GEO 511 MASTERS THESIS

University of Zurich
Department of Geography, Geographic Information Systems
Winterthurerstrasse 190
8057 Zurich

Evaluation of spatial Relevance in Geographic Information Retrieval

Author

Raphael Kunz
Kapfstrasse 12
8032 Zürich
rkunz@geo.uzh.ch

Supervisor

Dr. Ross. S. Purves
ross.purves@geo.uzh.ch

Faculty Representative

Prof. Dr. Robert Weibel
robert.weibel@geo.uzh.ch

Submitted: November 30, 2008

Abstract

Information retrieval (IR) is concerned with the finding and ranking of documents with respect to their relevance to a corresponding query, by searching in relatively unstructured data collections. Geographic Information Retrieval (GIR) includes a spatial component to classic IR. The queries are composed of a thematic and a spatial concept and the documents are ranked with respect to their thematic as well as to their spatial relevance to the corresponding query. In order to evaluate IR and GIR results, documents need to be assigned a relevance value for the respective query. This process is known as relevance judgement.

The evaluation of GIR results is typically identical to the evaluation of classic IR results, which is based on precision and recall measurements. The spatial properties of GIR results have so far not been respected for evaluation purposes. In this work, a new methodology is proposed, which combines the binary document relevance judgement with the result ranking evaluation and additionally includes measures for the description of the result ranking's spatial component.

The spatial measures include the comparison between a preliminary defined expected extent and the footprint distribution's resulting extent as well as Kernel Density Estimation in order to describe the footprint distribution and footprint clustering within the resulting extent. The definition of expected extents is examined in detail for different spatial operators, different geographical granularities, and different thematic concepts. The results of the new methodology combine precision and recall with these spatial measures and further include the corresponding ranking positions.

The new methodology is applied to a set of SPIRIT as well as to a set of GeoCLEF data in order to test the applicability of the proposed methods and to investigate the "spatialness" of the different systems. The main problem for spatial evaluation is that many of the results include a low number of unique footprints and therefore are uninteresting for most spatial measures. The combination of query, gazetteer and document collection is found to be insufficiently spatially considerate in many cases, which creates a geographical granularity level in which classic IR often outperforms GIR. Furthermore, toponym distributions in documents are strongly biased towards the region in which the authorship is located. One possibility to improve the performance of GIR is the diversification of the document collection, by including documents located in the query region in order to achieve a finer geographical granularity level.

Zusammenfassung

Information Retrieval (IR) hat zum Ziel, Dokumente in unstrukturierten Dokumentkollektionen zu finden und entsprechend ihrer Relevanz zu einer Abfrage zu sortieren. Geographic Information Retrieval (GIR) berücksichtigt bei der Suche nach Dokumenten geographische Suchbedingungen, erkennt geographische Namen in Abfragen und ordnet gefundene Dokumente entsprechend ihrer thematischen, als auch ihrer räumlichen Relevanz. Für die Evaluation von IR und GIR Resultaten ist es notwendig zu wissen, welche Dokumente für eine Anfrage relevant sind und welche nicht. Die Zuweisung von Relevanzwerten an Dokumente ist bekannt als Relevance Judgement.

Die Evaluation von GIR Resultaten beruht typischerweise auf denselben Massen wie diejenige von IR Resultaten, Precision und Recall. Die räumliche Komponente der Resultate wurde in bisherigen Evaluationen nicht explizit berücksichtig. In dieser Masterarbeit wird eine neue Methodologie entworfen, die das binäre Relevance Judgement und die Resultatevaluation kombiniert und dabei Masse für die Beschreibung der räumlichen Komponente des Resultates mit einschliesst.

Die räumlichen Masse sind aufgeteilt in Flächenvergleiche zwischen einem vorgängig definierten erwarteten Extent und dem tatsächlich bedeckten Extent der Footprintverteilung einerseits, sowie Kernel Density Estimation zur Beschreibung der Footprintverteilung andererseits. Die Definition des erwarteten Extents spielt dabei eine gewichtige Rolle und ist für verschiedene räumliche Operatoren, Granularitäten und thematische Konzepte detailliert aufgezeigt. Die Resultate dieser neuen Methodologie vereinen Precision und Recall mit diesen räumlichen Massen und berücksichtigen ausserdem auch die Rankingpositionen.

Die räumlichen Masse werden auf je ein Set von SPIRIT und von GeoCLEF Daten angewendet, um die Verwendbarkeit der Methodik zu überprüfen und gleichzeitig Aussagen über die Berücksichtigung der Geographie in den verschiedenen Systemen zu ermöglichen. Die grösste Schwierigkeit besteht dabei darin, dass viele Ergebnisse nur wenige nicht-duplizierte Footprints aufweisen und deshalb für eine räumliche Auswertung uninteressant sind. Die Kombination aus Abfrage, Gazetteer und Dokumentkollektion schafft oftmals keine feine geographische Granularität, was dem IR bessere Resultate ermöglicht als dem GIR. Weiter ist die Toponymverteilung in Texten stark durch den Standort der Autorschaft beeinflusst. Eine Möglichkeit um geographische Suche stärker zu gewichten, ist demnach die Erweiterung der Dokumentkollektion um Texte, die aus der abgefragten Region stammen.

Acknowledgments

I would like to thank my supervisor, Ross Purves, for his much appreciated advice and contributions; Simon Overell for providing us with the data and necessary support; and, finally, Nadia Fries, for proof-reading my paper.

Contents

Figures

Tables

1. Introduction

1.1. Context

The use of search engines for the retrieval of web documents has become a matter of course among the user community of the World Wide Web. The underlying research field is called Information Retrieval (IR), which is characterised by the content based querying of unstructured collections of documents and the probabilistic relevance ranking depending on a document's responsiveness to a query (Van Rijsbergen, 1979).

Much of the information available on the World Wide Web contains references to geographic locations. McCurley (2001: 222) states that around 10% of the web pages gathered in a "*fairly large partial web crawl*" contained a reference to either a US Zip code or telephone number. Martins & Silva (2004) found an average of 2.2 references to the Portuguese municipalities per document in a corpus of 3.7 Million web pages.

Since "*spatial information obtained from web documents is often incomplete and fuzzy in nature*" (Silva et al., 2006: 379), the field of Geographic Information Retrieval (GIR) has grown out of classic IR. GIR includes the detection of place names (toponyms) and the assignment of their corresponding geographic coordinates (footprints) to documents in unstructured collections as well as the processing of incomplete spatial queries, the initial retrieval and the probabilistic relevance ranking of retrieved documents.

The performance of GIR systems is typically evaluated with classic IR measures, expressing how many retrieved documents are relevant and how many off all relevant documents are retrieved. For the evaluation, documents need to be assigned a relevance value for the respective query. This process is known as relevance judgement or relevance assessment.

Current relevance judgement approaches distinguish between thematic and spatial relevance, use binary or ternary relevance schemes and focus on single documents. For the evaluation of GIR systems, the same measures are used as for classic IR evaluation; the geographic or spatial properties of GIR results have so far not been respected for evaluation purposes. Methods for quantifying point patterns, point distributions and clustering behaviour have been described and applied in many GIScience related publications (Openshaw et al., 1987, Diggle, 1990), but so far no effort was made in order to use such measures in GIR relevance judgement and result evaluation.

1.2. Hypotheses

When looking at how GIR documents and result rankings are currently judged and evaluated, one notices that the spatial component is not much respected. The following hypothesis, as one which is to be assumed as underlying current spatial relevance judgement and result evaluation, is to be investigated in this masters thesis:

H_1: All relevant characteristics of a GIR ranking result can be revealed using precision and recall measurement and spatial considerations do not benefit the process of GIR result evaluation.

The approach taken in this work in order to test this hypothesis is to find out what spatial measures are interesting for evaluation purposes and to propose a new methodology which includes a spatial component to standard evaluation.

The uncritical consideration of the query's spatial component's extent and geographical granularity in relation to those of the document collection leads to the following assumption encountered in GIR research: A spatially aware system should achieve better results than a pure text search, if a query has a thematic and a spatial component. Based on this assumption, the second hypothesis to be investigated is:

H_2: As soon as a query is built following the concept@location form, geographical aware information retrieval outperforms text based retrieval and independently allows for spatial relevance judgement of documents and evaluation of ranking results.

The first step to approaching this hypothesis is to look at GeoCLEF results and find out about the spatial quality of the queries, of the document collection and of the footprint distributions. This might explain some of the performance differences between text based and geographically aware systems from a geographical point of view. The second part of this hypothesis is based on the fact that relevance judgement and result evaluation are static. In other words, relevance judgement and evaluation are assumed to be independent of the query region, the query region's granularity, the spatial operator, the document collection's granularity and its toponym distribution.

A more concrete compilation of this work's aims based on these hypotheses and including conclusions of the chapter Current State of Research can be found in 2.3.

1.3. Outline

The chapter Current State of Research introduces the field of GIR with a special focus on its spatial component. The chapter's second part focuses on point pattern analysis techniques and their application in GISciences. This work's methodology and practical part is separated into two different sections. In the first one, Evaluating SPIRIT data, a number of pattern and extent analyses are applied and tested in GIR footprint distribution evaluation and a suitable methodology for spatial relevance judgement and evaluation is outlined. In the second methodology section, Evaluating GeoCLEF Data, the conclusions of the first methodology chapter are applied in combination with standard IR measures and to a set of GeoCLEF results. This work's main idea for spatial relevance evaluation is also introduced and explained in that chapter. The chapter Results and Interpretation presents results of the second methodology part only, concerning the GeoCLEF data basis and the usefulness of the new relevance judgement and evaluation method. The Discussion consistently focuses on the ability of the second part's methodology to answer to spatial problems of search topics, and on the appropriateness of the GeoCLEF setup. Findings are summarised in the Conclusions chapter.

2. Current State of Research

2.1. Geographic Information Retrieval

2.1.1. Introduction

2.1.1.1. What is GIR?

The article "As We May Think" (Bush, 1945), which often constitutes an entry point into the field of IR, is the first to reflect upon the idea of "*automatic access to large amounts of stored knowledge*" (Singhal, 2001: 1). The idea of using words as units for indexing and ranking documents is introduced by Luhn (1957). According to Lancaster (1968), an information retrieval system does not answer to users' questions but informs him/her about the "*existence and whereabouts of documents*" relative to the required topic. Van Rijsbergen (1979) characterises IR, compared to Data Retrieval (DR), as a probabilistic system, attempting to find the best matching documents in terms of their relevance to a request.

Geographic Information Retrieval (GIR) is a relatively recent field, resulting from IR, while adding a spatial dimension to it. A first definition of GIR is proposed by Larson (1996: 82):

"*Geographic information retrieval (GIR) is concerned with providing access to geo-referenced information source*"

Arguing that this definition is too broad, Purves & Jones (2006a: 375) give an alternative definition:

"*...we define GIR as the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope*".

According to the differentiation of information retrieval and data retrieval, Purves & Jones (2006a) underline the distinction between geographic information retrieval and geographic data retrieval and emphasize the relatively unstructured nature of the document collections. According to their definition, the authors list a number of key issues to be addressed in GIR that can be grouped as follows:

- Identification of geographic terms and assignment of geographic coordinates
- Thematic and spatial indexing and suitable retrieval algorithms
- Combined thematic and spatial relevance ranking
- Usability (Query formulation, user interface, visualisations)

The result evaluation occurs at the end of the GIR processing chain. It tries to validate and evaluate the quality of the ranking produced by the previous steps. It can be considered as standing outside of the above scheme, but has an influence on these processes through feedback. The evaluation of GIR results requires that one know which and how many documents are relevant to a respective query. The assignment of relevance values to documents in a document collection is called relevance judgement or relevance assessment.

Who does GIR and where is it done? An important institution is the Association for Computing Machinery (ACM[1]), where many GIR relevant articles are published. Furthermore, the ACM hosts the Conference on Information and Knowledge Management (CIKM[2]) and has a Special Interest Group on Information Retrieval (SIGIR[3]). Important platforms are the GIR workshops, which take place alternately at the CIKM and the SIGIR conference (Jones & Purves 2005, Purves & Jones 2004, 2006b, 2007) and the GeoCLEF[4], as introduced in more detail in chapter 2.1.4. Participants are academic research groups from the U.S.A, Europe and Asia as well as commercial suppliers like Microsoft, IBM or MetaCarta[5].

The following chapters introduce the processes of identification of geographic names, assignment of geographic coordinates and basic IR evaluation methods. Relevance ranking and evaluation methods with an emphasis on the spatial component are presented in chapter 2.1.2.

2.1.1.2. Toponym Recognition

Toponym recognition is the process by which geographic terms are found in unstructured text. This process is also known as Named Entity Recognition (NER) and encounters three major problems:

- Toponyms describing other things than geographic entities
- Toponyms used in a representative way
- Aliases and former notations

Amitay et al. (2004) introduced the term geo/non-geo ambiguity for toponyms that do not describe geographic places, but names (Jack London) or common words (Turkey). Lakoff & Johnson (1980: 35) first described the problem of metonymy as a figure of speech that uses „*one entity to refer to another that is related to it*". Market & Nissim (2002) classify

[1] http://www.acm.org/
[2] http://www.cikm.org/
[3] http://www.sigir.org/
[4] http://ir.shef.ac.uk/geoclef/
[5] http://www.metacarta.com/

metonymic expressions in place-for-event (*Beijing* 2008 was even bigger than *Germany* 2006), place-for-people (*Italy* eats the most pasta) and place-for-product (a 1984 *Bordeaux*) problems. Aliases can occur from different spellings ("St. John", "Saint John") or toponyms in different languages ("Lisbon", "Lisboa", "Lissabon"). Furthermore, several toponyms in the same language may point to the same referent ("Rome", "Eternal City") or toponyms may change over time ("Yugoslavia", "former Yugoslavia", etc.). Leidner (2008) names the process of resolving these alias problems place-name normalisation.

