boundaries for vague places from knowledge acquired from the Web. Evaluation on both precise and imprecise regions shows promising results and application of this work include Geographic Information Retrieval (GIR) systems such as SPIRIT [12].

Keywords

Imprecise regions, Geographic Information Systems (GIS), Geographic Information Retrieval (GIR)
their search [17]. However, using the web as a geographic knowledge source for deriving region boundaries is not well researched. Most work so far on imprecise 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 [18]. For example, in an investigation of people’s perception of the vague place “downtown Santa Barbara” Montello et al [19] 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.

A method that depends heavily upon the text found on maps is described by Lam et al [21] 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 [22] 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.

In a theoretical approach to the delineation of topographic features, Fisher et al [20] showed how extent could be delineated automatically by analysing terrain models for features such as peaks and passes across a range of scales. Although many locations were classified as more than one type of feature, according to the scale of 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 a database of names for topographic features (e.g. mountain peaks) exists.

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, of course, the web. Data mining techniques to identify texts mentioning vague place names in association with other place names are a potential approach which have been subject to an initial set of experiments by Arampatzis et al [23], who have performed some preliminary experiments to search for such associations using a web search engine. Their method depends upon access to 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 mining methods in this paper (see Section 3 onwards).

2.2 Modelling the Boundaries of Vague Places with Empirical Data
Having identified as possible candidates, locations which may lie within or outside an imprecise region, it may be necessary to derive measures for whether any other location lies within the imprecise region. By definition a vague place cannot be expected to have a single precise boundary. However, as indicated above, 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 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. [19] 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 [24] 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
A qualitative approach to representing vague places has been presented by Vogele et al [11]. Imprecise 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, following Worboys [26]. 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.

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 [27] 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 cells. The Voronoi cell edges that lie between these two sets constitute an approximate boundary.

If a set of candidate points for membership of an imprecise region exist, 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 approach was adopted by Purves et al [28] 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 weighted according to different measures, and some distance weighted interpolator applied, whereby for each interpolated point the nearest neighbouring sample points are summed to created a weighted average.

3. GENERATING REGION BOUNDARIES
Web pages often contain references to place names in order to provide geographic context. The place names used include names associated with well-defined boundaries and those with vague boundaries. If a document contains a place name it is often the case that there will be multiple occurrences of other names referring to a particular region of interest and to possibly disparate places. It is hypothesised that the co-occurrence of frequent place names may be used to find the identities of relatively precise places that are associated with, and ideally inside, vague places [23,28]. 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 imprecise place names as the target regions. By using precise place names we can evaluate the efficacy of our method in modelling precise regions, before moving on to consider how the method can be extended to imprecise regions. To determine the approximate boundary of either a precise or imprecise region from web search, we perform the following:
1) Search the web for pages containing 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 place and use it to extract an approximate boundary.

3.1 Searching the Web
To gather candidate region members, queries were submitted to Google containing a reference to a target region. The goal of searching is to find 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. 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 (this approach was used in [28]);
- **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 [23]).

During initial experiments, queries based on region and concept appeared to find the most geographically-rich pages, e.g. directory listings. This 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 related to the target region (e.g. contact-us pages), these approaches also generated far fewer locations. Importantly, our 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 the UNIX lynx command was used to extract plaintext which we found to help to reduce 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 ANNIIE, the default Information Extraction (IE) system that comes with GATE (General Architecture for TEXT Engineering) [29,30]. ANNIIE was used to perform NER using both internal and external evidence (see [31]) in the form of gazetteer lists (e.g. common names of people and places) and context rules to disambiguate between named entities (also called referent class ambiguity [32]). For example, if we found the sequence “<Forename> Location=” where Location and Forename exist in the gazetteer lists, 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\textsuperscript{1} (Seamless Administrative Boundaries of Europe) dataset and (2) the Ordnance Survey 1:50,000 Scale Gazetteer\textsuperscript{2}. 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 also identified postcodes which could also be used to provide valuable spatial information. Effective geo-parsing is a challenging task and the methods employed in the present work have considerable scope for improvement. The emphasis here has been upon an initial evaluation of the potential for using these methods for delineating vague places. Improved geo-parsing would then result in improved results for co-locating place names.

### 3.3 Assigning Coordinates (Grounding)

Once has 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. We also used the Getty Thesaurus of Names (TGN\textsuperscript{3}), a hierarchical geographical thesaurus of over 1 million names to provide global geographical knowledge. Having additional knowledge helps with the disambiguation of places where the same name is used in different countries. For example, 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 assigning 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 (e.g. as organisation or person names) [32]. 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 [32], by population of the place name [33] or by semi-automatic extraction from the Web [34]. 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 (e.g. World > Europe > United Kingdom > England > Lancashire). If matches between the local contexts are not found, then places are assigned a default sense (or coordinate). In these experiments we assign the default sense to the “largest” location based on feature types and hierarchy depth as provided by the gazetteers. UK postcode data is particularly useful as these are unambiguous and thereby introduce less error.

### 3.4 Generating Approximate Boundaries

The first step in generating approximate boundaries was to derive density surfaces modelling the distribution of point references associated with a target region. We used a standard interpolation technique, kernel density estimation, where density values are interpolated over a regular grid from the irregularly distributed dataset of point data [35]. Points can be assigned weights in the density estimation, for example term frequency (the total number of occurrences of a point) or document frequency (the number of documents a point occurs in). Having derived a surface, if sharp boundaries are required, threshold values for membership of a region must be selected.

