Web-based Delineation of Imprecise Regions

Avi Arampatzis, Marc van Kreveld, Iris Reinbacher
Institute of Information and Computing Sciences, Utrecht University
PO Box 80.089, 3508TB Utrecht, The Netherlands
Tel: +31 (30) 253 9128, Fax: +31 (30) 251 3791
{avgerino,marc,iris}@cs.uu.nl

Christopher B. Jones, Subodh Vaid
School of Computer Science, Cardiff University, UK
{c.b.jones,subodh.vaid}@cs.cardiff.ac.uk

Paul Clough, Hideo Joho, Mark Sanderson
Department of Information Studies, University of Sheffield, UK
{p.d.clough,h.joho,m.sanderson}@sheffield.ac.uk

Abstract

This paper describes several steps in the derivation of boundaries of imprecise regions using the Web as the information source. We discuss how to use the Web to obtain locations that are part of and locations that are not part of the region to be delineated, and then we propose methods to compute the region algorithmically. The methods introduced are evaluated to judge the potential of the approach.

Keywords

Geographical Information Systems (GIS), World-Wide Web (WWW), Imprecise Regions,
Fuzzy Boundaries, Geometric Algorithms.

1. Introduction

The World Wide Web is a major source of geographical information. Much of the existing technology for accessing geographical data is based on structured digital map data within geographical information systems (GIS) and as such is not well adapted to the unstructured, largely text-based resources of the Web. Documents on the Web can be categorized geographically according to the their textual content (Smith, 2002), but there are considerable challenges in interpreting the geographical terms. A major problem is the vague and imprecise nature of place names that are commonly employed within documents and by users of the Web when formulating a query. Terms such as “Midwest” in the US and “Midlands” in the UK have no formal geometric boundary and may be interpreted differently by different people. Because such imprecise terms occur so frequently, it is an important challenge to develop techniques to approximate their extent in a manner that enables them to be interpreted intelligently for purposes of information retrieval.

Once an approximate boundary has been determined it can be stored for subsequent use within a GIS database or a geographical ontology such as a gazetteer. An attempt in this direction has been made in the SPIRIT project (Jones et al., 2002; http://www.geo-spirit.org/). Using a structured query interface, a user may ask the SPIRIT search engine for hotels in the Midlands, or castles in the Cotswolds.

This paper describes several steps in the derivation of boundaries of imprecise regions using the Web as the information source (see also (Markowetz et al., 2003)). We discuss how to obtain locations that are part of and locations that are not part of the region to be delineated, and then we propose methods to compute the region algorithmically in Section 2. In Section 3, we present experimental results that show how well our approach works. Both precise (Wales) and imprecise (East Anglia, Midlands, and South East) regions in the UK are used during evaluation. In Section 4 we provide a discussion of the approach and directions for future research.

2. A method for deriving a boundary for imprecise regions

We use the following steps to determine a possible boundary for an imprecise region:

1. Use the Web to find points (cities, towns) inside the unknown region (Section 2.1).
2. Find the coordinates of these points, and a bounding box. Use the bounding box to find coordinates of other cities and points, apparently lying outside (Section 2.2).
3. Compute a boundary of the imprecise region using the points in- and outside (Section 2.3).

This approach is based on the idea that humans, who are responsible for the contents of Web pages, have an idea which locations are part of the imprecise region. The cognitive reasons or attributes for the choice of categorizing a location as inside are not known and not important. Categorization for geography is for example described by MacEachren (1995).

2.1 Using the Internet to obtain references to geographical locations

The first step in our approach is to generate a candidate list of geographic references or *geo-references*, which are primarily in the form of place names. We do this by searching for documents on the Web that contain a reference to the imprecise region, also referred to as the *target region*. We assume that geo-references which co-occur will in some way be related to the region and enable us to define it spatially.

Web searching is performed using the Google search engine, accessed through the Google API (http://www.google.com/apis). A set of trigger phrases is used to query Google and obtain a set of search results (Section 2.1.1). From these, geo-references are extracted (Section 2.1.2) and assigned spatial coordinates automatically (Section 2.1.3), a process known as *geo-parsing* and *geo-coding* respectively (Larson, 1996).

2.1.1 Performing Web page searches using trigger phrases

To find a set of member places within a geographic region, we use a set of linguistic patterns called *trigger phrases* to capture geographical relationships. For example, membership could be identified using the pattern “is a city in” (e.g., “Birmingham [is a city in] the Midlands”). Trigger phrases are used for searching, rather than the name of the imprecise region itself, because it helps to create more geographically-focused queries. The outcome is for documents containing geo-references within the target region to be ranked highly in the search results.

Trigger phrases are used to capture regular linguistic patterns, which identify relationships between geographic locations. From a linguistic point-of-view, one can think of these patterns as lexico-grammatical frames (Moon, 1998) where word (or *lexis*) order is fixed and used within a fixed structure (a *frame*). For example, the trigger phrase "X is located in Y" will typically extract X and Y as noun phrases which identify places (e.g., "Birmingham is located in the Midlands"). By defining these patterns as regular expressions, we can capture specific information about a target region. Table 1 lists the trigger phrases used in these experiments (where R is the target region, "*" matches anything and [X | Y] will match X or Y). These patterns have been generated through our initial investigations and from previous work on question-answering (Joho and Sanderson, 2000), (Joho et al., 2001), (Dumais et al., 2002). Although these patterns are generic and could be filtered to match country-specific geographical regions (e.g., *county* in the UK and *province* in France), for simplicity we use all patterns when searching.

Each trigger phrase is submitted to the Google API[1] as a search request using quotes to match the pattern as an entire phrase (e.g., "* is located in the South East"). Search results follow a standard format and contain the following metadata: page title, followed by a brief extract from the site (called a *snippet*), the page URL and links to a cached version of the page and similar pages if found (see Figure 1). For each search up to 100 results are retrieved. We extract the title and snippet text and merge the results from different searches together to create a single set of results. In the merging process, duplicate results are removed based on the URL and snippet text. Metadata from the search results is used to find candidate region members rather than the Web pages themselves because: (1) the snippet captures the local context of the target region in the Web page thereby generating more likely region members and (2) downloading and parsing the Web pages takes much longer than using the metadata itself. In Figure 1, both the title and snippet contain suitable geo-references for the search "* is located in the Midlands", these are Birmingham and Tamworth respectively.

Google snippets and trigger phrases have been used successfully before in tasks such as question-answering (Joho and Sanderson, 2000). One of the reasons behind the success of such approaches is due to the use of large amount of texts that are indexed by Web search engines. While the occurrence of trigger phrases can be rare, we only need a couple of matching sentences to extract related names/descriptions.

2.1.2 Extracting geo-references from Web page metadata

Given a set of search results, we extract geo-references from the title/snippet (geo-parse) and ground them (geo-code). For extraction, we use a version of the GATE (General Architecture for Text Engineering, `http://gate.ac.uk/`) information extraction (IE) system (Cunningham et al., 2002). GATE provides a framework (in Java) within which to develop custom Language Engineering (LE) applications. The system provides a Collection of REusable Objects for Language Engineering (CREOLE), a reusable family of language and processing resources such as a default IE system called ANNIE (A Nearly New Information Extraction system).

GATE is highly flexible and enables us to perform both gazetteer lookup and language-dependent processing, such as co-reference resolution and semantic tagging. This helps to deal with ambiguity between named entities (e.g., between locations and people). This is known as referent class ambiguity and proves problematic when geographical names overlap with names of organisations, people, buildings, etc. We use a default version of GATE (version 2.2), which includes limited gazetteer lists of global regions. To improve geo-parsing and enable us to ground locations, we use two specific UK resources: (1) the SPIRIT ontology, and (2) a gazetteer list from the UK Ordnance Survey (OS) company. In addition, we have also adapted grammar rules for semantic annotation to capture organisational names beginning with a location. Using only text identified as locations with the IE system, we would otherwise miss annotations containing potentially useful geo-references such as “Cardiff City Council” or “Cambridge University”.

The SPIRIT ontology (Jones et al., 2003) is based on SABE (`http://www.eurogeographics.org`) data and contains 10,275 unique UK names of which approximately 10% are ambiguous. Locations include regions such as towns, cities and counties represented spatially as polygons. Places are defined by a

[1] Searches are submitted to google.co.uk. Therefore a search on South East will tend to return results for the South East of England. This keeps the pattern as general as possible.

geographical hierarchy (e.g., /United Kingdom/England/Sheffield/Bromhill). The OS resource used is the gazetteer list from the Landranger© 1:50,000 scale map (`http://www.ordnancesurvey.co.uk/oswebsite/products/50kgazetteer/`). This contains about 260,000 UK locations defined by type such as town, city, water feature, hill and place of interest. Approximately 10.9% of names are ambiguous and we use a subset of 80,635 names based on the features: city, town, other settlement (e.g., village), antiquity (e.g., Stonehenge), forest, hill and water. We also parsed the name list to improve gazetteer lookup by removing text in parenthesis (e.g., Ackling Dike (Roman Road) → Ackling Dike), create separate entries for alternate place names listed together (e.g., Scalpay/Scalpaigh → Scalpay and Scalpaigh) and expanded abbreviations (e.g., Trentham Gdns → Trentham Gardens). The OS data contains spatial coordinates in point form.