2.1.1.3. Toponym Resolution

Correctly recognised toponyms from an unstructured text can still be ambiguous. Therefore, toponym resolution is concerned with finding out which concrete instance of "Springfield" or "Sheffield" is meant in the document and assigning it the according pair of coordinates. Contrasting with the above mentioned geo/non-geo ambiguity, Amitay et al. (2004) introduce the term geo-geo ambiguity if a recognised place name is not unique. According to the number of publications and described approaches, toponym resolution appears to be one of the best investigated research fields within GIR. Quite a number of methodologies or heuristics can be found when investigating the different publications and authors. The following non-exhaustive table has been compiled according to chapter 2.3 in Brunner (2008). It gives a brief overview of the heuristics, their fundamental ideas and the corresponding authors. Most authors, however, use a combination of methods and ideas for their toponym resolution.

Heuristic	***Principle***	***Author***
Hierarchical context	Use of hierarchical properties and relationships between unique and ambiguous toponyms.	Hauptmann & Olligschlaeger (1999) Amitay et al. (2004)
Gazetteer reduction	Places are removed from the gazetteer depending on their population	Pouliquen et al. (2004) Krupka & Hausman (1998)
Described content	Probability of a toponyms correctness influenced by occurrence of keywords in the context	Rauch et al. (2003) Li et al. (2003)
Default Referent	Definition of one single default referent per toponym based on highest probability (mostly by population)	Ferrés (2007) Li et al. (2003)
Co-occurrence	List of toponym Co-occurrences for each toponym. Refined by geometric constraints	Overell & Rüger (2006)
One Referent per Discourse	An ambiguous toponym with multiple appearances in one text is used for one instance only	Amitay et al. (2004) Hauptmann & Olligschlaeger (1999)
Spatial information / correlation	Spatial information of toponyms is used: Minimum kilo-metric distance to next unique toponym, minimum distance to centroid, minimum convex polygon of unique toponyms	Pouliquen et al. (2004, 2006) Smith & Crane (2001) Leidner et al. (2003)

Tab. 2-1: Overview of heuristics for toponym resolution

2.1.1.4. Query Expansion and Relevance Feedback

Query expansion and relevance feedback are both concerned with finding additional terms and geographic concepts in order to create a more complete list of terms for the searching algorithms. A definition of query expansion is given in Fu et al. (2005) as "*a process of supplementing a query with additional terms as the assumption is that the initial query as provided by the user may be an inadequate representation of the user's information needs*". Moreover they classify query expansion techniques into two categories, one of them based on initial search results using relevance feedback (Gey & Petras, 2005, Bischoff et al. 2006b) and the other one using additional knowledge structures such as domain and geographic ontologies (Fu et al. 2005). Inadequate or incomplete user needs area explained by Cardoso et al. (2007): "*A user may not know the exact name of the location of interest and use an indirect description (5 star Hotels in the capital of Portugal), or a user may use an alias of a group of scopes (5 star Hotels in the Portuguese islands)*". Their approach for query expansion relies on geographical feature types, such as cities, airports, seas, islands, etc. To find out which feature type is the most promising for geographical query expansion, the frequency of feature types within the initially top ranked documents is analysed.

Relevance Feedback (RF) or Blind Relevance Feedback (BRF) consists in looking for terms within a set of initially top ranked documents, in order to expand the initial query. A definition of relevance feedback is found in Chen & Gey (2004): "*First, an initial search using the original queries is performed, after which a number of terms are selected from the top-ranked documents that are presumed relevant. The selected terms are weighted and then merged with the initial query to formulate a new query*". The selection and weighting of terms to be added for an expanded search is crucial to the method's success. Gey & Petras (2005) define "*"Good" terms to be added are terms that are relevant to the query and add new information to the search, for example synonyms of query terms but also proper names or word variations*".

2.1.1.5. Evaluating GIR

In this chapter, the IR and GIR evaluation, which stand at the end of the IR/GIR processing chain, are looked at. The relevance ranking, which is performed before the evaluation, with an emphasis on spatial relevance ranking is introduced in chapter 2.1.2. Spark et al (1997) differentiate between a system-centred and a user-centred IR evaluation strategy. The former is concerned with measuring system performance and comparison of different systems, the latter with usability issues involving end user requirements. The most frequently quoted

measures that can be considered system centred are precision and recall, as first introduced by Kent et al. (1955). Precision is the fraction of retrieved documents that are relevant with respect to all documents in the ranking.

$$p = \frac{relevant\ \ documents \cap\ \ retrieved\ \ documents}{retrieved\ \ documents}$$

Recall stands for the fraction of relevant documents retrieved compared to all relevant documents in the document collection.

$$r = \frac{relevant\ documents \cap retrieved\ documents}{relevant\ documents}$$

Fallout is the probability of a non relevant document being found. These measures are also employed for toponym detection or resolution evaluation. They use correctly resolved toponyms in a document as a measuring unit instead of relevant documents in the ranking.

Van Rijsbergen (1979) introduces the F-Value that is the weighted mean of precision and recall with β as weighting parameter.

$$F_\beta = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

If β is 1, the weighted mean value becomes the weighted harmonic mean value, also known as F_1 value (Van Rijsbergen 1979).

$$F_1 = \frac{2 * P * R}{P + R}$$

Another measure based on precision and recall is the precision/recall at *n* concept. Precision and recall are measured at *n* documents retrieved or the *n*th position of the ranking. A measure that emphasizes ranking positions is Average of precision (AP). For each position at which a relevant document is found, the retrieved precision is computed, whereas the precision of an unretrieved relevant document is 0 (Buckley & Voorhees 2000). To find the Mean Average Precision (MAP), AP is computed at predefined recall levels over a group of processed queries of equal relevance. The two most important frameworks for IR/GIR evaluation are the Text REtreival Conference (TREC[6]) and the Cross Language Evaluation Forum (CLEF[7]) that will be introduced in more detail in chapter 2.1.4

[6] http://trec.nist.gov/
[7] http://www.clef-campaign.org/

2.1.2. Space in GIR

2.1.2.1. Introducing Space and spatial Relevance

Cai (2002) suggests a geographic and a thematic subspace for GIR, representing the two different aspects of relevance. This chapter focuses on the spatial subspace of GIR. To find spatially relevant documents to a query, a spatial relevance referring to the query and the document footprints has to be defined. Spatial relevance is an important concept for the spatial result ranking, but also for the spatial relevance judgement of documents. Two geographical principles underlie most of the approaches in spatial relevance ranking as well as in spatial relevance judgement. They are implicitly found in all studies introduced in the next chapters: Tobler's first law of Geography, "*Everything is related to everything else, but near things are more related than distant things*" (Tobler 1970: 236) and the "*topology matters, metric refines*" principle by Egenhofer & Mark (1995: 9). An interpretation of these two rules applied in a GIR context could be as follows:

- Documents containing footprints near the query footprint are more relevant to the query than documents containing distant footprints and
- Topological information (contain, overlap, touch, equal, disjoint, etc.) about query and document regions is considered more important, than metric properties such as area, distance and shape.

2.1.2.2. Spatial Relevance Ranking

Ranked lists are the most common way to return a set of documents obtained from a query. The aim of a ranking is to place documents that are most likely to be relevant for the user first in the list. A score is calculated for each document according to the query and then used to sort out the documents in decreasing order. Since queries in GIR are more complex and composed of several relevant aspects, sorting them out by one single score might not be the most useful approach. Instead, an overall score can be divided into a thematic score and a spatial score (Kreveld et al. 2005).

A frequently found concept for creating spatial ranking scores is the calculating of spatial similarity between the query and the document footprints. Larson & Frontiera (2004a) list three basic approaches for spatial similarity computation: simple overlap, topological overlap and extent of overlap. In Frontiera et al. (2008) these are more precisely termed as metric spatial characteristics, topological spatial relationships and directional spatial relationships. Andrade & Silva (2006) compute the similarity between two scopes based on the information

found in the geographic ontology. Inclusion, proximity and siblings are calculated by using hierarchical and metric information, and combined to a geographical similarity value ranging from 0 to 1. Their system is based on the "one scope per discourse" principle, meaning that one single encompassing scope is assigned per document.

A different approach to assessing spatial similarity uses probabilistic models as initially proposed by Maron & Kunhs (1960). The logistic regression model of information retrieval introduced by Cooper et al. (1992) and found in Larson & Frontiera (2004b) and Frontiera et al. (2008) is based on probabilistic ranking (Robertson, 1977) and treats the spatial similarity score as the probability according by which a particular document is relevant to a particular query. The estimated probability of relevance is calculated as the log-odds of relevance and converted from odds to probability. Frontiera et al. (2008) compare their logistic regression approach to five different GIR methods that calculate spatial similarity as a function of single spatial metric property, and conclude that the former outperforms all of the non-probabilistic models in terms of precision and recall.

Yu & Cai (2007) compute a term- and a geographical query specificity in order to determine the relative weights of the respective relevance scores. The thematic and geographical specificity attempts to measure how specific or general the queried concepts are. The term specificity is calculated by using a combination of Inverse Document Frequency (Jones, 1972) and an ontology based computation employing conceptual information in hierarchical structures. The geographic specificity is based on the ratio between the query scope and the collection's total covered scope. The calculated specificity is used for an effective combination of the thematic and the spatial relevance scores between the query and the documents. Thematic relevance is calculated by Lucene[8], geographic relevance is a binary 1/0 score if topological containment between query- and document scope is true, if not, Euclidian distance is used. The authors further investigate how to combine the two types of relevance and conclude that the Dempster-Shafer theory based merging achieves highest agreement with human judgement.

Kreveld et al. (2005) introduce the multi-dimensional scattered ranking. The aim of their efforts is to generate GIR results with a high spreading in space and to de-cluster result point distributions. So far, documents were judged on their thematic and spatial relevance to a query only. In this method, however, the position of all other points is taken into account in order to calculate a point's relevance. A point p from a set of not yet ranked points is chosen

[8] http://lucene.apache.org/java/docs/index.html

as the best point to rank next, if it achieves the highest value in a scoring function depending on the distance to the query and the distance to the closest already ranked point. Once a new point is added from the unranked to the ranked point set, distances to the closest ranked point of all remaining unranked points need to be updated with the last ranked point and replaced if the new smallest distance is smaller than the current. A number of variations of that basic idea are presented as well as two extensions improving the selection of the next "best to rank" point. The authors state that their methods appear to perform well in terms of distance to query and high spreading, but acknowledge that the evaluation is based only on the visual inspection of the outcome. The implementations of these methods are part of the relevance ranking unit of SPIRIT, a GIR system that will be introduced in chapter 2.1.3.

2.1.2.3. Spatial Relevance Judgement

In order to evaluate a GIR ranking result, knowledge about which documents in the collection are relevant to a query is necessary. The relevance judgment consists in assigning documents a value indicating an amount of relevance to a query. The spatial relevance judgement thus takes into account the spatial relevance of a document to a query only. In general, relevance judgement appears to be a little elaborated field within GIR so that little detailed literature is available. Studies concerned with relevance judgement can be found in relation with the SPIRIT project and the geoCLEF campaigns, both of which will be introduced in more detail within the next chapters.

In order to evaluate the effectiveness of their spatial ranking approach, Frontiera et al. (2008) consider a document geographically relevant if the document footprint has any area of overlap with the query footprint. A similar approach is taken at the GeoCLEF, where documents are judged in a binary system (2.1.4.4). The idea of separating thematic relevance from geographical relevance (multidimensional relevance) is applied in a series of studies concerned with the SPIRIT system. Bucher et al. (2005) and Clough et al. (2006) use two ternary schemes and have spatial and thematic relevance judged by human assessors. In the former work, inter-annotator agreement is found to be significant for spatial and thematic relevance, while in the latter, inter-annotator agreement is relatively poor. Furthermore, Bucher et al. (2005) remark that "*many assessors also commented on this measure of thematic relevance being unnecessarily cumbersome and expressed a preference for a binary scheme*". Clough et al. (2006) also use the reduced binary scheme for further analysis of their results. Purves & Clough (2006) continue the previous efforts by splitting up spatial relevance once more into the spatial relevance of a document's content location (Wang et al. 2005) to the

query footprint and the relevance of the retrieved footprints to the content location of the document they are associated with. They found that inter-annotator agreement for spatial relevance is strongly related to the query type. Queries containing the “in” operator achieved a much higher agreement value than queries built with near, around or directional spatial operators. These can be interpreted more variably and will cause additional difficulties for judging spatial relevance.

2.1.3. SPIRIT

2.1.3.1. Introducing SPIRIT

Since the test data for the first evaluating block in chapter 3 are results produced by the SPIRIT system, a more detailed introduction of SPIRIT is given in this chapter. SPIRIT (Spatially aware Information Retrieval on the Internet) is a complete solution to geographic information retrieval (Jones et al. 2002) that uses a footprint based approach of document retrieval for unstructured text. SPIRIT was a common project developed by a variety of research groups located at the Universities of Zurich, Hannover, Sheffield, Cardiff, Utrecht and the Institut Géographique National (IGN), France. One half of the participating research groups belongs to the Information or Computer Science Departments of their Universities. Zurich’s Department of Geography, the Hannover Institut für Kartographie und Geoinformatik (IKG) and the IGN represent geographic research institutions. Nevertheless, the fundamental hypothesis to all SPIRIT related research activities is “*that search techniques which take explicit account of geographic content and spatial relationships will provide more accurate results than pure text search for queries which include geographic content*” (Purves et al. 2007: 721). The main parts of the SPIRIT architecture are the user interface, a broker component that controls and schedules the information flow through the system, the core search engine, the relevance ranking and the geographically and textually pre-processed data basis. The multi-modal user interface supports textual (<theme><relationship><location>) input and interactive map feedback of the context of retrieved documents. The range of spatial operators includes in, near, outside, north, south, east, west and within distance of (Purves et al. 2007).