This approach requires selection of a number of parameters, namely surface resolution, kernel radius and threshold point density for selection of sharp boundaries. Resolution of the surface (i.e. grid cell size) 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 these experiments kernel radius was between a half, and an order of magnitude less than, the maximum diameter of the regions under investigation. Finally, threshold point density (required to generate an approximate boundary) was identified 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.

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;
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 the three imprecise UK places, the Cotswolds, Mid Wales and the Highlands of Scotland. Purves et al. [28] describe the evaluation of imprecise regions involving human participants showing a strong correlation between the human-generated boundaries and the boundary generated by the methods described in this paper for the Mitteland in Switzerland. In this paper we assess the imprecise regions by visual inspection.

<table>
<thead>
<tr>
<th>Region</th>
<th>Points inside region</th>
<th>Points outside region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leicestershire</td>
<td>310 (1425)</td>
<td>653 (3189)</td>
</tr>
<tr>
<td>Hertfordshire</td>
<td>213 (1253)</td>
<td>592 (2605)</td>
</tr>
<tr>
<td>Surrey</td>
<td>225 (1109)</td>
<td>660 (3469)</td>
</tr>
<tr>
<td>Devon</td>
<td>358 (3667)</td>
<td>954 (5488)</td>
</tr>
</tbody>
</table>

\[\textsuperscript{1}\text{http://www.eurogeographics.org/eng/03_projects_sabe.asp}
\textsuperscript{2}\text{http://www.ordnancesurvey.co.uk/oswebsite/products/50kgazetteer/}
\textsuperscript{3}\text{http://www.getty.edu/research/conducting_research/vocabularies/tgn/}
4.1 Evaluating Precise Places

The first set of tests measured the number of associated place points found within the borders of the target administrative regions. Table 1 shows counts for unique associated points for each region and also gives point counts taking repeated points into account. Between 30 and 50% of 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.

![Figure 1: Point set retrieved for top 100 documents for query "hotels Devon" – boundaries of administrative region corresponding to Devon are shown.](image1)

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 places that are external to the region.

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 2 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 an overestimation of the total area of between 70 and 40%. However, large areas of this overestimation 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 discussed in further work.

![Figure 2: Four precise regions and actual boundaries for two threshold values of derived polygons](image2)

### Table 2 Percentage agreement between areas of derived boundaries and the corresponding administrative boundaries.

<table>
<thead>
<tr>
<th>County</th>
<th>Area of county classified</th>
<th>Area of county not classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devon</td>
<td>98%</td>
<td>2%</td>
</tr>
<tr>
<td>Surrey</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Leicestershire</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>Hertfordshire</td>
<td>99.9%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

4.2 Application to Imprecise 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 surface derived to different weighting terms for the Highlands of Scotland, assuming the vague region for this term rather than the unitary authority. 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. Thus term frequency is ignored and the peak of the surface corresponds...
to the area where most named points occur. Figure 3b is based on interpolating with the term frequency used as a weight, whilst Figure 3c illustrates the result of weighting using document frequency of place names. Figure 4b also displays a spurious peak in the surface in the southerly county of Fife. This is the result of a grounding error, whereby the surname of Cameron, which occurs in the Highlands, is also present in the southerly Scottish county of 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 weighting term can significantly bias results through a falsely assigned place name. The use of unweighted points or document frequency gives broadly similar results, and the following regions are derived through the use of document frequency weighted 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 as running “through six counties, particularly Gloucestershire, Oxfordshire and southern Warwickshire”^4^ 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 described - in particular the ridge 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 duplicate place names – for example the area to the south east is once again a consequence 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. This result suggests that consideration be given to the use of snippets in assessing the association of place names with imprecise region names in web pages (i.e. limiting the extent to which a geographic reference applies). 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, at least in part, where few hotels are found.

---

^4^ http://en.wikipedia.org/wiki/Cotswold
5. APPLICATION TO GIR
The aim of the techniques described in this paper, is to extract automatically from web data representations of vague or imprecise regions for storage in digital gazetteers. Having identified a boundary, or a set of boundaries, these can then be used in natural language queries for geographical information retrieval.

Search engines such as SPIRIT [12], within whose research framework this work was undertaken, or Google Local, which use query interfaces based on natural language, can usefully use such interpretations. 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.

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. However, any systems allowing queries to specify regions with natural language should allow such names to be interpreted when they are used to describe or to enquire about geographically-specific information. In this paper we have described and evaluated a set of techniques which identify the borders of such regions based on selection of an initial set of candidate locations through web queries, and derivation of a surface representing these regions through interpolation. An initial evaluation of the techniques was carried out by obtaining boundaries for precise regions by thresholding surfaces 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 region’s borders. These techniques are similar to the empirical approaches described by Montello et al [19], 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 imprecise regions, which suggests it may be suitable for use in populating gazetteers with boundary information for such regions.

When applied to imprecise regions the method gave results which are in agreement with the authors’ notion of these regions, and initial experiments with human subject testing [28] suggest that the extents calculated are plausible. Further work is underway with human subject testing, to further compare our results with those gained from human subjects.

The quality of the results produced is sensitive to the quality of geoparsing 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 places found in that part of space, due for example to low population levels. In future work the web search methods will be extended to find named topographic features in addition to named populated places. 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 imprecise region itself. Three approaches may improve results here. Firstly, place names which appear with similar frequencies for any query in a specific country might appropriately be automatically discarded. Secondly, better identifying only place names which are directly relevant to the imprecise region in the trigger phrase would automatically filter such locations. Finally, automatically removing distant regions (outliers) with no topological connection to the imprecise region should be investigated.

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.

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

8. REFERENCES
[34] Li, H., InfoXtract location normalization: a hybrid approach to geographic references in information extraction. HLT-NAACL 2003 Workshop on Analysis.