2.1.3 Assigning coordinates to extracted geo-references

After extracting geo-references we assign them spatial coordinates. In some cases, the same name can refer to multiple locations. This is called *referent ambiguity* and can occur for places between different countries (e.g., Sheffield exists in the UK and US) or within the same country (e.g., Cambridge in the UK appears four times in the OS gazetteer list: Scottish Borders, Leeds, Cambridgeshire, and Gloucestershire). For disambiguation, we apply the notion of a default sense. That is, we predetermine a default location for an ambiguous place name. This can be achieved using, e.g., 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 et al., 2003).

In our experiments, we assign the default sense to the "largest" location as estimated from information provided by the geographical resources. For the SPIRIT ontology (based on SABE), which organizes places in a hierarchy of administrative levels, we use the location with the shortest hierarchy. For example, between Cambridge with the hierarchy United Kingdom>England>Cambridgeshire (depth 4) and Cambridge with hierarchy United Kingdom>England>Gloucestershire>Stroud (depth 5), our approach would select Cambridgeshire as the default sense. For the OS gazetteer we order coordinate references by their feature type (e.g., city → town → village). For multiple coordinates with the same feature type we order these randomly. Our method of disambiguation is very simple and therefore susceptible to errors. For example, maybe an ambiguous location should be assigned the coordinates of a smaller geographical region and not the largest, or maybe the UK version of Google returns results for geographic regions that are not in the UK (e.g., the Midlands in USA, or places in South East France).

2.2 Determining geo-references that lie outside a region

After identifying members in an imprecise region, possibly with some noise, we obtain their coordinates by looking up their names in a geographic ontology. The ontology stores the coordinates for every geographic feature, so this gives a set of points with coordinates that are inside the region to be determined. We define these points to be red. For the red points, we compute a bounding box BB, which we enlarge by 20% in all directions to get the surroundings of the region of interest as well. Again using the ontology, we identify geographic features and their coordinates that lie in the bounding box BB, but were not found in step 1. These apparent non-members are likely to be outside the imprecise region because they did not appear in a trigger phrase. The coordinates of these locations give a set of points as well that we define to be blue. A reasonable boundary of the imprecise region is a polygon that contains (most of) the red points but not (most of) the blue points.

Most geographic features have an extent, and therefore cannot be represented well by a single point with two coordinates. They are better captured by polygons. However, our algorithms for step 3 assume that only points are given. This problem is remedied easily: we can choose all vertices of a polygon representing the feature. Or, for efficiency reasons, it will be better to choose a small set of points on the polygon. A simple choice is the set of four points where the polygon touches its bounding box.

2.3 Delineating the boundary of a region

We need to find a region (polygon) that has (nearly) all red points inside and (nearly) all blue points outside. We denote the set of red points by R and the set of blue points by B. The polygon that we want to define should have properties such as compact, simply-connected, smooth boundary, etc.

Algorithms to compute such polygons have been proposed before (Alani et al., 2001), where Voronoi diagrams are used. The idea is to compute the Voronoi diagram of $R \cup B$, the union of the two point sets.

The boundary between the red and blue cells defines the polygon. In the application of Alani et al., the input was assumed to be correct, that is, all colors were correctly assigned. We propose two algorithms for our application, where we cannot assume correct coloring of the points. False positives and false negatives are likely to occur, since the information is obtained from the Web.

2.3.1 The α-shape algorithm

The first algorithm starts with an α-shape of the red points (Edelsbrunner et al., 1983). Only the red component with the largest number of red points is maintained, the other red points are outliers (false positives) and are discarded. The remaining component is a simple polygon (Figure 2; red points are shown as discs and blue points are shown as squares). Then we adapt the polygon to transfer more blue points to the outside (if none are inside, we are done). We do this incrementally, while keeping the compact shape of the polygon. We choose a blue point close to the polygon boundary and change the shape. If no blue point lies close to the boundary, or the compact shape cannot be maintained, we stop and report the polygon. Blue points remaining inside are assumed to be false negatives.

There are several possibilities for which point to choose to bring outside, and also when to stop changing the shape of the polygon. Two natural choices of the first type are: (a) choose the blue point closest to the boundary of the polygon, and (b) choose the blue point that, when brought outside, gives the smallest additional perimeter length. Two natural choices of the second type are: (a) the additional perimeter length when bringing another point outside is large, and (b) the ratio of the squared perimeter to the area of the polygon exceeds a certain value. The latter choice is related to well-known shape measures for polygons like compactness and elongation (O'Sullivan and Unwin, 2003).

When a blue point p is brought to the outside, one edge of the polygon is chosen and replaced by new edges. The edge that is replaced is the one that is closest to the blue point, or the one that had the least increment in perimeter, whichever was the criterion for selecting p. Often, the new edges will be the two edges from the endpoints of the chosen edge to the point p. However, this could bring red points outside the polygon, which is not allowed. So instead, we do the following (Figure 3). Let u and v be the endpoints of the edge of the polygon to be replaced. Let w be the point on edge uv that is closest to p;

possibly, w is u or v. Now Δpuv is a triangle that is partitioned into two triangles Δpuw and Δpwv. If Δpuw does not contain any red points, then the new polygon will contain the edge pu. Otherwise, the new edges come from the shortest path from u to p that keeps all red points that are inside the triangle Δpuw in the polygon. This path necessarily is a convex chain and can be determined using a convex hull computation. The triangle Δpwv is handled the same way: either edge pv is new, or else the shortest path from p to v that keeps all red points in triangle Δpwv in the polygon.

2.3.2 The recoloring algorithm

The second algorithm to determine a reasonable boundary between the red and blue points is based on the Delaunay triangulation. We compute the Delaunay triangulation of R ∪ B, the red and blue points, and give all edges one of three colors. To describe the algorithm, an edge is called *blue* if both endpoints are blue, an edge is *red* if both endpoints are red, and an edge is *green* otherwise. If we connect the midpoints of the green edges around the biggest red component we get a possible shape for the polygon (Figure 4). This shape is very similar to, but not precisely the same as the shape obtained by (Alani et al., 2001). We will improve the polygon by changing the colors of points that seem to be falsely colored.

Note that a red point only has red and/or green incident edges, and a blue point only has blue and/or green incident edges. We define for each point p its green angle (Figure 5 shows the green angle for four of the points): it is the largest angle between two green edges incident to p that have no red or blue edge in between. We incrementally recolor any point whose green angle is larger than some well chosen value A, which must be larger than 180 degrees. Intuitively, a red point with green angle larger than 180 degrees is partially “surrounded” by blue points, and hence its color may have been wrong. A similar statement is true for a blue point with green angle larger than 180 degrees.

Recoloring a point (red to blue, or blue to red) changes the color of all the incident edges. For a red-to-blue recoloring, the red edges become green and the green edges become blue. For a blue-to-red recoloring, the blue edges become green and the green edges become red. Furthermore, the green angle of the neighbor points of a recolored point may change.

We continue this process until all points have green angle at most the pre-specified value A (Figure 6; only two points needed to be recolored). Then we take as the boundary of the imprecise region the connection of the midpoints of the green edges around the largest red component.

2.3.3 Potential adaptations to the algorithms

When we use trigger phrases to get points and their colors, the evidence that a point is inside or outside can be stronger or weaker. A name that appears very often in the trigger phrase gives a point that should not be recolored, but a name that appears only once or twice may well be falsely colored red. The methods described in this section do not take the strength of the evidence into account yet. However, both methods can be adapted for this. For example, if there is strong evidence that a point is inside the imprecise region, then the recoloring algorithm is not allowed to change its color from red to blue even if it is surrounded by blue points.

To delineate an imprecise region that is adjacent to the sea, or any large region in which no blue points are generated, we must take extra care to obtain good output. One general way to do this is to generate blue points randomly in regions that are void of red and blue points. The default is that if there is no evidence that some location is part of the imprecise region then it is not inside. For natural boundaries like coast-lines, additional methods are needed to respect them.

3. Evaluation

In this section we evaluate various aspects of our method using four regions: Wales, Midlands, South East, and East Anglia. Of these, Wales is not an imprecise region, but this in fact helps with the evaluation because we can therefore determine how much the region delineated corresponds to the true region. This is not possible for Midlands and South East. The fourth region, East Anglia, is also an imprecise region, but its extent is mostly defined nevertheless (http://en.wikipedia.org/wiki/East_Anglia).

We evaluate geo-parsing, geo-coding, and trigger phrases and snippets first for all four regions. Then we show delineated polygons resulting from both algorithms.