2.1.3.2. SPIRIT Routines

The basic processes of any GIR system were introduced in the previous chapters. In this section, a closer look at their implementation within the SPIRIT system is taken. SPIRIT routines can be roughly classified into two groups: Pre-processing and run-time operations. In order to actually allow geographic search, the underlying document collection needs to be textually and spatially pre-processed. The Data basis of the SPIRIT systems is one terabyte collection of approximately 94 million web pages (Joho & Sanderson, 2004).

Pre-processing includes assigning spatial footprints to these web documents and building thematic and spatial document indexes. Assigning spatial footprints can be further divided into the processes of geoparsing and geocoding (McCurley, 2001) that correspond to 2.1.1.2 (Toponym Recognition) and 2.1.1.3 (Toponym Resolution). SPIRIT uses the GATE (General Architecture for Text Engineering) for Information Extraction for geoparsing (Cunningham et al. 2002) which combines a place names gazetteer lookup with additional context rules to filter out metonyms. The underlying geographic data sources are the SABE (Seamless Administrative Boundaries of Europe) dataset and the Ordnance Survey 1:50’000 scale gazetteer. For the geocoding and the resolving of place name ambiguity, which results in around 89% of correctly assigned place names, a default sense approach and global geographical knowledge are used. After the stage of geoparsing, document collection size decreased to 885’502 web pages covering the United Kingdom, France, Germany and Switzerland (Purves et al. 2007).

The main components of run-time operations can be summarized in query handling, retrieving initial documents by the search engine, relevance ranking and the display of results. A query, when it is first entered needs to be disambiguated and expanded based on the geographic ontology, which contains knowledge (aliases, place type, geographic footprint and topological relationships) of all places within the geographic coverage of the search engine. Given an ambiguous toponym in a query, the user will be prompted to specify which instance of the toponym should be used. The user is supported by the system providing additional geographical and hierarchical information for all instances. For query expansion, the same geographical ontology is accessed in order to expand the spatial query terms. Enlarged geometric query footprints are generated with respect to the initial place name in the query and the spatial operator (Purves et al. 2007).

As seen in 2.1.2.2, relevance ranking is divided into two main tasks, textual and spatial relevance ranking. A spatial footprint similarity score is first produced for each footprint in

the document. A document's spatial similarity score is then generated based on all footprint similarity scores contained in the document and combined with textual BM25 (Robertson et al. 1998) weighting function scores. Since SPIRIT supports several spatial operators, the calculation of spatial similarity score is dependent on the spatial relationship applied in the query. In order to combine the two relevance types into one, two different approaches are mentioned, both of which rely on the Euclidean distance from their specific scores (spatial [normalised to 0-1], thematic [normalised to 0-1]) to point (1,1), which is assumed to be the most relevant document possible. The non distributed method ranks the documents strictly in descending order of their distance to point (1,1), while the distributed method tries to de-cluster documents having similar score components (Purves et al. 2007, Kreveld et al., 2005).

2.1.4. GeoCLEF

2.1.4.1. Introducing GeoCLEF

GeoCLEF is an evaluation campaign for the performance measuring of GIR systems. CLEF stands for Cross- Language Evaluation Forum and "*promotes research and development in multilingual information access, by developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European Languages (…) and creating test-suites of reusable data which can be employed by system developers for benchmarking purposes*"[9]. Until now, after each GeoCLEF, a summary of search topics, participants and results has been published, on two of which this chapter is based: Gey et al. (2007b) and Mandl et al. (2008).

CLEF arose from the Cross Language Information Retrieval (CLIR) track, a task at the Text REtrieval Conference (TREC[10]), which focused on cross lingual information retrieval in European, Arabian and Asian languages. CLEF can be seen as European counterpart to TREC and first took place in 2000 (Agosti et al., 2007, Kluck et al., 2002). CLEF before 2005 and other campaigns were not explicitly concerned with geographical aware searching. "*The aim of GeoCLEF is to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects*" (Gey et al. 2007a). GeoCLEF was first conducted in 2005 as a pilot track in English and German, with 11 participating groups. In 2006, document languages were English, German, Portuguese and Spanish, with 17 participating groups, whereas in 2007 document collections and topics were only available in English, German and Portuguese and the number of participants decreased to 13.

[9] http://www.clef-campaign.org/
[10] http://trec.nist.gov/

Participant research groups were from the United States, Western Europe Asia and Australia with the United States, Germany and Spain as most represented countries.

2.1.4.2. Search Tasks

At every GeoCLEF so far, 25 search topics were defined by the organizing groups. These topics are the main task by which the participant's performance is measured. All participants share the same data basis. The document collections consist of newspaper and news agency articles of *The Glasgow Herald* and *The Los Angeles Times* in English, *Der Spiegel*, *Frankfurter Allgemeine Zeitung* and *Schweizer Depeschen Agentur* in German, and *Público* and *Folha de São Paulo* in Portuguese, dating from 1994 and 1995. In 2007 the document collection included 169'477 articles in English, 294,809 articles in German and 210,734 articles in Portuguese.

The topics are created using the DIRECT system (Di Nunzio & Ferro, 2005) provided by the University of Padua. Each group creates initial versions of proposed topics in their language, which are subsequently checked for relevant documents in the other collections. If too few relevant documents are found in the other collections, topics need to be improved on. After the final 25 queries are marked, they are translated into all other languages: In 2005, English and German, in 2006, English, German, Portuguese, Spanish and Japanese and in 2007 only English, German and Portuguese. Furthermore, search tasks are divided into monolingual (English to English, German to German, etc.) tasks and bilingual (language X to language Y) tasks.

2.1.4.3. Geographical Issues

Gey et al. (2007) came up with a tentative classification for geographical topics at the 2006 GeoCLEF. The following table represents how topics depend on a place or can be considered "geographic":

1	non-geographic subject restricted to a place (music festivals in Germany) [only kind of topic in GeoCLEF 2005]
2	geographic subject with non-geographic restriction (rivers with vineyards) [new kind of topic added in GeoCLEF 2006]
3	geographic subject restricted to a place (cities in Germany)
4	non-geographic subject associated to a place (independence, concern, economic handlings to favour/harm that region, etc.) Examples: independence of Quebec, love for Peru (as often remarked, this is frequently, but not necessarily, associated to the metonymical use of place names)
5	non-geographic subject that is a complex function of place (for example, place is a function of topic) (European football cup matches, winners of Eurovision Song Contest)
6	geographical relations among places (how are the Himalayas related to Nepal? Are they inside? Do the Himalaya mountains cross Nepal's borders? etc.)
7	geographical relations among (places associated to) events (Did Waterloo occur more north than the battle of X? Were the findings of Lucy more to the south than those of the Cromagnon in Spain?)
8	relations between events which require their precise localization (was it the same river that flooded last year and in which killings occurred in the XVth century?)

Tab. 2-2: Classification of relations between topics and geographic space

From 2005 to 2007, the topics became geographically more challenging each year. This means that for the successful retrieval of relevant documents, explicit geographic knowledge is necessary and keyword-based approaches should not be favoured by the topics. Difficulties that were in introduced in 2006 and kept in 2007 include ambiguity (St. Pauls Cathedral, exists in London and São Paulo), vague geographic regions (Near East), geographical relations beyond IN (near Russian cities, along Mediterranean Coast), cross-lingual issues ("Greater Lisbon", Portuguese: "Grande Lisboa" , German: "Großraum Lissabon"), granularity below the country level and imprecise regions (Northern Italy), and complex region shapes (along the rivers Danube and Rhine) (Mandl et al. 2008). Within the 2007 topics, even more complex geographic relations were formed: East Coast of Scotland, Europe excluding the Alps, main roads north of Perth, the Mediterranean coast or Portuguese islands. Additionally, two more difficulties were introduced: Politically defined regions; smaller than countries (French speaking part of Switzerland, the Bosphorus) or larger than countries (East European countries) and finer geographic subjects, being lakes, airports or F1 circuits (Mandl et al. 2008).

2.1.4.4. Relevance Assessment

The first step in relevance assessment is to narrow down the entire document collection to so-called document pools. These document pools are extracted from monolingual and bilingual runs and assembled by language using the DIRECT system. The English pool in 2007 consisted of 15'637 documents that were used for human relevance assessment. Human assessors, mostly volunteers organised by the participating groups, judge the remaining documents manually. In this context relevance is treated as binary. A document can either be judged relevant or irrelevant, without the explicit distinction between thematic and spatial relevance being made. A document is judged relevant if any piece of the document is relevant, regardless of how small the piece is in relation to the rest of the document[11]. The outcome of the human relevance assessment is the qrels file. In the qrels file, all documents out of the pool are grouped by query (approx. 600 documents per query) and added a binary code for relevance. The order of documents in the qrels file is incidental and implies no indication of relevance. If a document does not occur in the qrels file, it is assumed to be irrelevant and is not judged by the human assessors[11].

Topic	Iteration	Document#	Relevancy
001	0	LA113094-0044	1
001	0	LA113094-0130	0
001	0	LA050794-0225	0

Tab. 2-3: Extract of the qrels file

The column Topic stands for topic or query number, iteration is the feedback iteration, document# is the official document ID, and relevancy is a binary code of 0 standing for not relevant and 1 for relevant[11].

The results of the relevance assessment in the qrels file are used to evaluate the participants' retrieval performance. The evaluation at GeoCLEF is based on standard IR measures such as mean average precision and recall versus average precision. The fraction of relevant versus irrelevant documents is easily found by comparing the document ID's of the retrieved documents to the relevant document ID's in the qrels file.

2.1.5. Improvement and Challenges for current GIR Evaluation

Gey et al. (2007b) divide challenges to evaluation within GeoCLEF into two areas. The first area concerns general problems which occur within the area of ad-hoc retrieval of

[11] http://trec.nist.gov/data/qrels_eng/

multilingual collections, the second specific problems due to the particular character of geographic information retrieval. An example for the first group is insufficient judgment pools. In GeoCLEF 2006, only three groups participated in Portuguese, four groups in Spanish and German, while 16 groups did English GIR. Problems of the second area arise because space and geography are not appropriately included in relevance judgement, which is a key motivation for the present masters thesis. The authors describe this problem as "*finding a consistent basis for relevance assessment*" (Gey et al. 2007b) and use the query "Cities along the Danube and the Rhine" from GeoCLEF 2006 as an example for it. They conclude that "*a complex geospatial query should have been applied to the GNS* [Geographic Names Server] *by taking the digital lat-lon coordinates the Rhine and Danube Rivers and computing a geospatial cover with a perpendicular line to each line segment specifying the river, together with circles covering the join points on the where the polygonal curve changes to a new line segment*" (Gey et al. 2007b). In other words, a spatial extent in which relevant results are expected is defined and used for spatial relevance judgement. This is one approach that is reflected and applied in this present work. These mentioned and other evaluating challenges might be an issue at the International Workshop on Evaluating Information Access EVIA[12], which first took place in Japan in 2007.

2.2. Point Pattern Analysis

2.2.1. Point Patterns in geographical Space

A point pattern in space or geographical space can be thought of as a set of recorded points or events with assigned coordinates in a defined study region of arbitrary shape (Gatrell et al. 1996). Point patterns are created by processes, either real or hypothesised. O'Sullivan & Unwin, (2003: 54) define a spatial process as "*a description of how a spatial pattern might be generated*". The simplest process that can generate a point pattern is one where no spatial dependencies exist and space itself does not influence the distribution. Such a process is called Independent Random Process (IRP) with its underlying assumption of Complete Spatial Randomness (CSR).

According to O'Sullivan & Unwin, (2003: 58), the IRP postulates two conditions: First, the condition of equal probability, meaning that a point has an equal chance of lying at any position inside of the study area. Second, the condition of independence, meaning that the position of any point is independent of the position of any other point.

[12] http://ntcir.nii.ac.jp/index.php/EVIA-2008/

Most real world patterns, however, are not created by IRP and CSR processes, but show some kind of spatial dependence or mutual dependence. According to O'Sullivan & Unwin (2003: 65), there are two ways in which real world patterns differ from IRP and CSR: a first order effect and a second order effect. First order effects can be described as the dependence of the point distribution on the environment, whereas second order effects refer to the dependence of an event's location on other events' locations. A definition of first and second order processes is given by Gatrell et al. (1996: 259): "*Very informally, the first-order properties describe the way in which the expected value (mean or average) of the process varies across space, while second-order properties describe the covariance (or correlation) between values of the process at different regions in space*".

According to Gatrell et al. (1996), a point process is stationary if the overall point density (first order property) is constant throughout the study region. The Second order intensity only relies on the direction and the distance between two points, not on their absolute location. A process is called isotropic if the second order intensity only depends on the distance, not on the orientation of the difference vector between two points.