3.1 Evaluation of geo-parsing, geo-coding, and trigger phrases

3.1.1 Evaluation of geo-parsing

First we evaluate the success of the geo-parsing method. We did this through the manual analysis of titles and snippets for each region to identify *all* possible geo-references (not all are necessarily members of the target region). We use the GATE Graphical User Interface (GUI) to annotate the texts manually. A GATE evaluation tool called AnnotationDiff is used to compare two sets of annotations: a manually-generated set (key-set) and a system-annotated version (response-set). AnnotationDiff creates several measures of annotation overlap, including: correct (C), precision (P), recall (R), F_1-measure (F1), false positives (FP) and missing (M). The F_1-measure is a variant of the F_β-measure which gives equal weighting to precision and recall. The precision, recall and F_1-measure are computed from the number of annotations found to match correctly, the number of annotations missing from the key-set, and the number of false positives (Equation (1)). From the annotations in the response-set, which are automatically generated, precision measures the proportion of these matching the manually assigned annotations. From the annotations defined manually, recall measures the proportion of these which are also correctly identified by the geo-parser. The F1 score is a single-valued summary of both precision and recall and enables much simpler comparison of different geo-parsing methods. Precision, recall and F1 are commonly used measures in the evaluation of Information Retrieval systems.

$$P = \frac{C}{C + FP} \qquad R = \frac{C}{C + M} \qquad F1 = \frac{(\beta^2 + 1)PR}{(\beta^2 P) + R} = \frac{2PR}{P + R} \qquad (1)$$

AnnotationDiff is able to deal with partially correct responses (i.e., either partial text matches, or exact matches with slightly different byte offsets). For example, suppose we identify the location “Vale of Clwyd”. If the extraction system only identifies “Clwyd” within the same text span, this is classed as partially correct. This is useful in cases where a partial match would be useful (e.g., “Cardiff” rather than “Cardiff Castle”); although there are cases where partially correct annotations are not useful, e.g., “Letchmoor Bridge” and “Bridge”). Evaluation measures are computed based on whether partially correct responses are considered correct or not (*lenient* or *strict,* respectively).

Table 2 summarizes the results of geo-parsing using GATE where F1 Avg. is the average of the strict and lenient scores, and correct, partially correct and missing are given as a proportion of the total of these. Based on the highest number of correct (gazetteer lookup only with OS and SPIRIT data), 81% of all locations (across all regions) are identified using our geo-parsing method. The results vary across region where, for example, 88% of the locations are found in the results for Wales, 81% for the Midlands, 76% for the South East, and 59% for East Anglia. For this last region, the results are more greatly affected by a smaller number of total locations. Further points to make include the following. Firstly, having more fine-grained geographical resources improves the accuracy of markup. Geo-markup using both the SPIRIT ontology and OS gazetteer gives a 21% increase in the average F1 score and a 75% decrease in locations missed.

Secondly, using Information Extraction (IE) helps to reduce referent class ambiguity by ignoring geo-references that are used as people or organisations. In addition, using additional grammar rules to include organisations which begin with a location, we increase the number of correct and partially correct by 7% and reduce the number of missing locations by 30% (compared with using full IE without additional grammar rules). Thirdly, gazetteer lookup offers high recall (i.e., the highest number of correct/partially correct and fewest missing), but at the cost of more false positives.

Fourthly, false positives are typically due to limited context of snippets in which the extraction system can decide whether an entry in the gazetteer list is being used as location or not. For example, false positives in the markup for the Midlands include: Over, Lords, Gemini, Watch, Lee and West. Although these could be used within a geographical context (e.g., Lords the cricket ground, Lee a village in Lewisham and Gemini a water feature in Warrington), in the snippets they are not (e.g., “House of Lords”, “Lee is a 35, living in West Midlands” and “I’m a bubbly Gemini with a love of life”). Finally, locations that are missed are generally because of:

- improper use of capitalization (e.g., “rugby” rather than “Rugby”),
- misspellings (e.g., “Swanleigh” rather than “Swanley”),
- dictionary mismatch (e.g., “Stoke-on-Trent” vs. “Stoke-Upon-Trent”),

- no entry in the gazetteer lists (e.g., “Leamington Spa”), and
- formatting problems (e.g., multiple new-line characters inserted between locations).

3.1.2 Evaluation of region membership and geo-coding

In the previous evaluation we did not consider how many of the locations found were possible members of the target region or how many could be assigned spatial coordinates; in this evaluation we consider both. As before, we create a reference set of locations for each region which we consider being members of the target region and for which we assign the correct spatial coordinates (i.e., perform manual disambiguation in the case of referent ambiguity). Of course deciding whether locations are within an imprecise region relies on both an individual’s geographic knowledge and perception (e.g. most people will agree that points nearer to the centre of a region are members; however, points on the boundaries are less clear and much more subjective).

To establish place name membership, we used a variety of geographic resources including the gazetteers from these experiments and authoritative online sources: weather sites from the UK Met Office (http://www.metoffice.gov.uk/weather/europe/uk/uk.html) and the BBC (http://www.bbc.co.uk/weather/ukweather/), and the Wikipedia (http://en.wikipedia.org/wiki/) encyclopedia. These resources were used to provide a list of place names considered as region members in the following manner: (1) lists of region members were generated from the authoritative sources, (2) the administrative boundaries of these places were established from the gazetteers (i.e. the UK county), and (3) place names not listed in the authoritative sources (e.g. smaller towns and villages), but located within the administrative boundaries were included as region members. Alternative methods to using authoritative sources would be to question individuals and establish a boundary based on shared consensus (Montello et al., 2003). To verify the grounding of place names, we used the context provided by the geographical resources used in these experiments (e.g. the hierarchy of places at different administrative levels as provided by the SPIRIT ontology), together with online mapping tools such as Map24.com and Multimap (`http://www.multimap.com`) to visualize locations.

Table 3 summarizes the locations found using the best method in Table 2 (full IE using additional grammar rules). Many of the locations found occur multiple times; therefore to obtain a more accurate view of the grounding we count multiple occurrences once (unique). The second column in Table 3 shows the number of unique locations extracted using the geo-parser. Many of these locations, however, cannot be grounded using the SPIRIT or OS resources. There are many reasons for this, including:

- foreign names (e.g. Australia) which are found due to the default GATE gazetteer lists,
- locations such as "North West" found by the grammar rules of the semantic tagger,
- locations which are treated as "stopwords"[2] and removed before grounding (e.g., "Watch", "Links", "Castle", "Hall" and "Travel"), and
- locations found which do not match the gazetteer entry (e.g., "South Yorks" rather than "South Yorkshire").

The number of unique locations found is much smaller than the total number found (C+PC+FP) because many locations occur more than once (particularly in Wales and the Midlands).

The third column in Table 3 shows the number of unique locations grounded. For some regions, e.g., the South East, only a small proportion of unique locations found are actually grounded (37%), drastically reducing the number of potentially useful locations. The fourth column identifies the number of unique locations which are possibly correct, i.e., they are members of the region, although in the case of ambiguous locations they may be assigned wrong spatial coordinates. The fifth column shows the number of locations which are region members and have been grounded correctly (judged manually). The final column in Table 3 shows the number of ambiguous locations and the proportion of these disambiguated correctly. In some instances the simple default sense disambiguation method works well (e.g., for "Cambridge" in the East Anglia region); in other cases the default sense is not correct (i.e., the location is not the largest). Out of 7 ambiguous locations for Wales, only 14% are correctly disambiguated. This demonstrates the need for a better disambiguation method which takes into account the context (i.e., could distinguish between the same place name located in England and Wales).

[2] These are the top 250 most frequent words found within a 20,000 document test collection sampled from a 1TB Web collection which are either commonly used in general language or part of HTML markup.

Overall we find that 58% of the unique locations identified by the geo-parser are actual region members (average correct). Two reasons to explain this are: (1) the query is under-defined, and (2) the snippet contains irrelevant locations. We purposely use general search queries (e.g. “the Midlands” rather than “the Midlands of England”) to retrieve the largest number of results. However, this will also produce irrelevant search results. For example, “the Midlands” search results contain documents about locations in the Midlands of Ireland (as well as other countries). However, making the query more specific (e.g., using “the Midlands of England”, “British Midlands”, or adding “England” to the query) results not only in fewer results, but also many potentially useful results are not expressed in a more specific way. In part, this is because of colloquial language usage (i.e., people often just write “the Midlands” rather than the more explicit “the Midlands of England”).

The second problem is the scope of the target region in the snippet. For example, a snippet for the region “South East” (where <SNIPPET> demarcates the snippet text) is: “<SNIPPET> region. Gateshead is under Tyne and Wear, which is in the North Region. Colchester is under Essex, which is in South East Region. The </SNIPPET>”. The snippet contains both relevant (underlined) and irrelevant locations (e.g., “Gateshead” and “Tyne and Wear”). Therefore, to alleviate this problem, we tried a method whereby we extracted names from only the sentence containing the target region. In the previous example, we obtain “<SENTENCE> Colchester is under Essex, which is in South East Region </SENTENCE>.”

Table 4 shows the results of using locations found in the same sentence as the target region. Although the number of correct locations is lower than using the whole snippet, the number of unique and grounded locations are also much less (i.e., the number of irrelevant locations is reduced) causing the proportion of correct unique locations to rise from 58% to 70%. Sometimes, however, this technique is unsuccessful, e.g., “<SNIPPET> Carmarthenshire. Carmarthenshire (Welsh: Sir Gaerfyrddin) is a county in Wales. Its main towns are Carmarthen, Llanelli and Ammanford. </SNIPPET>.” In cases such as these, using language processing techniques such as co-reference resolution, would resolve “Wales” with “Its” in the second sentence and be included as part of the local context surrounding the target region.

Table 5 shows the top 20 locations (ranked by ascending order of frequency) extracted from the snippet sentences and titles for each region (using the full IE method with OS and SPIRIT gazetteer lists). The number of correct locations is typically 75% and above. Ignoring the term "England", frequently occurring locations are often good indicators that a candidate member belongs to a region. However, there are exceptions to this such as London in the Midlands which occurs four times, e.g., "<SNIPPET> short. Wolverhampton is a town in the midlands of England, and West Ham is a part of the East End (the east of London). Gwyn ap Nudd. </SNIPPET>." To reduce the effects of commonly occurring place names, we could re-rank the place names by the classic Robertson and Spärck Jones F4 formula which takes into account term frequency and the number of documents containing that term in a document collection (Robertson and Spärck Jones, 1976). The effect of this will be to reduce the impact of commonly occurring words and phrases.

3.1.3 Evaluation of trigger phrases and snippets

In this section, we analyze the snippets and trigger phrases used to generate candidate member regions. We manually identify all snippets that contain target region members. On average across all regions, we find 64% of snippets (and titles) that contain at least one target region member. Figure 7 shows a breakdown by region where total is the total number of documents resulting from searching all trigger phrases, and useful the number of results which contain at least 1 or more target region members. The number of documents returned varies dramatically with each region depending on how well the target region is represented in the Google index. The number of useful snippets is much lower, on average, than the total number of snippets retrieved, mainly because of queries picking up results from unrelated geographical areas, or not mentioning any additional location apart from the target region. The following examples illustrate these:

```
<TITLE>Wallace West Virginia - Finance Pages</TITLE>
<SNIPPET> Wales Wales is a principality west of England. Wales is a town in Walla County
Washington, USA Wallace Wallace is a city in Shoshone </SNIPPET>

<TITLE>The Quest for the Holy Ale: Welsh Ales</TITLE>
<SNIPPET> Your best chance of finding this, aside from beer festivals, is in the North-
East of Wales, also a good hunting ground for Plassey beers. </SNIPPET>
```

Based on these results, we can determine which of the lexical patterns are retrieving most correct locations which we show in Figure 8. This shows the total number of results returned and those containing at least 1 correct location (useful) summed over all regions. The trigger phrase categories are those given in Table 1, and Figure 8 shows that the class of patterns which, on average, return the most correct locations is *which_i*s (the pattern working best is actually "which is in" and gives the most useful snippets for each region). The pattern *is_a* also retrieves many useful locations (67%); although the pattern with the best accuracy is *is_direction* of which 86% of the results retrieved contain at least 1 correct location.

3.2 Evaluation of the algorithms determining the imprecise region

In Section 2.3 we presented two methods of generating a possible polygon for an imprecise region from a set of points colored red or blue, providing evidence that the point is inside or outside, respectively. Both algorithms were implemented, and we show the results for two of the data sets, namely, Wales and the Midlands. We did not use all locations from the ontology that were not in trigger phrases to create blue points. Especially smaller locations inside the region of interest may not be in a trigger phrase on the Web. To avoid these false negatives, we only chose bigger locations as candidates for the blue points.

Figures 9 and 10 show the outcome of the α-shape method for Wales and the Midlands, respectively. We tried four different values of α to obtain different initial shapes. It appears that this has a large influence on the outcome of the imprecise region. As the stopping criterion we chose to continue bringing blue points to the outside as long as the perimeter of the resulting polygon is no more than five times its diameter. As mentioned before, other possibilities exist as well.

It appears that the α-shape indeed eliminates red outliers, assuming that a suitable value of α is chosen. Visual inspection shows that a value of 600 (Wales) or 700 (Midlands) is best for the two test cases. The process of bringing blue points to the outside by changing the polygon also works well, assuming that these blue points are really points that are outside the imprecise region. Our algorithm can handle incorrectly colored blue points in the middle of the delineated polygon, but incorrectly colored blue points that are close to the α-shape can lead to adapting the polygon when this is not appropriate. Similarly, incorrectly colored red points close to the correctly colored red points give problems, which can be seen in the figures. Due to the rapid growth of information on the Web, the number of false negatives may decrease, and this problem may be solved automatically.

Figures 11 and 12 show the results of the recoloring approach. In the top left of both figures, the delineated polygon is shown if no points are recolored; this corresponds to choosing the angle $\alpha>360$ degrees. It is clear that recoloring helps to generate more reasonable polygons for both Wales and the Midlands. As expected, values not much larger than 180 give a better shape of the polygon that is delineated. However, the results are not satisfactory overall. This is partly due to the large number of false positives, red points that lie close to the region of interest, but not inside. This makes the polygons for Wales and the Midlands to be too large (except at the east part of the Midlands, where it is too small because red points are missing completely). Some problems occur because there are no blue points in the sea, causing ill-shaped triangles in the triangulation and affecting the recoloring. There are several possible ways in which the shortcomings can be remedied. For example, we can give preference to red-to-blue recolorings because false positives (red) appear more problematic than false negatives (blue). Secondly, we can use different angles for recoloring for the red and blue points. Thirdly, we can extend the definition of green angle to take a larger neighborhood into consideration, which allows the method to deal with small groups of outliers. Finally, the rapid growth of the Web may also help to improve the shape of the polygons that are delineated. False positives will then appear to be most problematic.

Figures 13 and 14 show the regions East Anglia and South East with the best settings of the α-shape method (left) and the recoloring method (right). The outlier for South East and the recoloring method would have been recolored if there were some extra blue points South of the mainland of Great-Britain.

4. Discussion and future work

It appears that our approach to provide candidate members for a target region is successful. Our approach is to generate several searches based on lexical patterns and extract geo-references from the metadata returned by searching the Web using Google. We have shown that our method of geo-parsing is accurate for a number of different target regions and actual region members can be found using this approach. It appears that our assumption that region members will appear within the same local context of the target region is correct and useful to extract useful geo-references. We would like to explore this approach further, in particular we would like to use relevance feedback to perform multiple search iterations using locations either identified manually by a user, or using a pseudo relevance feedback approach (e.g. the most frequently occurring places from an initial search). We would like to experiment with different ranking approaches for predicting reliable region members. We would also like to experiment with extracting locations from the web pages themselves and compare this with using the Google metadata only. Also, Google provides a link to "similar" pages which we may be able to exploit in order to find more useful locations. Finally, we noticed that results returned by Google were biased, e.g. many results for the Midlands were from a dating agency. We would like to experiment with trying to pick up either more varied pages, or Web pages which may provide a better and more reliable source of geo-references, e.g. directory lists, encyclopedias, or "about/contact us" pages. These may provide more reliable snippets with more geo-references.

Our implementation of the algorithms to determine the boundary of imprecise regions by finding a polygon that includes many red points but few blue points show promising results. The methods can deal with falsely colored red and blue points, but the quality of the output will still be influenced negatively if there are many falsely colored points. At the moment the parameters have to be tuned by hand to get good polygons. Experiments on more data sets and on variations of the methods are needed to obtain more insight and better results. For example, experiments can reveal which blue point selection rule and which stopping criterion gives the best results in general. Also, more research and experiments are needed to refine the polygon delineation method. The strength of evidence of a point being red or blue can be taken into account, for example. At the moment it appears that the α-shape method gives better polygons than the recoloring method, but it is preliminary to see the experiments in this paper as conclusive evidence.

5. Acknowledgements

This research is supported by the EU-IST Project No. IST-2001-35047 (SPIRIT). Iris Reinbacher was also partially supported by a travel grant of the Netherlands Organization for Scientific Research (NWO). The authors thank Alexander Wolff, Marc Benkert, and Markus Völker for implementation support and related research that affected this paper as well. Furthermore, several useful suggestions on this paper were given by Ross Purves and an anonymous reviewer.

6. References

Alani H., Jones, C.B., and Tudhope, D.S. (2001). "Voronoi-based region approximation for geographical information retrieval with gazetteers". International Journal of Geographical Information Science, 15(4), 287-306.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). "Web question answering: is more always better?" In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 291-298, Tampere, Finland: ACM.