Applying quantitative analysis to spatial data includes a number of difficulties. One of these is called the modifiable areal unit problem (MAUP), as described by Openshaw & Taylor (1981). The MAUP illustrates that results of distribution and cluster measurements will vary by changing the study area. Another problem are edge effects. When measuring point pattern properties based on the number and distance of adjacent events, points lying near the study area border are denied the possibility of neighbours outside the study area (Gatrell et al. 1996). O'Sullivan & Unwin, (2003: 95) propose the use of a guard zone around the study area which is included for the search of nearest neighbour events. Most of the measures, however, are aware of edge effects and offer some kind of edge correction. Nevertheless, the definition of the study area requires special attention since it has a large influence on most point pattern measures.

2.2.2. Density based Measures

Density based measures of a point pattern characterise its first order properties. A simple density based approach is the pattern's crude density or intensity, as the mean number of point events per unit area. An improvement on overall intensity measurement are quadrat count methods. A regular grid of quadrates is superimposed over the event distribution and events per quadrat are counted and compiled into a frequency distribution. According to this idea, each single quadrat can be assigned its calculated density value and a value-coverage for the

entire study area can be created. Kernel density estimation is a commonly used procedure that converts a point pattern into a field representation giving an estimated point density at each point of the study area. The estimation is computed by counting the number of point events within a specified region or kernel. The kernel moves across the study area and calculates estimations for each kernel-centre. (O'Sullivan & Unwin, 2003: 81-88).

According to Gatrell et al. (1996), the estimated intensity is defined as

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right)$$

where k represents the standardised kernel form, s a vector of points where the function is evaluated, s_i a vector of observed events and τ the kernel bandwidth. The most sensitive parameter is the kernel bandwidth, which defines the size of the radius in which point events will be counted. A too large bandwidth will result in a surface with very similar values close to the mean, while a too small bandwidth gives a surface focused strongly on individual events (O'Sullivan & Unwin, 2003: 85-87). Seaman & Powell (1996) explain several methods for bandwidth definition and conclude that the least squares cross validation produced the most accurate estimates. This method performs several thoroughfares by using different bandwidths and chooses the one producing a minimum estimated error. "*This score function is an approximation of a jacknife estimator and essentially uses subsets of the data to determine the bandwidth that gives the lowest mean integrated squared error for the density estimate*" Seaman & Powell (1996: 2077). The final shape is created by multiplying the found bandwidth by the standard deviation of the x and the y dimension of the data and results in an asymmetrically elongated kernel.

Not only different bandwidths, but also different kernel shapes, which produce essentially equivalent results, are available. The kernel form is a decreasing radially symmetric bivariate function providing a total weight of unity over the region of influence and has a volume integrating to one (Gatrell et al.1996). Furthermore, a kernel can be defined as adaptive kernel with more smoothing in low density areas and less smoothing in high density areas (Seaman & Powell, 1996).

2.2.3. Distance based Measures

Distance based measures rely on distances between points of a pattern and estimate the second order properties of a pattern's underlying spatial process (Gatrell et al.1996). The Clark and Evans aggregation index is a measure of the degree to which a distribution in a given area departs from that of a random distribution. The basis for this measure is each point's nearest neighbour distance. The resulting R-value is the ratio of the observed nearest neighbour distance mean to the mean distance of a random distribution with an equal point density. Values of R vary between 0 and 2.1491, indicating completely aggregated and perfectly uniform distributions respectively. R = 1 is evidence for a random distribution (Clark & Evans, 1954: 445-447).

The G, F and K functions are three examples for measures that try to go further than the nearest neighbour approach. The G function is the most simple of them and is based on the nearest neighbour distance of each event. The G function is the cumulative frequency distribution of the nearest neighbour distances and reveals what fraction of all nearest neighbour distances is less than a distance d, plotted on the x-axis. If events are clustered, G increases rapidly for low distances; if events are evenly distributed, G increases sharply for those d values that correspond to the spacing between the points (O'Sullivan & Unwin, 2003: 89-91).

The F function is similar to the G function, but instead of summarising nearest neighbour distances between all events in the pattern, it uses a set of randomly selected points at any location in the study area. Values of the F function correspond to the cumulative frequency distribution of the distances between the random points and their nearest neighbour event in the original pattern. An F function's graph for a clustered pattern has a rather constant gradient, since distances between the random points and the cluster vary and produce no clear increase at one particular distance. (O'Sullivan & Unwin, 2003: 91).

Ripley's K function (Ripley, 1976) is based on all distances between events in the study area and might therefore carry more information about the distribution than the two above functions. It uses an increasing radius d, centred on each event, which defines the circular search area around every event in the pattern. Fig. 2-1 illustrates an event pattern with different stages of d around each event.

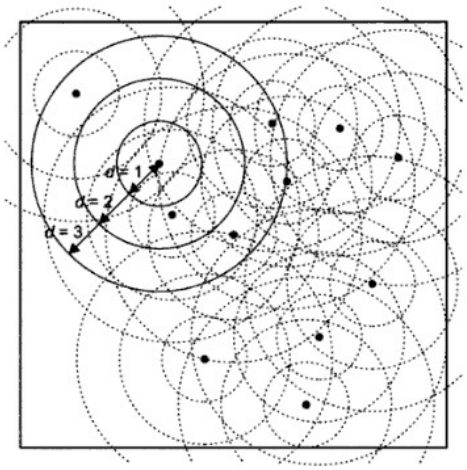


Fig. 2-1: K function principles (O'Sullivan & Unwin, 2003: 94)

At each stage, the number of points found within the search radius is counted. For each d, the mean count for all events is calculated and divided by the overall event density. A clustered pattern produces a graph with a high gradient for small radiuses d. The G, F and K functions' graphs can be compared to an expected graph, formed by a uniform Poisson process under IRP and CSR conditions. The form of the expected graph under IRP/CSR conditions for the G and F function is given by

$$E[G(d)] = 1\text{-}e\text{-}\lambda\pi d^2$$

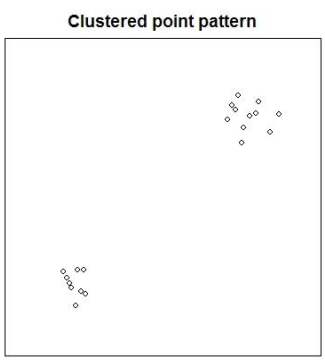
$$E[F(d)] = 1\text{-}e\text{-}\lambda\pi d^2$$

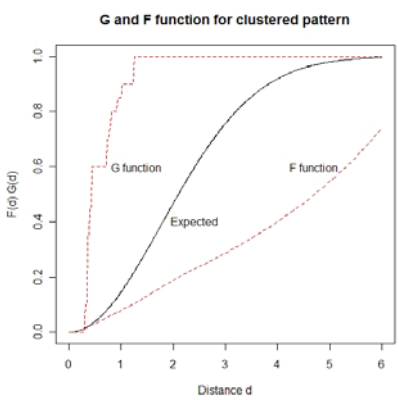
and for the K function by

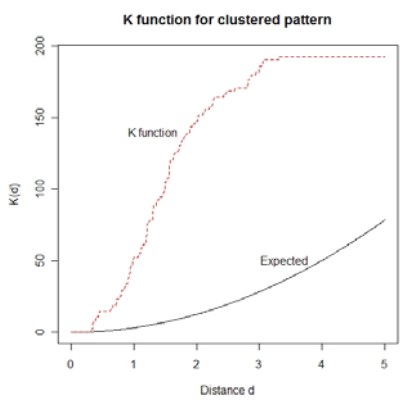
$$E[K(d)] = \pi d^2$$

with λ representing the mean density (O'Sullivan & Unwin, 2003: 101-103). Tab. 2-4 illustrates these three expected graphs compared to their according graph under clustering conditions.

G, F and K function; expected under IRP/CSR conditions versus cluster







Tab. 2-4: G, F and K function under IRP/CSR conditions versus cluster conditions

2.2.4. Related Work

Gatrell et al. (1996) divide cluster detection and cluster analysis into two different groups with differing underlying null hypotheses: The first group is an exploratory data analysis approach that checks for clusters in a given pattern within a given study area. The second group's underlying idea is to verify whether a cluster can be found in the vicinity of a predefined location of interest, a so-called *a priori* location or not. The goal of both approaches, however, is to produce a statistically significant result proving the existence of a cluster. Both methods test against a Poisson distribution. Openshaw et al. (1991) outline possible computation algorithms for both processes.

Examples of the first group are Openshaw et al. (1987) or Kulldorff (1997). Openshaw et al. (1987) introduce the Geographical Analysis Machine (GAM) for the analysis of point pattern data. "*The original GAM involved the use of a circular pattern detector and a locationally comprehensive search based around intersection points on a lattice*" (Openshaw et al., 1991: 399). A GAM procedure examines several circle sizes and the corresponding point frequencies were checked against a Poisson probability. Circles with the lowest probability of being due to chance were used for identifying subregions in which point distribution was not random. In addition to the mapping of the phenomenon of interest, population base data is necessary to compute the case frequency under the null hypothesis of uniform distribution and to eliminate effects that are caused by any different variable than the one investigated (Openshaw et al., 1991).

Kulldorff (1997) introduces the "*spatial scan statistic*", which shares a number of similarities with the GAM. The method creates a regular or irregular grid of centroids covering the whole study region. Then, an infinite number of circles around each centroid are created. The radius may vary between zero and a maximum, meaning that at most 50% of the population is included. A number of actual and expected cases inside and outside the circle is obtained. The circle with the highest likelihood function, depending on the number of cases, is declared the Most Likely Cluster and compared to equivalent random distributions under the null-hypothesis of no clusters. The spatial scan statistic is applied in Kulldorff et al. (1997) in order to detect breast cancer clusters in the northeast United States. A significant result is found, which claims that the mortality rate in the New York City – Philadelphia metropolitan area is 7.4% higher than in the rest of the northeast.

An example of the second group is Diggle (1990), who investigates the relationship between the spatial distribution of larynx cancer cases and a formerly used industrial incinerator in the

Chorley-Ribble area. The null hypothesis states that cancer intensity is equal to the intensity of the background population. The calculation method is based on a multiplicative model of case intensity, which corresponds to a function composed of two elements: the background intensity and the distance from the point source. Evidence is found for elevated cancer risk in the vicinity of the incinerator. Diggle differentiates his method from the above approaches, saying that "*it shifts the focus of the analysis away from the somewhat artificial definition of clusters of events and towards a quantitative description of variation on local intensity around a prespecified point*" (Diggle, 1990; 360).

2.3. Conclusions and Motivation

In current relevance judgement, documents are mostly judged on a binary scale and independently of the geographic scopes and ranks of other documents. As seen in chapter 2.1.2.3, the relevance judgement of documents can be refined by judging spatial and thematic relevance separately and by using a ternary scheme. Bucher et al. (2005), however, state that annotators prefer, at least when judging thematic relevance, a binary scheme and Clough et al. (2006) find unsatisfying inter-annotator agreements for the ternary schemes. Furthermore, the problem of ignoring other documents and their footprint distributions is not resolved by using this approach. The evaluation of GIR results at GeoCLEF is equal to the evaluation of classic IR results, which is based on precision and recall. The spatial aspects of GIR result rankings have not yet been explicitly evaluated.

The spatially aware evaluation methodology to be developed in this present work should thus provide a simple document relevance judgement (binary scheme), refer to the whole set of ranked relevant documents and footprints, and reveal differences in spatial quality (covered extent, clustering behaviour) between single documents in the final result. Methods for explorative descriptions of point patterns used in GEO sciences are therefore tested for suitability and integrated into the relevance judgement and evaluation processes.

Why are precision and recall not good enough? Given a query "Milk consumption in Europe", a ranking result that covers only one European country would not be considered different from a result covering large parts of Europe, if precision and recall are equal. This contrasts with Kreveld et al. (2005), who propose two requirements for a good spatial ranking: Small distance to the query and high spreading. A more sophisticated GIR evaluation should thus include measures beyond precision and recall. The spatial quality of a GIR ranking needs to be judged with spatial indicators such as the covered extent of the ranked document footprints or characteristics of the spatial footprint distribution. A partially comparable technique, which

measures the area of relevant overlapping bounding boxes and convex hulls is mentioned in Frontiera et al. (2008).

Considering the ranking position of a document is important for two reasons. First, eye tracking experiments performed by Joachims et al. (2005) reveal that users often focus just on the first few results and ignore the rest. Second, given a set of perfectly working GIR systems accessing the same data basis and retrieving all thematically and spatially relevant documents, the resulting point distributions of the entire rankings, irrespective of the rank, will always look the same, since the same documents containing the same footprints are found. A general rule for considering the ranking positions is that a relevant document appearing early in the ranking is better than a relevant document appearing late.

An additional aim derived from this current state of research overview is to investigate the presuppositions for spatial search that are provided in the GeoCLEF campaigns. The aim of GIR efforts is to produce better retrieval results than purely text based search for queries with a spatial condition. But are the GeoCLEF queries and document collection spatially considerate enough and can spatially aware search outperform text search under the given circumstances?

The aims of this present work can be summarised in three points:

- To develop and test point pattern measures for applicability in GIR footprint distribution evaluation.
- To apply spatial measures in a new evaluation system beyond precision and recall.
- To investigate the “spatialness” of GeoCLEF queries, document collection and its footprint occurrences, by using the SPIRIT system as benchmark.

3. Evaluating SPIRIT data

3.1. SPIRIT Data

In order to develop a methodology for the description of GIR footprint distributions, GIR data is necessary. 20 queries and their corresponding results produced by the SPIRIT system are used as test dataset in this chapter. Tab. 3-1 gives an overview of the used queries.

museums in cardiff united kingdom
hotels in cardiff united kingdom
mountaineering in scotland united kingdom
oil industry in aberdeen united kingdom
camping in Highland,Scotland,United Kingdom
beaches in cornwall united kingdom
walking in fife united kingdom
pubs in edinburgh united kingdom
shipping in liverpool united kingdom
schools in norwich united kingdom
climbing near aviemore united kingdom
camping near Lancaster,Lancashire,United Kingdom
hotels near edale united kingdom
walking near beauly united kingdom
canals near stroud united kingdom
red kites near Cromarty,Upper Ythan,United Kingdom
walking outside edinburgh united kingdom
cycling south london united kingdom
castles east edinburgh united kingdom
hotels west fort william united kingdom

Tab. 3-1: 20 Queries performed with SPIRIT

SPIRIT raw data comes as a *txt* file structured with records in rows and attributes in columns. The records can be assigned to a query by the column *docID*. A document itself can contain various relevant place names so that several records can be associated to the same *docID*. The attribute *Rank* gives the number of the document according to the spatial and the thematic ranking. The dataset contains the first ten ranks for each query. Each record has a set of coordinates of the location it has been assigned to. The coordinates are represented in WGS84 and in a projected Lambert Conformal Conic system, whose details can be found in the appendix. Conformal projections preserve original shapes, but parts of the projection will be relatively enlarged or reduced. The Lambert Conformal Conic projection's area distortion is small near and between the standard parallels (Robinson et al. 1995: 74-77). It thus provides "*exceptionally good directional and shape relationships for an east-west mid-latitudinal zone*" (Robinson et al. 1995: 77) and is also easily interpretable, since it is in meters. The

columns *x* and *y* contain the Lambert coordinates and are used for the georeferencing and further data analysis.

3.2. Analysing spatial Extents

3.2.1. Building Result Areas from SPIRIT Point Data

The first step to describe GIR footprint distributions is to explore their locations and spatial extents. The footprint distribution's minimum bounding rectangles allow the quantifying of the extent size, using its side lengths as well as the geographic location, by using its centroid coordinate.

SPIRIT raw data is converted into the *dbf* format and imported to ArcMap 9.2, by using the *add xy data* procedure. The columns x1 and y1 of the SPIRIT data sheet are assigned to the x and y coordinate filed of the import function. The parameters of the SPIRIT Lambert projection are also imported into ArcMap in order to define a correspondent projection. This projection is then used for the entire dataset. Once the point clouds are imported in the correct projection, their minimum bounding rectangles are created. The computation of the minimum bounding rectangles is performed with the ET Geo Wizards function *features to bounding rectangle*, which aligns the new feature with the longest side of the original geometry. Since the creation of bounding rectangles with point input is not supported in ET Geo Wizards, a convex hull feature has to be created first. The minimum bounding rectangle is then computed around the convex hull feature. Once the bounding rectangle features are drawn, xy coordinates of the four corner points are extracted by using ArcToolbox *feature vertices to points*.

Corner point coordinates are exported as *dbf*, converted to *csv* and imported to R 2.6.2[13]. These four corner coordinates are then used to create an instance of the spatstat[14] class *owin*, which represents a two-dimensional observation window. Centroids of the result areas are calculated based on the *owin* object, by using the *centroid.owin* method. The side lengths are extracted by using Pythagoras and the corner coordinates.

[13] http://www.r-project.org/
[14] http://www.spatstat.org/

The point coordinates of the SPIRIT result point patterns are imported into R by using a *csv* import script. By joining the window feature and the correspondent query's point coordinates, an instance of the spatstat class *ppp* (planar point pattern) is created. These *ppp*-objetcs contain the information of the minimum bounding rectangles as well as all the point coordinates of the SPIRIT point pattern, and are the basis for the upcoming spatial measurements.

Tab. 3-2 gives a summary of some of the basic properties of the 20 result set areas. The column "unique points" represents the number of points left after all the duplicated points in the point pattern have been removed.

Query	*Points*	*Unique Points*	*Centroid WGS84*	*Centroid Lambert*	*Area (km^2)*	*X-dist (km)*	*Y-dist (km)*
museums in cardiff united kingdom	23	4	51.47246 / -3.23401	-1846695.24 / 301702.33	49.4	4.5	11.1
hotels in cardiff united kingdom	20	11	51.47994 / -3.161355	-1845107.12 / 301698.71	97.6	10.6	9.2
mountaineering in scotland united kingdom	50	29	57.78978 / -3.604741	-1623970.60 / 952595.93	110564.2	205.9	536.9
oil industry in aberdeen united kingdom	12	3	57.13928 / -2.124643	-1559007.09 / 843701.51	26.2	3.5	7.5
camping in Highland,Scotland,United Kingdom	21	15	57.44483 / -4.879424	-1723904.96 / 933857.27	25132.7	123.5	203.4
beaches in cornwall united kingdom	51	25	50.43077 / -5.075778	-2001324.55 / 230140.64	4632.7	114.1	40.6
walking in fife united kingdom	14	10	56.18552 / -3.253356	-1662823.16 / 774371.46	1343.4	55.2	24.3
pubs in edinburgh united kingdom	12	4	55.941 / -3.23585	-1669621.36 / 746675.70	9.9	6.6	1.5
shipping in liverpool united kingdom	10	1	53.39554 / -2.917015	-1752671 / 485115.5	-	-	-
schools in norwich united kingdom	21	11	52.63828 / 1.271298	-1524632.27 / 315054.76	30.9	5.9	5.2
climbing near aviemore united kingdom	21	15	57.23361 / -3.848522	-1650462.05 / 890363.80	3984.7	75.1	53.1
camping near Lancaster,Lancashire,United Kingdom	20	16	54.05779 / -2.721256	-1722320.68 / 544978.69	3687.2	40.9	90
hotels near edale united kingdom	21	17	53.37305 / -1.833509	-1688180.56 / 454887.30	1467.6	44.2	33.2
walking near beauly united kingdom	47	34	57.4898 / -4.585344	-1677073.41 / 933982.77	9193.2	83.5	110.3
canals near stroud united kingdom	30	22	51.67444 / -2.142843	-1771424.50 / 295939.39	2893	55.7	51.9
red kites near Cromarty,Upper Ythan,United Kingdom	11	8	57.54017 / -2.313963	-1549535.61 / 890204.75	1616.7	33.4	48.4
walking outside edinburgh united kingdom	19	12	55.91782/ -3.400946	-1676899.31 / 748262.36	1436.7	34.1	42.1
cycling south london united kingdom	15	14	51.12457 / 0.1814926	-1647327.78 / 185935.75	3888.2	96.7	40.2
castles east edinburgh united kingdom	20	12	55.96767 / -3.028612	-1659783.32 / 746166.39	615.1	24.4	25.6
hotels west fort william united kingdom	6	3	56.75349 / -5.341452	-1750749.41 / 877564.28	157.4	5.9	26.7

Tab. 3-2: Summary of result areas from SPIRIT point data

3.2.2. Building expected Areas of Interest

In this section, the spatial extents of the expected result point distributions are simulated for all queries. The idea is to compare the bounding box of the result point pattern to a defined area of interest, in which the results are expected. Local knowledge as well as geographic and thematic conditions are explicitly considered. For each query, an expected spatial extent is defined by using background data from the Digital Chart of the World (DCW[15]), GeoNames[16], online maps (Google Maps[17]) or satellite imagery (Google Earth[18]).

The Digital Chart of the World is a comprehensive 1:1'000'000 scale vector base map of the world that was developed for the US Defense Mapping Agency. The data package contains several thematic layers of which the most interesting for this work are water and country boundaries, populated places and urban areas, roads and railroads. The DCW is a relatively old dataset, dating from the early 1990s and containing even older data in some regions[19]. For the purposes of spatial orientation and the defining of areas of interest, however, the actuality of the data is good enough, since the decisions to be taken using the DCW affect large areas and boundaries that are unlikely to have changed, even within a period of up to twenty years.

Great Britain's administrative divisions are extracted from Nomenclature of Territorial Units for Statistics (NUTS[20]) datasets. NUTS was established by the statistical office of the European Communities in order to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union[21]. NUTS data are organised in different hierarchical levels and the NUTS 3 level mostly corresponds to Great Britain's second order administrative divisions.

3.2.2.1. Queries containing spatial Operator "in"

The first areas of interest are defined for queries containing the spatial operator "in". Finding a bounding box for "in queries" turns out to be rather intuitive, since an "in" relationship implies relatively clear borders between inside and outside. Moreover, different granularities

[15] http://www.maproom.psu.edu/dcw/
[16] http://www.geonames.org/
[17] http://maps.google.com/
[18] http://earth.google.com/
[19] http://www-sul.stanford.edu/depts/gis/DCW.html#desc
[20] http://epp.eurostat.ec.europa.eu/portal/page?_pageid=2254,64099847,2254_64185160&_dad=portal&_schema=PORTAL
[21] http://ec.europa.eu/comm/eurostat/ramon/nuts/introduction_regions_en.html

do not influence the topological relationship "inside", so that the results of "in" are not affected by changing scales.

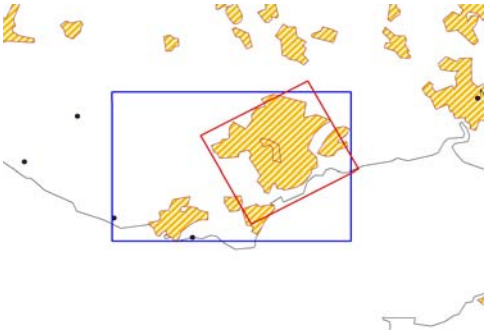
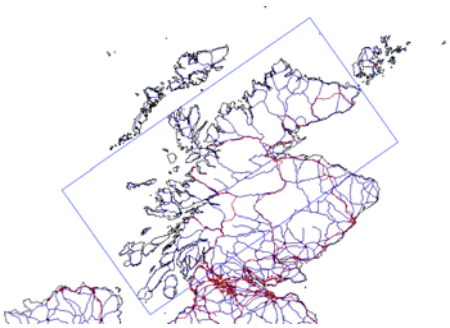
Expected areas of interest are either defined manually by using DCW background information or resolved by geometric algorithms within ArcGIS. ET Geo Wizards are used to obtain minimum bounding rectangles from DCW and NUTS shapefiles, if the concerned region is either an administrative division (Cornwall, Fife) or a country (Scotland). Firstly, a convex hull is built around the interesting feature; secondly the minimum bounding rectangle is created around the convex hull.

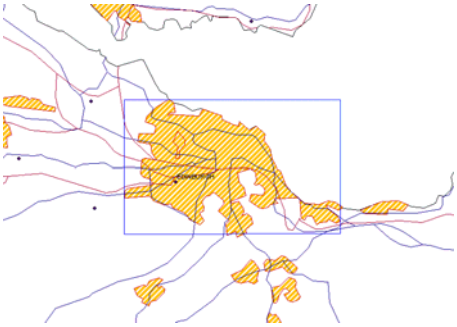
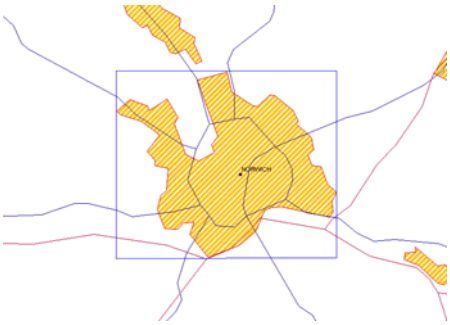
The "to convex hull to bounding rectangle" method is applied for a number of reasons. Firstly, the bounding rectangles of the SPIRIT result point patterns were computed that way. Keeping the methodology for the expected area features is more consistent than changing it. Furthermore, features of the "to convex hull to bounding rectangle" method lie tighter around the original feature and contain less useless area. This is especially interesting for costal areas, where the amount of area covered with water is smaller by using this method. Finally, the resulting bounding rectangles appear to obviously give a better representation of the interesting feature's shape than bounding rectangles aligned with x and y gridlines.

Hence, why not use convex hull directly, instead of converting it back into its bounding rectangle? The first reason is because of the minimum bounding rectangle's simplicity especially in view of the calculations to come. The second and more important reason is because of the intuitive defining of expected areas of interest for those queries that are not built with "in". As we are about to see, it is a difficult task to define an expected area of interest. Defining it with a convex hull is arbitrary and makes it even more subjective than doing it by using a possible bounding rectangle.

The following table gives a summary of how the expected areas of interest were built and which features they contain.

Museums in Cardiff United Kingdom	
Greater Cardiff area, including university. Manually extracted from DCW.	

Hotels in Cardiff United Kingdom	
Greater Cardiff area and airport. Manually extracted from DCW.	
Mountaineering in Scotland United Kingdom	
Scotland, first-order administrative division. Shape extracted from DCW, to convex hull, to bounding box.	
Oil industry in Aberdeen United Kingdom	
Greater Aberdeen area, nearby industrial estates along the coast. Manually extracted from DCW.	
Camping in Highland,Scotland, United Kingdom	
Highland, second-order administrative division. Shape extracted from NUTS2 (Highland and Islands), to convex hull, to bounding box.	
Beaches in Cornwall United Kingdom	
Cornwall, second-order administrative division. Shape extracted from NUTS3 (Cornwall and Isles of Scilly), to convex hull, to bounding box.	

Walking in Fife United Kingdom	
Fife, Clackmannanshire, second-order administrative divisions. Shape extracted from NUTS 3 to convex hull, to bounding box. Clackmannanshire and Fife is one NUTS 3 region and cannot be further separated.	
Pubs in Edinburgh United Kingdom	
Greater Edinburgh area. Manually extracted from DCW.	
Shipping in liverpool united kingdom	
Only 1 point, no area extracted.	
Schools in Norwich United Kingdom	
Greater Norwich area. Manually extracted from DCW.	