Edelsbrunner, H., Kirkpatrick, D.G., and Seidel, R. (1983). "On the shape of a set of points in the plane". IEEE Transactions on Information Theory, IT-29(4):551-559.

Joho, H., and Sanderson, M. (2000). "Retrieving Descriptive Phrases from Large Amounts of Free Text". In: Proceedings of the 9th International Conference on Information and Knowledge Management, 180-186, McLean, VA: ACM.

Joho, H., Liu, Y.K., and Sanderson, M. (2001). "Large scale testing of a descriptive phrase finder". In: Allen, J. (Ed.), Proceedings of the 1st Human Language Technology Conference, 219-221, San Diego, CA: Morgan Kaufmann.

Jones, C.B., Abdelmoty, A.I., and Fu, G. (2003). “Maintaining ontologies for geographical information retrieval on the web”. In Meersman, R., Tari, Z., Schmidt, D. C. (Eds.) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE Ontologies, Databases and Applications of Semantics, ODBASE'03, Catania, Italy, Lecture Notes in Computer Science 2888, 934-951.

Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M.J., and Weibel, R. (2002). “Spatial information retrieval and geographical ontologies - an overview of the spirit project. In Proc. 25th Annu. Int. Conf. on Research and Development in Information Retrieval (SIGIR 2002), 387-388.

Larson, R.R. (1996). Geographic Information Retrieval and Spatial Browsing. In GIS and Libraries: Patrons, Maps and Spatial Information, Linda Smith and Myke Gluck, Eds., University of Illinois.

Li, H., et al. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Kornai, A. and Sundheim, B. (eds.) Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada: ACL, 39-44.

MacEachren, A. M. (1995). How maps work. Publisher: The Guilford Press, New York.

Markowetz, A., Brinkhoff, T., and Seeger, B. (2003). "Exploiting the Internet as a Geospatial Database". ISPRS WG IV/5 Workshop on Next Generation Geospatial Information.

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

Moon, R. (1998). “Fixed expression and idioms in English”. Clarendon Press, Oxford.

O’Sullivan, D., and Unwin, D.J. (2003). "Geographic Information Analysis". Wiley, Hoboken.

Rauch, E., et al. (2003). A confidence-based framework for disambiguating geographic terms. In: Kornai, A. and Sundheim, B. (eds.) Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada: ACL, 50-54.

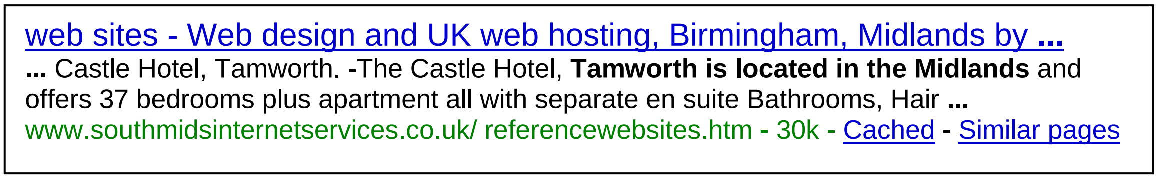
Robertson S E and Spärck-Jones K. (1976). Relevance Weighting of Search Terms. Journal of the American Society For Information Science, 129-146

Smith, D. A., and Mann, G. S. (2003). Bootstrapping toponym classifiers. In: Kornai, A. and Sundheim, B. (eds.) Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada: ACL, 45-49.

Smith, D. (2002). "Detecting and Browsing Events in Unstructured text". In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 73-80, Tampere, Finland: ACM.

<table>
<tr><th>ID</th><th>Trigger phrase</th><th>Examples</th></tr>
<tr><td>in</td><td>* in [R]</td><td>Birmingham in the Midlands</td></tr>
<tr><td>which_is</td><td>which is [in | in the * of] [R]</td><td>West Ham which is in London</td></tr>
<tr><td>is_a</td><td>* is a [city | county | province | region | state | town | village] in [R]</td><td>Paris is a city in France</td></tr>
<tr><td>is_direction</td><td>* is [in | located in | situated in] the [center | centre | north | south | east | west | north east | south east | north west | south west] of [R]</td><td>Canterbury is located in the south east of England</td></tr>
<tr><td>such_as</td><td>[cities | towns | villages | counties | provinces | regions | states] in [R] [such as | including] *</td><td>Cites in the Midlands such as Birmingham</td></tr>
<tr><td>and_other</td><td>* and other [cities | towns | villages | counties | provinces | regions | states] in [R]</td><td>Staffordshire and other counties in the Midlands</td></tr>
</table>

Table 1: Trigger phrases used to identify geo-references



web sites - Web design and UK web hosting, Birmingham, Midlands by **...**
... Castle Hotel, Tamworth. -The Castle Hotel, **Tamworth is located in the Midlands** and offers 37 bedrooms plus apartment all with separate en suite Bathrooms, Hair **...**
www.southmidsinternetservices.co.uk/ referencewebsites.htm - 30k - Cached - Similar pages

Figure 1: Example Google search result for "* is located in the Midlands"

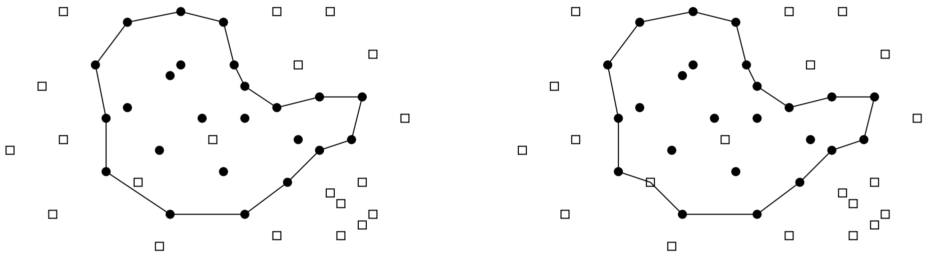


Figure 2: α-shape of a set of red points (circles) and its adaptation so that a blue point (square) is no longer inside

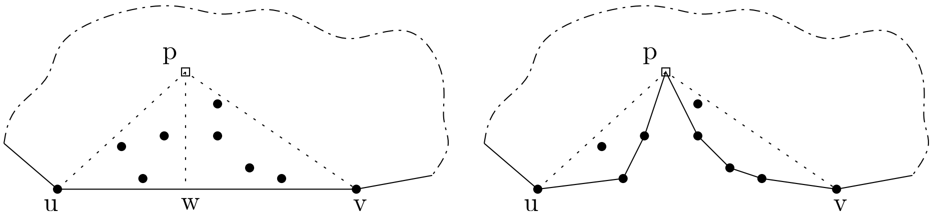


Figure 3: Construction illustrating how a polygon is adapted so that the blue point p is no longer inside

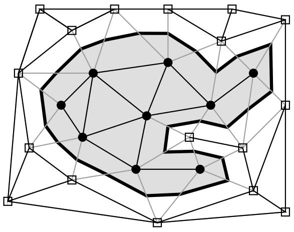


Figure 4: Delaunay triangulation of a set of red and blue points, and a polygon that separates them by connecting midpoints of Delaunay edges

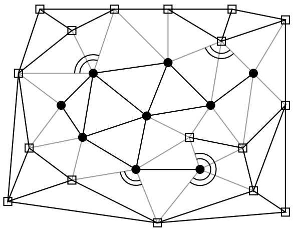


Figure 5: Illustration of the green angle of four of the points

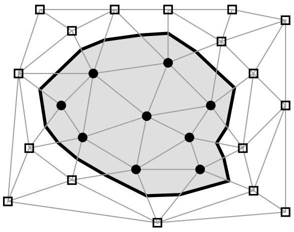


Figure 6: The polygon obtained by two recolorings of the points in Figure 5

Region	C (%)	PC (%)	M (%)	FP	F1 Strict	F1 Lenient	F1 Avg.
Gazetteer lookup only (SPIRIT)							
Wales	61	3	36	12	0.7289	0.7651	0.7470
Midlands	43	7	50	6	0.5673	0.6590	0.6132
South East	55	11	34	8	0.6565	0.7856	0.7211
East Anglia	38	3	59	0	0.5417	0.5833	0.5625
Total	*54*	*7*	*39*	*26*	0.6232	0.6983	0.6610
Gazetteer lookup only (SPIRIT and OS)							
Wales	88	5	7	82	0.8249	0.8719	0.8484
Midlands	81	7	12	33	0.7965	0.8701	0.8333
South East	76	11	13	40	0.7588	0.8682	0.8135
East Anglia	59	38	3	7	0.5405	0.8919	0.7162
Total	*81*	*9*	*10*	*162*	0.7302	0.8755	0.8029
Full Information Extraction (SPIRIT and OS)							
Wales	84	2	14	46	0.8499	0.8702	0.8601
Midlands	75	6	19	19	0.7907	0.8512	0.8209
South East	68	9	23	11	0.7496	0.8526	0.8011
East Anglia	41	6	53	1	0.5490	0.6275	0.5882
Total	*75*	*5*	*20*	*77*	0.7348	0.8004	0.7676
Full Information Extraction (SPIRIT and OS) using additional grammar rules							
Wales	87	3	10	54	0.8550	0.8798	0.8674
Midlands	80	7	13	24	0.8115	0.8825	0.8470
South East	74	9	17	12	0.7807	0.8808	0.8307
East Anglia	47	38	15	2	0.4923	0.8923	0.6923
Total	*79*	*7*	*14*	*92*	0.7349	0.8839	0.8096
Avg Total	*72%*	*7%*	*21%*	*89*	*0.7058*	*0.8145*	***0.7603***