Tab. 3-3: Summary of expected areas for "in-queries“

3.2.2.2. Queries containing other spatial Operators

The remaining queries do not contain the “in” relationship. Instead, the other applicable operators “near”, “outside”, “south”, “east” and “west” are used, which complicates the definition of the expected area of interest. All results produced with one of these operators can change with an increasing or decreasing scale. Furthermore, the dimension and position of the extent fundamentally rely on the thematic question and the geographic situation as well as on local knowledge and common sense. The following expected bounding boxes are created with respect to these thematic and geographic constraints. In addition to the allowance of thematic und geographic peculiarities, metric distance and area considerations need to be included as

well. A spatial operator “near” can mean a small distance up to one or two kilometres, around the corner or two blocks down the street. At larger scales, “near” can still stand for a 30 minute car ride, a distance of 50 kilometres or even more. In this work, “near” is interpreted as being rather large. The expected areas of interest for “near” queries explicitly encompass a large amount of area around the queried place in order to underline the claim for declustered and spatially well distributed results.

The definitions of the bounding rectangles are documented in the following case studies. All queries are examined separately and further geographic sources and knowledge are used to define a reasonable expected extent. The approach consists in taking the thematic concept and asking “what is it about?”, in taking the geographic location and asking “where is it located and how is its environment?” and in taking the spatial operator and asking “how large could a reasonable extent be with respect to topic, geography and spatial operator?” These considerations about the dependence of distance on the *where*, the *what* and the spatial relation agree with those of Fu et al. (2005) who think about the interpretation of spatial relationships for spatial query expansion. In order to make the process of defining expected areas of interest more transparent, reproducible and less subjective, a number of objective rules are followed:

- Distances of bounding rectangles range from 30 to 120 kilometres.
- Possible result points should not lie in the catchment area of other places that are of a higher hierarchical level.
- Possible result points should not be separated by large water bodies. Instead, they should be easily accessible by land, without long diversions.

Climbing near Aviemore United Kingdom

Topic: Climbing, mountains, rocks.

Geography: Aviemore lies in Highland, Scotland. Climbing areas could be expected to be located in the south and the southwest of the town in the Grampian Mountains. Furthermore, the Cairngorms and the Cairngorms national Park as the peak of Ben Macdui can be found in the south area of the town[24]. Some foothills where climbing could be possible are located in the west of Aviemore and in the northeast, in the area of Grantown on Spey.

Operator and distance: High elevation and mountains do not necessarily imply climbing opportunities. Climbers need to be mobile and to travel to the appropriate locations, where maintained routes are provided. Therefore and because of the low village density in that potential interesting area, the extent is defined rather large, that about 100 x 75 kilometres.

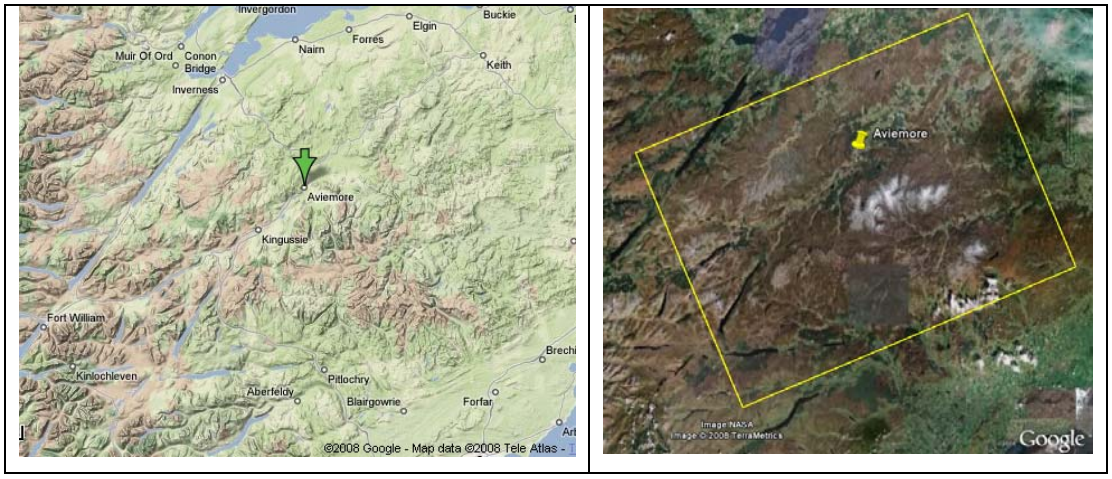


Fig. 3-1: Overview and expected extent; climbing near Aviemore

Camping near Lancaster, Lancashire, United Kingdom

Topic: Camping, nature, outside, holidays.

Geography: Lancaster lies on Britain's west coast at about 70 km north of Manchester. Possible camping areas include the forest of Bowland and Ward's Stone in the west of the town. Further away we find two National Parks where camping, in a restricted way, could be possible. The Yorkshire Dales national park is found in the east and the Lake District national park in the north and northwest, both of them between 50 to 80 kilometres from Lancaster[22]. In the south, the interesting area for this query is bounded by the two cities of Preston and Blackburn. Campsites in that area would probably not be considered "near Lancaster", but "near Preston" or "near Blackburn". Finally, potential camping areas along the coast towards the north and the south of Lancaster are also covered by this bounding rectangle.

[22]U.K. http://www.nationalparks.gov.uk/

Operator and distance: For the topic "camping", the operator "near" can be interpreted wider, since a certain mobility and willingness to travel can be expected from people who are looking for campsites. The bounding rectangle including the two national parks and excluding the urban cluster in the south measures about 95 x 70 kilometres.

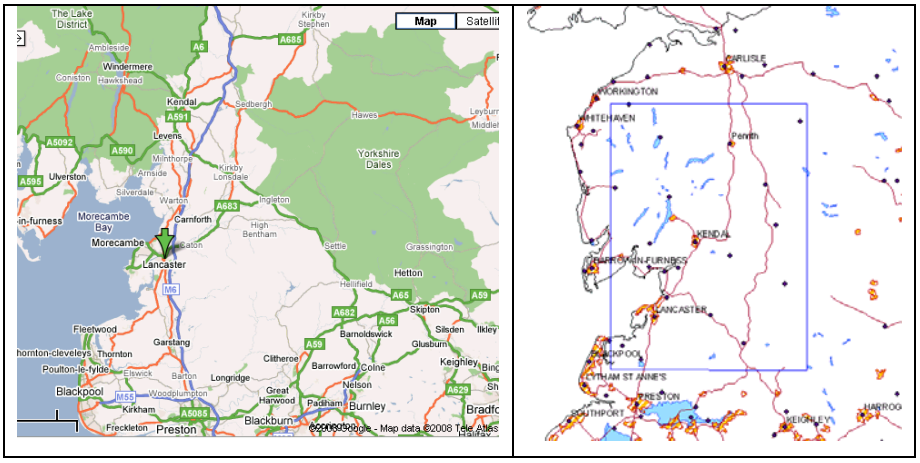


Fig. 3-2: Overview and expected extent; camping near Lancaster

Hotels near Edale United Kingdom

Topic: Hotel, tourism.

Geography: Edale lies in the middle of Peak District national park in Derbyshire, England[23]. The area is right in between the metropolitan area of Manchester in the west and Sheffield amd its surroundings in the east. In the north, the region is bounded by the urban areas of Bradford and Leeds. People looking for hotels near Edale would probably be interested in hotels within or near the national park borders rather than in hotels in Sheffield, Manchester or Leeds. This results in a narrow bounding rectangle jammed in between the three urban centres.

Operator and distance: Interesting places for hotels that lie near Edale, but not right inside the urban centres around it, are bordered by a bounding rectangle of about 40 x 50 kilometres.

[23] http://www.nationalparks.gov.uk/

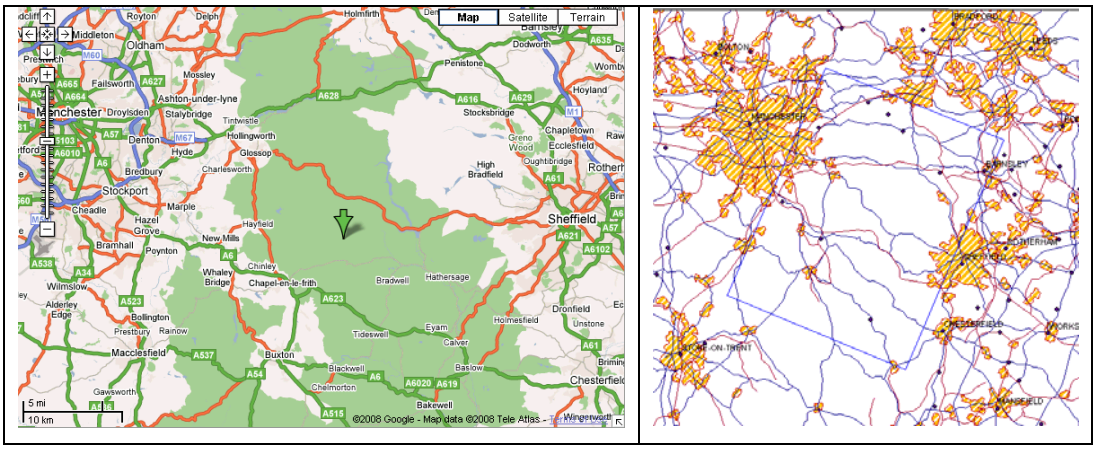


Fig. 3-3: Overview and expected extent; hotels near Edale

Walking near Beauly United Kingdom

Topic: Walking, flat, hilly, lakes, coast, countryside.

Geography: Beauly lies in Highland, Scotland, situated at about 15 kilometres northeast of Inverness. In the south of Beauly, we find the Loch Ness and its surrounding villages where a lot of walking possibilities can be expected. Other paths or trails can be reckoned with in the east of Beauly, along the Beauly Firth and on the flat "Black Isle" peninsula as well as in the north, where we find several lakes such as Loch Fannich, Loch Glascarnoch or Loch Luichart.

Operator and distance: The expected extent is bounded by the south end of Loch Ness in the south and by the Black Isle in the east. In the north and in the west, the region does not have very clear geographical borders. Still, a boundary of the expected area can be suggested in the region of Loch Glascarnoch. It appears to be the end of a relatively homogenous area, which is hilly and has numerous lakes that are interesting for walking. The defined interesting area with respect to the topic "walking" is bordered by a bounding rectangle of about 65 x 70 kilometres.

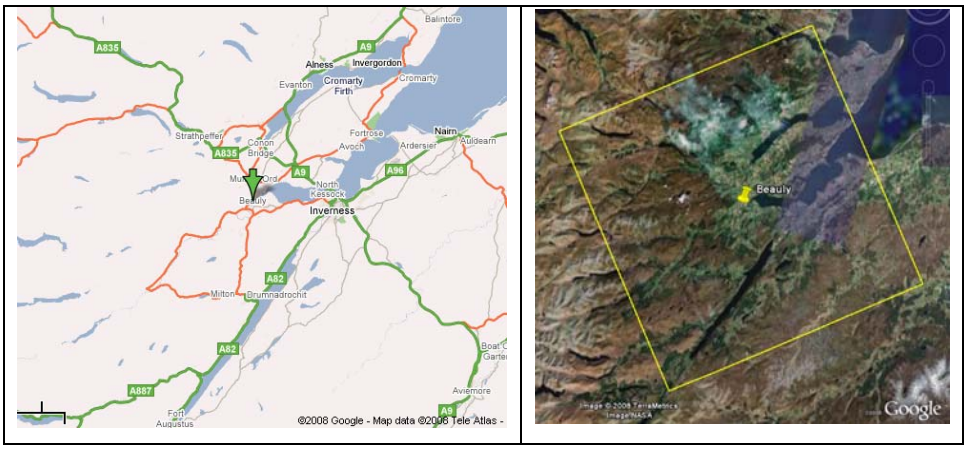


Fig. 3-4: Overview and expected extent; walking near Beauly

Canals near Stroud United Kingdom

Topic: Canals, water, flat, sea.

Geography: Stroud lies in Gloucestershire at about 35 kilometres northeast from the city of Bristol and is located at the meeting point of the so-called Five Valleys[24]. Following the elevation, water lines will converge in the area of Stroud, meaning that canals are likely to be found in the town's adjacencies. At a larger scale, canals could be expected to be located closer to the coast, namely along the river Severn and especially in the area of its estuary. The southwest border of the bounding rectangle is built by the area of Bristol, where canals are perceived "near Bristol" more than "near Stroud".

Operator and distance: This query allows a more narrow extent of interest than the previous "near" queries. The "near" in this case may include less surrounding area, since the desired topic is likely to appear right in the neighbourhood of the location. It can be enlarged, however, to the area of the river Severn's borders, covering a spatial extent of about 45 x 50 kilometres.

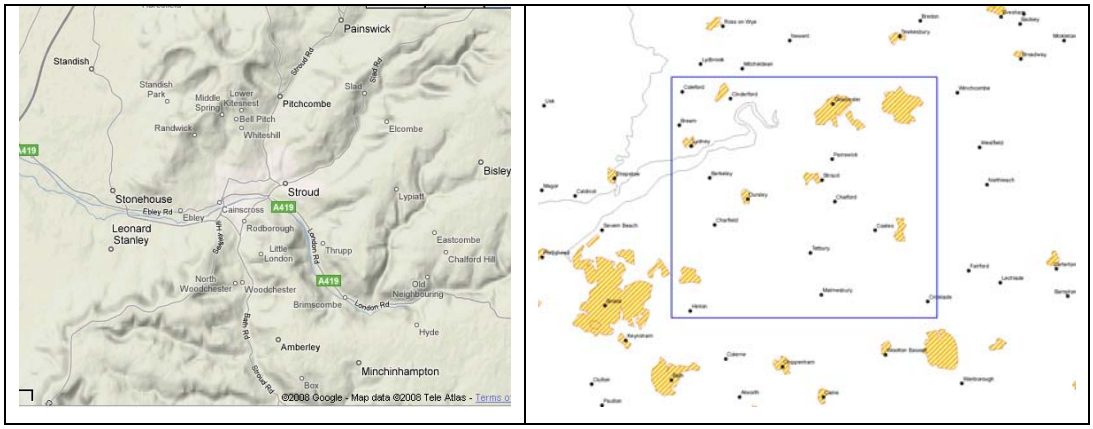


Fig. 3-5: Overview and expected extent; canals near Stroud

Red kites near Cromarty,Upper Ythan,United Kingdom

Topic: The Red Kite (Milvus milvus) is a medium large bird of prey in the Accipitridae family. It was almost extinct in the United Kingdom, except for a small breeding population in Wales. In England and Scotland, the bird has been successfully re-introduced[25]. The Red Kite is a wide-ranging species with a large habitat tolerance. It requires a large tree, which it can easily access and in which it can build a nest about 10 to 15 metres above ground[26].

Geography: We find Cromarty in Ross-Shire in the northeast end of Scotland's Black Isle. The whole Black Isle peninsula is flat and mostly covered by agricultural and park

[24] http://en.wikipedia.org/wiki/Stroud,_Gloucestershire
[25] http://en.wikipedia.org/wiki/Milvus_milvus
[26] http://www.forestry.gov.uk/forestry/redkite/

landscapes. As we leave the Black Isle, towards the west and the south, the terrain gets rougher and the vegetation more barren. The same applies for the north. A band along the coast with a milder climate appears to be covered by cultivated and park landscapes, towards the inland, terrain and vegetation get rougher. Remembering the Red Kite's habitat, the presence of the bird around Cromarty can be expected on the Black Isle as well as on the mild bands along the coasts in the north and south.

Operator and distance: Since the number and distribution of Red Kites in Scotland is limited and under continuous observation, the appearance of the bird is known very precisely. In Scotland, three places close to Cromarty are listed[27]: Blackmuir Wood near Strathpeffer, Monadh Mor near Conon Bridge and Ord Hill near North Kessock. Since wide areas around these places also accord to the Red Kite's habitat and the expected areas for "near queries" are interpreted largely, the bounding rectangle is enlarged to about 60 x 40 kilometres.

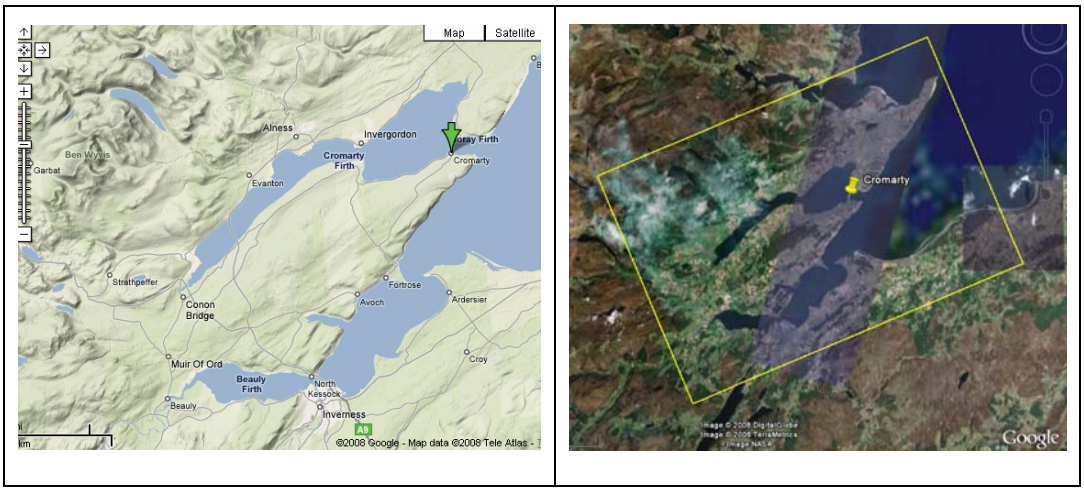


Fig. 3-6: Overview and expected extent; red kites near Cromarty

Walking outside Edinburgh United Kingdom

Topic: Walking, flat, hilly, lakes, coast, countryside.

Geography: The city of Edinburgh lies in Midlothian in the southeast of Scotland. It is situated on the south coast of the Firth of Forth, a few kilometres large estuary of the river Forth, whose water builds a geographical border in the north. Towards the south and the east, we find an open plain which is bordered by the southern uplands crossing the country from the west to the east. In the west, within a distance of 60 kilometres from Edinburgh, we find the urban area of Glasgow and its agglomeration that is connected to the Edinburgh zone by a number of middle-sized towns forming a large metropolitan area.

[27] http://www.forestry.gov.uk/forestry/redkite/

Operator and distance: A bounding box for this query includes the plane in the south and the southeast of Edinburgh as well as the area along the coast of the Firth of Forth. However, it does not include the town of Edinburgh itself. The area north of the river is not considered lying outside of Edinburgh, since the topological relationship "outside" in a geographical understanding does not imply the crossing of a water obstacle of such dimensions. The resulting rectangle has side lengths of about 35 x 55 kilometres.

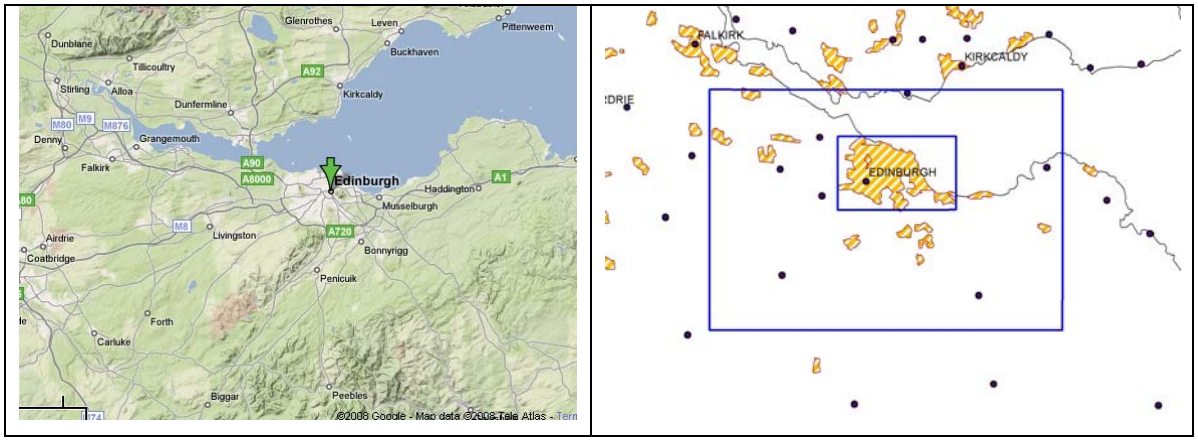


Fig. 3-7: Overview and expected extent; walking outside Edinburgh

Cycling south London United Kingdom

Topic: Cycling, bike trails, flat, hilly.

Geography: The area south of London basically includes the four counties of Surrey, West Sussex, East Sussex and Kent. In the south and east, the natural border is built by the English Channel; in the west, the "south of London" area is bordered by the two cities of Portsmouth and Southampton. This region certainly has enough importance to be seen as counterpart to the more anonymous term "south of London". Southern areas of Greater London should not be included either, since they are not lying "south of London", but still part of London.

Operator and distance: The interesting area for this query is rather predetermined by natural borders and an urban area in the west. One could argue that results are to be expected mainly in the centre of the southern London area around places like Crawley, Horsham or Tunbridge Wells rather than at the coast. People looking for cycling opportunities in and around costal towns like Brighton, Eastbourne or Hastings would probably enter a more specific query asking for those regions. The same applies for the region in the southeast around Dover, Ramsgate or Margate. The bounding rectangle including the expected area measures about 110 x 40 kilometres.

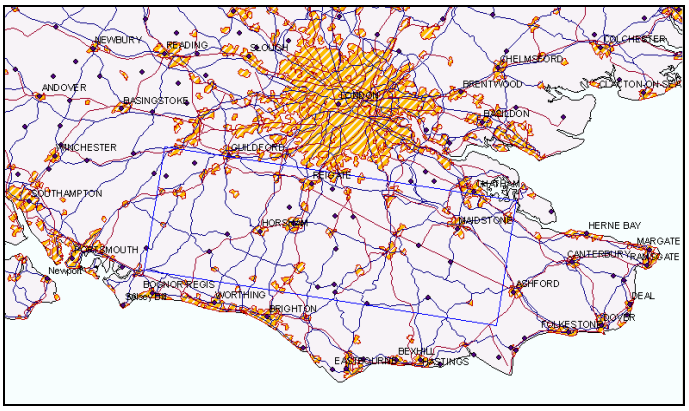


Fig. 3-8: Expected extent; cycling south London

Castles east Edinburgh United Kingdom

Topic: Castles.

Geography: The geography of Edinburgh has already been looked at two queries before. Still, the Firth of Forth constitutes a large water obstacle in the north of the city.

Operator and distance: The definition of an interesting area for the spatial operator "east" is determined by the geographical barrier formed by the Forth estuary. From a geographical point of view, points lying on the north side of the Firth of Forth would not be considered lying east of Edinburgh. The extent of interest would thus include the region east and most of all, southeast of the city. In the south, the expected area is bounded by the city of Galashiels and the English border. English castles are not perceived as lying east of Edinburgh and can therefore be excluded as an answer to this query. The bounding box enclosing the interesting area measures about 55 x 50 kilometres.

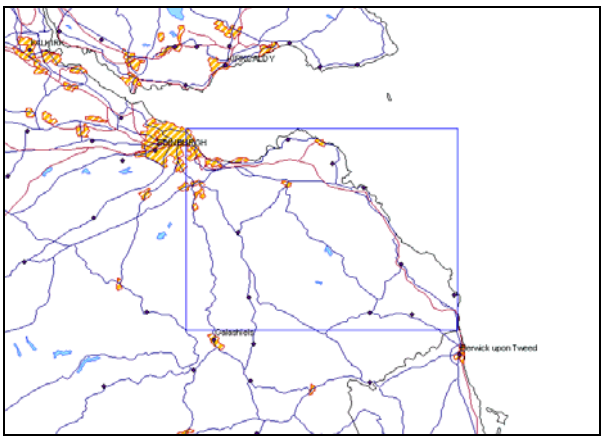


Fig. 3-9: Expected extent; castles east Edinburgh

Hotels west Fort William United Kingdom

Topic: Hotel, tourism.

Geography: Fort William lies in Inverness-Shire in the geographical centre of Scotland on the west coast. The town is situated on the shore of Loch Linnhe crossing the region diagonally from the southwest to the northeast. Being the largest town within the area and due to the proximity of several spectacular sites, Fort William can be considered the local economic and touristic centre. Since peaks like Ben Nevis and numerous lakes attract visitors, a fairly high number of hotels and similar accommodations are to be expected.

Operator and distance: The interesting area for the spatial operator "west" would probably be the peninsula cut out by Loch Linnhe right to the west and the northwest of Fort William. Furthermore, one could argue that hotels are also to be expected on the Isle of Mull and along the Loch Linnhe coast, in the south, towards the southwest. The Isle of Mull is not directly accessible by land, but the separating water body is rather narrow and the area can still be considered lying west to Fort William. In the north, the region of interest is bounded by the Isle of Skye, which constitutes a different geographical area of high weight. The appropriate bounding rectangle measures about 95 x 75 kilometres.

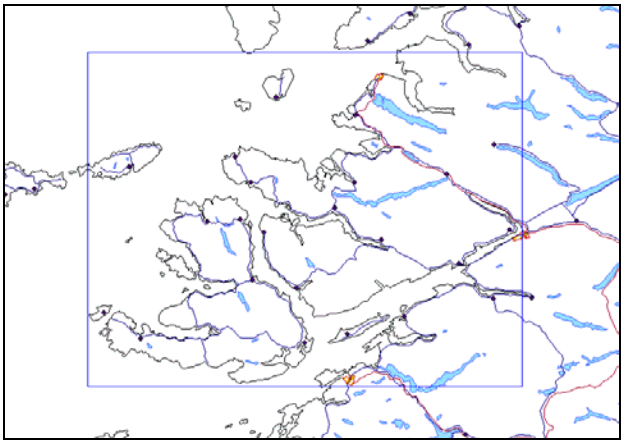


Fig. 3-10: Expected extent; hotels west Fort William

3.2.3. Evaluation of expected versus actual spatial Extents

In the last chapter, the expected spatial extents for all queries were defined in a qualitative way and with respect to local circumstances. In this section, the expected extends and the relationships between the result- and the expected extents will be evaluated in a more quantitative manner. The presented results are limited to a selection of five examples. These examples represent the variety of the results encountered in the complete result set. Tables containing the full set of results are found in the appendix. Tab. 3-4 gives a summary of some basic figures describing the expected extents.