Table 2: Evaluation results for geo-parsing where C = Correct; PC = Partially Correct; M = Missing; FP = False Positives; F1 Strict = F1 computed using correct; F1 Lenient = F1 computed using correct and partially correct; F1 Avg = average of F1 Strict and F1 Lenient

Region	**Unique (total)**	**Grounded Unique**	**Possibly correct**	**Correct (%grounded)**	**Ambiguous (% correct)**
Wales	120 (409)	74	43	37 (50%)	7 (14%)
Midlands	77 (223)	57	28	27 (47%)	3 (66%)
South East	141 (267)	52	37	34 (65%)	10 (70%)
East Anglia	19 (31)	14	10	10 (71%)	3 (100%)
Avg	*89 (233)*	*49*	*30*	*27 (58%)*	*6 (63%)*

Table 3: Number of locations identified which are region members and ambiguous (using full IE)

Region	**Grounded unique**	**Possibly correct**	**Correct (%grounded)**	**Ambiguous (% correct)**
Wales	53	40	35 (66%)	6 (17%)
Midlands	48	27	26 (54%)	3 (66%)
South East	38	31	29 (76%)	8 (75%)
East Anglia	12	10	10 (83%)	3 (100%)
Avg	*38*	*27*	*25 (70%)*	*5 (65%)*

Table 4: Locations extracted from the local context of the target region (the sentence)

Wales		Midlands		South East		East Anglia	
Location		Location		Location		Location	
Swansea	9	Ireland	9	England	32	Suffolk	3
Carmarthenshire	9	Birmingham	9	Brighton	10	England	2
Cardiff	7	Derbyshire	5	London	7	Cambridge	2
England	6	London	4	Essex	7	Cambridgeshire	2
Gwynedd	6	England	4	Oxfordshire	6	Sutton Bridge	1
Ceredigion	5	Coventry	4	Dorset	3	Sible Hedingham	1
Powys	5	Leicester	3	Poole	3	Lowestoft	1
Conwy Castle	4	Nottinghamshire	3	Dorchester	3	Newmarket	1
Pembrokeshire	3	Nottingham	3	Weymouth	3	Barnwell	1
Chester	3	Hinckley	3	South Woodham Ferrers	2	Thetford	1
Caernarfon	2	Leicestershire	2	South Oxfordshire	2	Thurston	1
Presteigne	2	Watch	2	Northamptonshire	1	Norfolk	1
Llandrindod	2	Lutterworth	2	Spelthorne	1		
Wrexham	2	Amington	2	Scotland	1		
Harlech	2	Moseley	2	Borough	1		
Monmouthshire	2	Stafford shire	2	Clough	1		
Carmarthen	2	Catherine	1	Hastings	1		
Kingdom	2	Shire	1	Gretna Green	1		
Corndon Hill	1	Warwick shire	1	West Lothian	1		

Table 5: Top 20 locations extracted from the Google snippets (sentence only) and titles ranked by ascending order of frequency

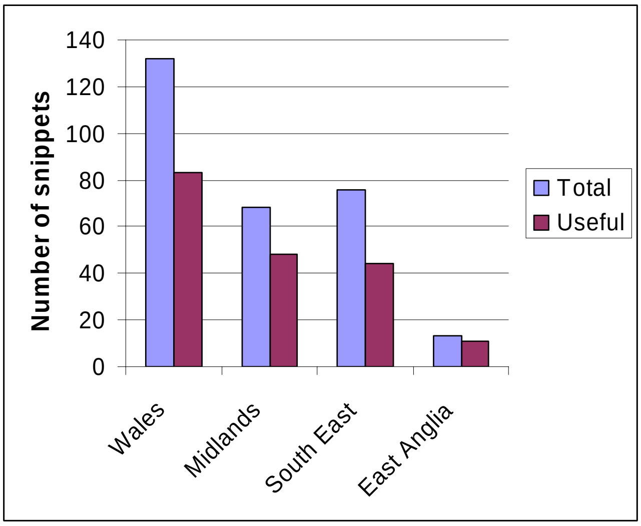


Figure 7: Number of snippets and those containing at least one correct location (useful) by region

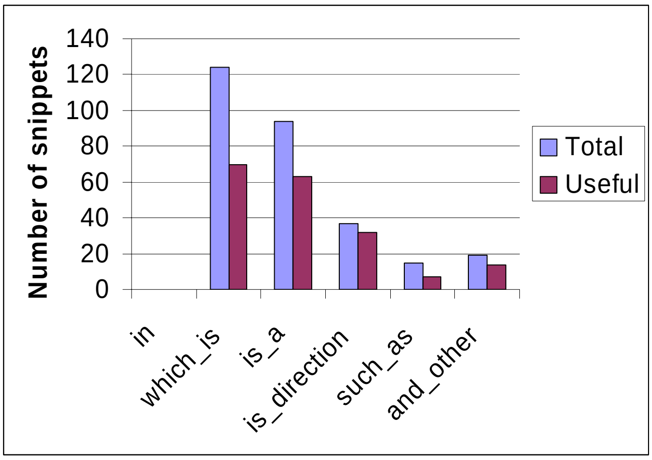


Figure 8: Number of snippets and those containing at least one correct location (useful) by trigger phrase

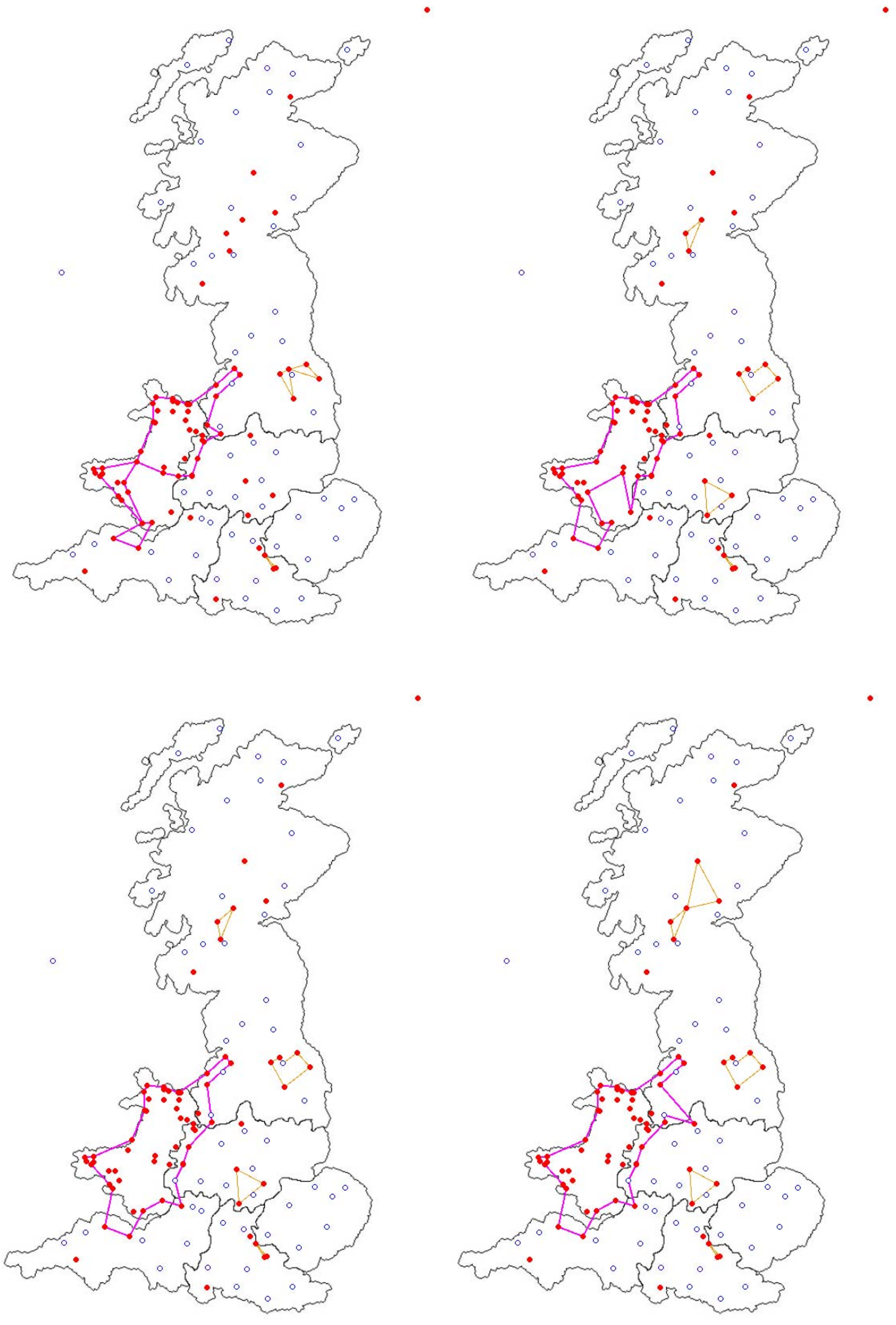


Figure 9: Different values of α affect the boundary obtained by the α-shape algorithm considerably. Shown are the outcome after choosing α=315, α=400, α=600, α=700 for Wales

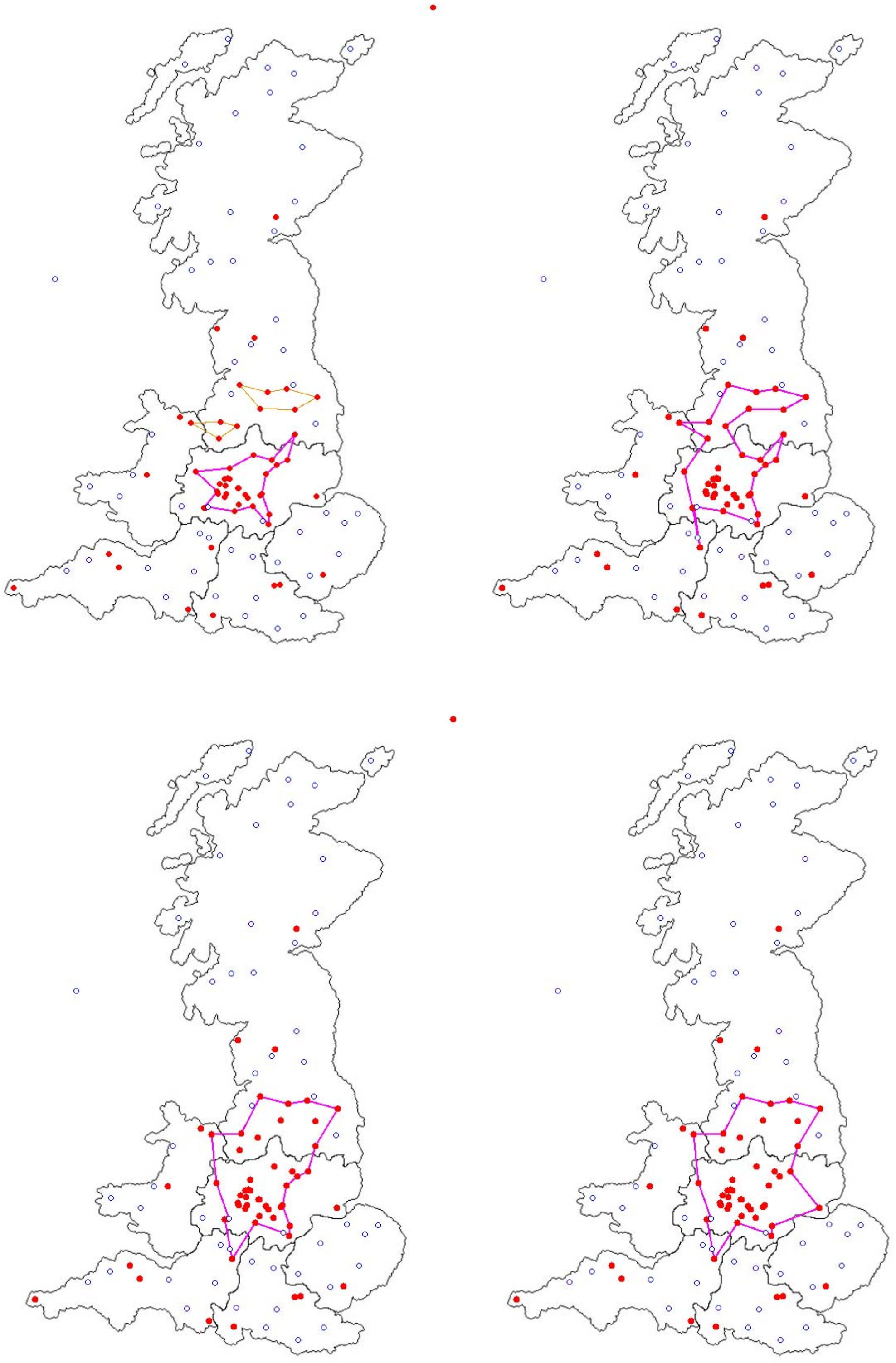


Figure 10: Different values of α affect the boundary obtained by the α-shape algorithm considerably. Shown are the outcome after choosing α=315, α=400, α=600, α=700 for the Midlands

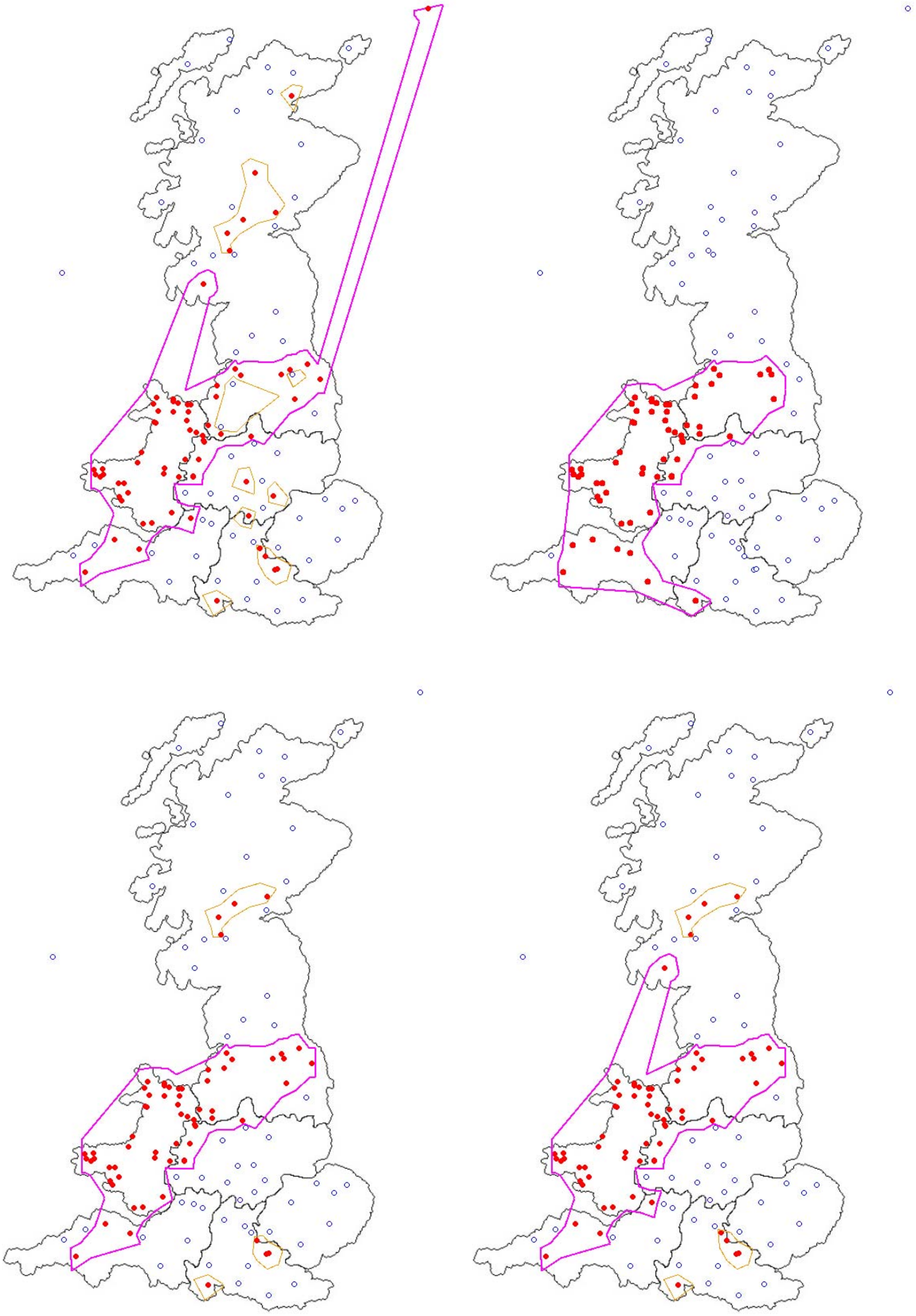


Figure 11: Delineated polygon for Wales before recoloring, and the outcome of the recoloring algorithm with angles 185, 215, and 260

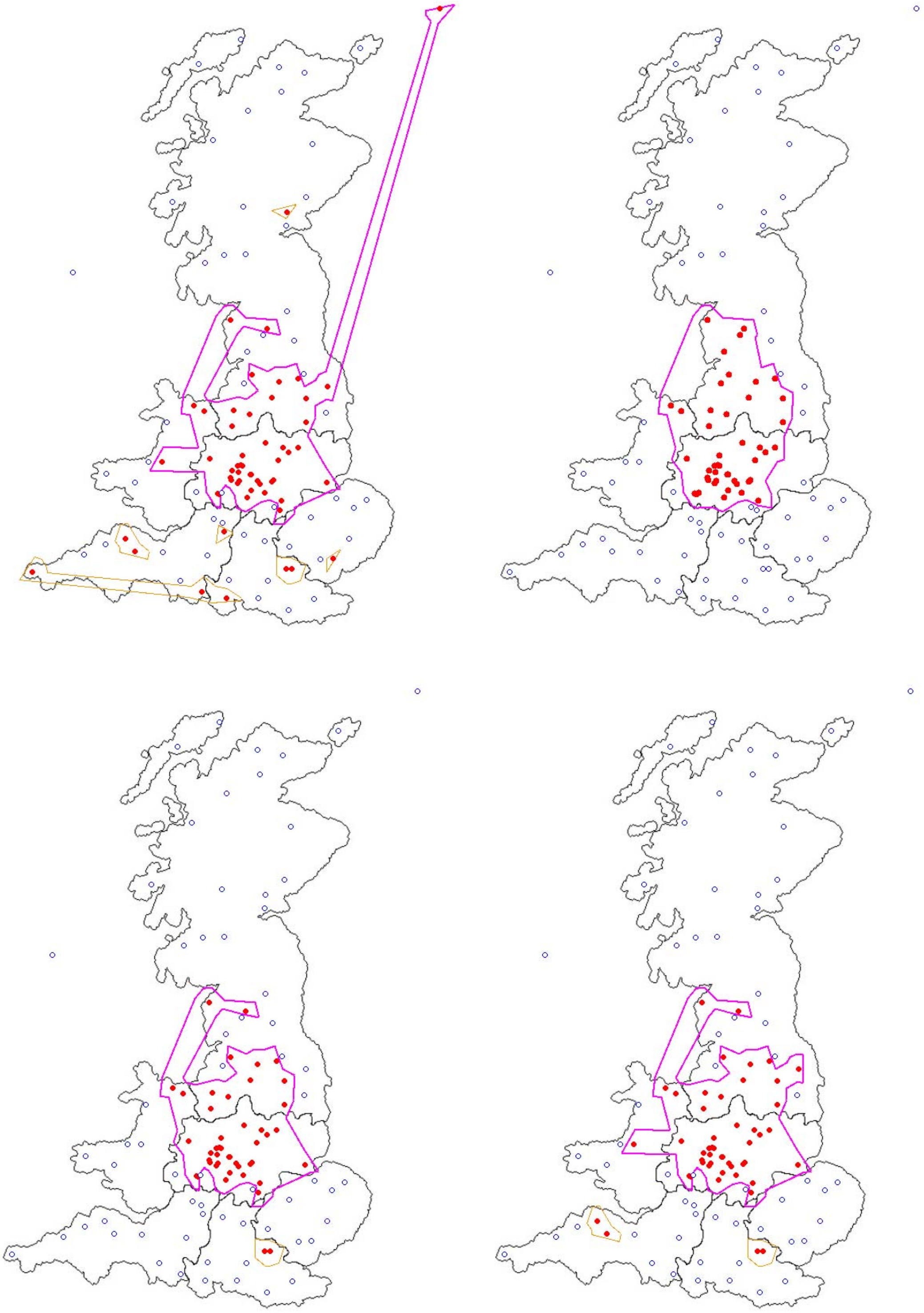


Figure 12: Delineated polygon for the Midlands before recoloring, and the outcome of the recoloring algorithm for the Midlands with angles 185, 215, and 260

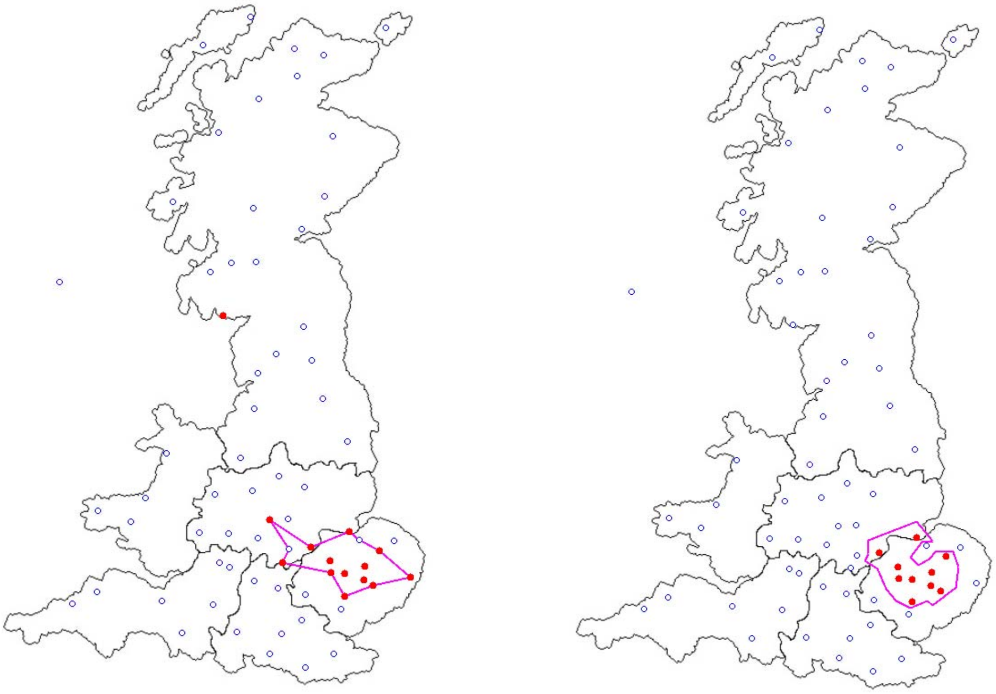


Figure 13: Delineated polygon for East Anglia with the α-shape algorithm (α=600) and the recoloring method (angle 215)

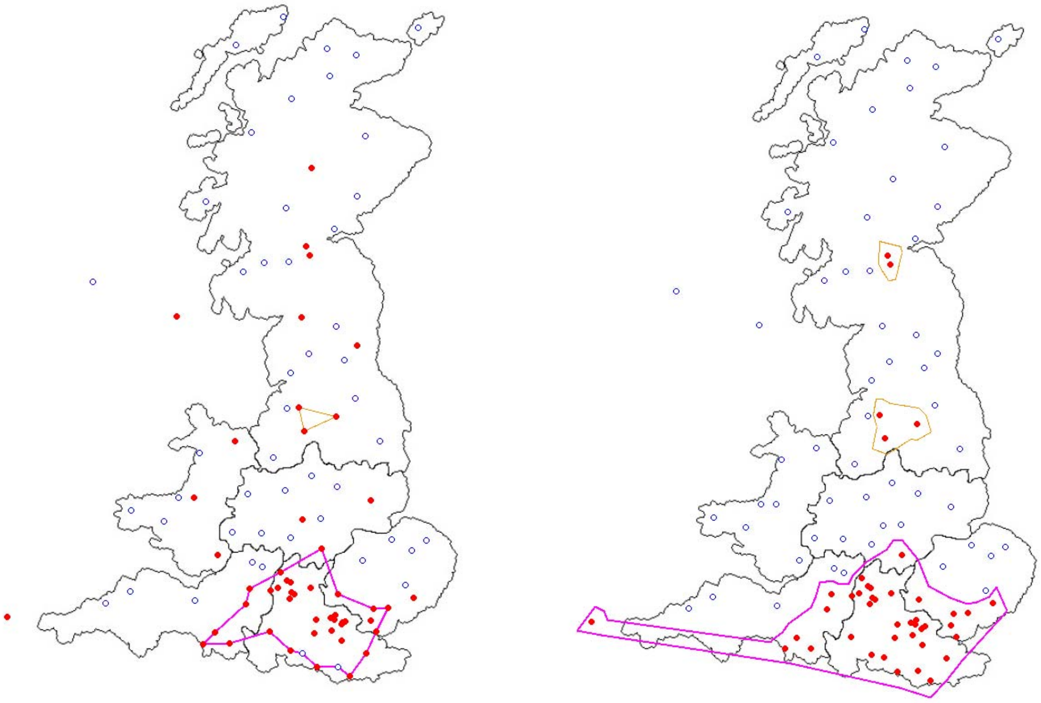


Figure 14: Delineated polygon for South East with the adaptation method (α=600) and the recoloring method (angle 215)