Query	***Centroid Lambert***	***Area (km^2)***	***X-Dist (km)***	***Y-Dist (km)***
mountaineering in scotland united kingdom	-1661637 / 968224.9	241577.5	346.6	696.9
beaches in cornwall united kingdom	-1994806 / 228159.0	8048.6	127.4	63.2
walking near beauly united kingdom	-1673715 / 928869.9	4552.7	66.7	68.4
canals near stroud united kingdom	-1776418 / 300443.3	1734.5	43.5	39.9
cycling south london united kingdom	-1658948 / 189415.2	4257.5	110.5	38.7

Tab. 3-4: Summary of expected areas of interest

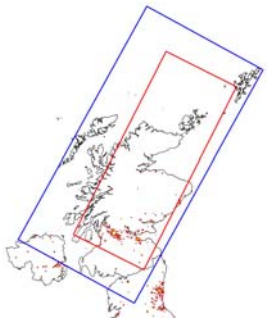
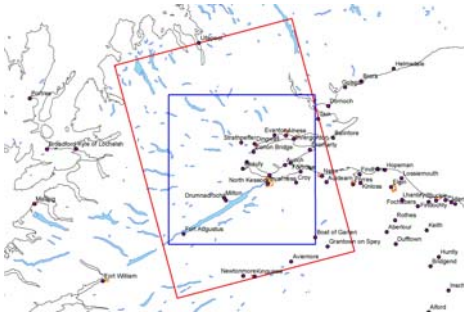
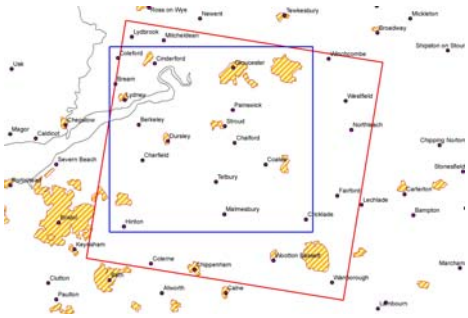
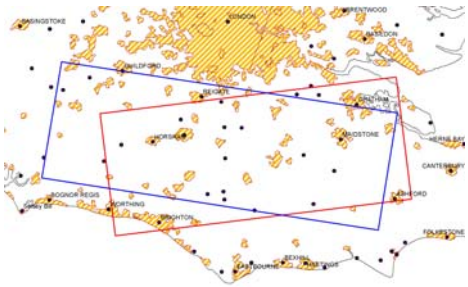
Once the expected extents have been investigated, the size and position of both, result and expected extents will be considered in Tab. 3-5. The ratio of the result area to the expected area in the second column gives an idea of the relative sizes of the two extents. The higher the percentage, the more similar the size of the two rectangles is. However, nothing is said about the position of the rectangles; in fact, they could have a high percentage, but lie completely apart from one another. The third column takes the location of the rectangles into account and gives an absolute size of the spatially intersecting area. The intersection area is normalised by the total covered area in the fourth column: The total covered area corresponds to the area that is covered by the union of the result area with the expected area. It is calculated by building the sum of the result and of the expected area and by subtracting the intersecting area.

Query	***Result Area / Expected Area (%)***	***Intersection Area (km^2)***	***Intersection Area / Total Covered Area (%)***
mountaineering in scotland united kingdom	45.77	110564.2	45.77
beaches in cornwall united kingdom	57.56	4632.7	57.56
walking near beauly united kingdom	201.93	4550.1	49.48
canals near stroud united kingdom	166.80	1712.6	58.75
cycling south london united kingdom	91.32	3061.3	60.21

Tab. 3-5: Spatial relations between expected and result areas

The expected and result extent for the query “cycling south London united kingdom” reaches the highest intersection versus total covered area ratio, meaning that 60% of the area covered by both extents is the intersection area. The smallest ratio is found for the “mountaineering in scotland” query, which has a very large expected extent, corresponding to the area of Scotland. In the next section, a closer look is taken at a number of selected queries and the spatial relations between their result and expected extents. Tab. 3-6 illustrates the amount of intersecting to non-intersecting area for both extents. It reveals how similar the two extents are in size and position, gives a comprehensible summary of the situation and also includes the topological relation between the two extents. An overview including all 20 results is found in the appendix.

The two “in query” result extents stand completely within their corresponding expected extent, which results in 100% intersecting result area for the topological relationship “inside”. For the “canals near stroud” and “walking near beauly” queries, the topological relationship is “intersect”, but close to “inside”. The fraction of non-intersecting result area is very small and usually notably higher for the remaining “other queries”. The two extents of the “cycling south London” query are similar in size and position and result in high values for the intersecting expected and result area.

Query		*Intersecting*	*Non-intersecting*	
Mountaineering in Scotland United Kingdom	***Result Area***	100 %	0 %	
	Expected Area	45.77 %	54.23 %	
	Topology	Inside		
Beaches in Cornwall United Kingdom	***Result Area***	100 %	0 %	
	Expected Area	57.56 %	42.44 %	
	Topology	Inside		
Walking near Beauly United Kingdom	***Result Area***	49.49 %	50.51 %	
	Expected Area	99.94 %	0.06 %	
	Topology	Intersect		
Canals near Stroud United Kingdom	***Result Area***	59.20 %	40.80 %	
	Expected Area	98.74 %	1.26 %	
	Topology	Intersect		
Cycling south London United Kingdom	***Result Area***	78.73 %	21.27 %	
	Expected Area	71.90 %	28.10 %	
	Topology	Intersect		

Tab. 3-6: Summary of spatial and topological relations between result and expected extent

3.2.4. Conclusions

In the last section, the results of the extent analysis for five example queries were illustrated and commented on. The results for the complete query set are found in the appendix. The description of the result extent and position by means of bounding rectangles, gives a simple but important overview about the footprint distribution's spatial dimensions. The combination with an expected extent used as benchmark is helpful to recognize and understand the geographic context and also allows for comparability between the results of different systems.

When analysing the results for all 20 queries, we see that the percentage of the result area that intersects with the expected area differs with the query type. The values for "in-queries" are significantly higher than those for other queries. Mann-Whitney U-Test for the two samples results in a U-value of 7.5 and a significance value of 0.001. Due to the high value of intersecting resulting area, a valuable conclusion posits that the result areas of "in-queries" are in fact lying inside of the expected area. The fact that the spatial relationships between actual and expected results are represented correctly is also an indication that the concept of minimum bounding rectangles works well enough and does not have to be replaced by convex hulls.

On the other hand, it is not possible to conclude how much of the expected area is covered by the result area. There is a tendency, however, towards higher values for the queries which use a different spatial operator than "in". Two massive outliers, the results for "red kites near Cromarty" and "hotels west Fort William", have a very small percentage of intersecting expected area and pull the mean value down, which avoids a significant result in the U-Test. In the first case, a mistake in toponym resolution occurred and the intersecting area is zero. In the second case, the result extent is very small, covering only 2.2% of the expected extent.

The ratio between intersection and total covered area indicates how much of the area covered by both, result and expected extent, consists of intersection area and how much is covered by either one of the extents. Low values occur either if one extent is considerably larger than the other one or if the intersection area is small due to the geographic positions and orientations of the extents. High values are generated if the result and the expected extent have similar sizes and similar geographical characteristics. The results of that ratio for the whole query set range from 0% up to 61%, which shows that a variety of constellations is possible. Furthermore, Mann-Whitney U-Test revealed no significant relationship between the intersection versus total covered area ratio and the topological relationship between the two

extents. It is not possible to say that results with the topological relationship "inside" have a bigger or smaller ratio than the results with the relationship "intersect" and vice versa.

The intersection versus total covered area ratios of the five examined examples were amongst the highest achieved in the complete query set and thus represent good results. This means that the produced footprint distributions create a result extent which answers to the requirement of a wide footprint spreading. Other examples, such as the mentioned Fort William query, were less successful and produced small result extents.

The conclusions of the entire set's extent analysis can be summarised in three points:

- Results of "in-queries" are significantly inside the expected area.
- Results of "other-queries" tend to cover a higher percentage of the expected queries than "in-queries".
- It is not possible to infer the result's intersection versus total covered area ratio from the topological relationship of its expected and result extent and vice versa.

3.3. Analysing Point Patterns and Distributions

3.3.1. Footprint Aggregation Analysis

In this section, geo-statistical and cluster analysis methods introduced in chapter 2.2 are applied in order to come to statements about pattern- and distribution type of GIR footprint sets.

The Clark and Evans aggregation index introduced in chapter 2.2.3 is based on nearest neighbour distances. The computation for the SPIRIT footprint distributions is performed in r, by means of the spatstat method *clarkevans*. No edge correction is applied, since no points are expected to lie outside the specified window. The study area is defined as the minimum bounding rectangle of the result point pattern as it was used in chapter 3.2.

Another measurement provided by the spatstat class performs a goodness of fit test and measures how well the uniform Poisson process fits the observed point pattern. This procedure does not rely on the nearest neighbour distances, but uses the classical Kolmogorov-Smirnoff test in order to assess the differences between a spatial covariate of the observed- and the predicted distribution under model conditions. The spatstat method *kstest* is called twice, once for the spatial covariate x (longitude) and y (latitude) respectively. By investigating both axes, it becomes clear how well the distributions of the x-values and the y-

values of the point's coordinates correspond to a Poisson distribution. The result of the procedure is the P-value, which is compared to the significance level 0.05. If the result is smaller, it is considered significant for the 5% significance level and the null hypothesis of IRP and CSR can be rejected.

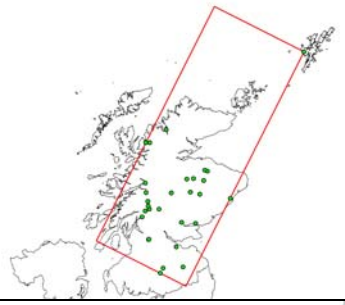
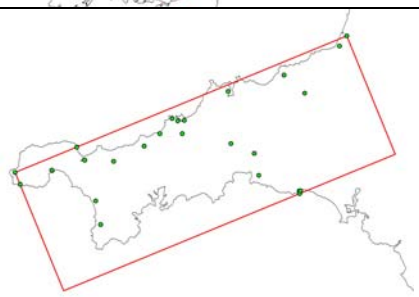
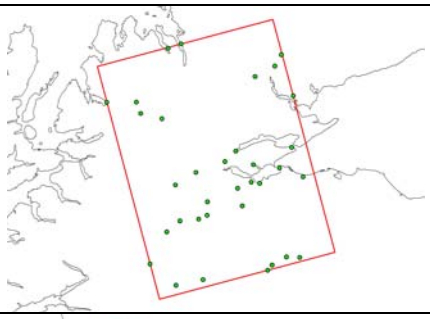
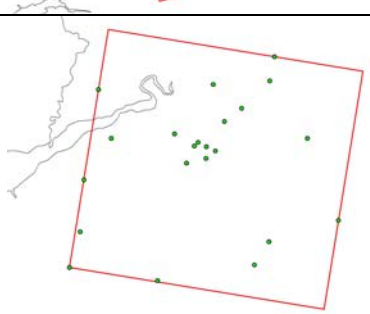
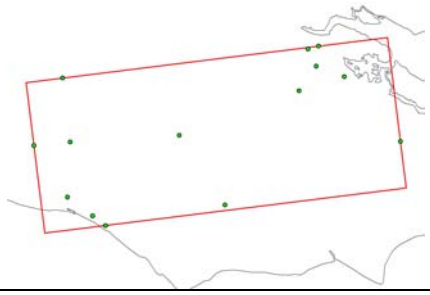
Tab. 3-8 illustrates the footprint distributions of the five example queries used in the previous chapters, surrounded by their minimum bounding rectangles. The Clark Evans aggregation index and the Kolmogorov-Smirnoff test are then computed for these point patterns. Note that all these footprint distributions have a high number of unique points and therefore represent interesting point patterns. Furthermore, we can see that the results have a fine geographical granularity. Along Cornwall's coast line or in central Scotland for instance, many toponyms of small scaled places were identified.

Tab. 3-7 summarises the results of the Clark Evans aggregation index and Kolmogorov-Smirnoff goodness of fit test for the five previous example queries.

Query	***Average Density (Points per km^2)***	***Clark Evans Agg. Index***	***K-S X***	***K-S Y***
mountaineering in scotland united kingdom	0.000452	0.48	Sig.	Sig.
beaches in cornwall united kingdom	0.011	0.35	Not Sig.	Sig.
walking near beauly united kingdom	0.00511	0.67	Not Sig.	Not Sig.
canals near stroud united kingdom	0.0104	0.89	Sig.	Sig.
cycling south london united kingdom	0.00384	1.13	Not Sig.	Not Sig.

Tab. 3-7: Summary of cluster analysis values

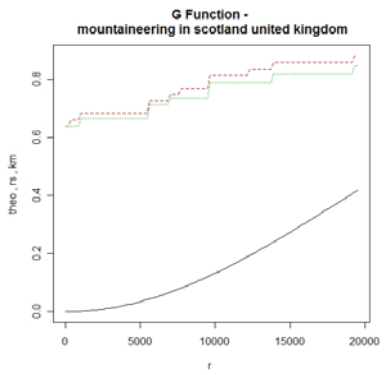
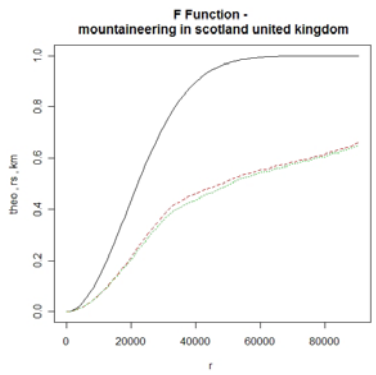
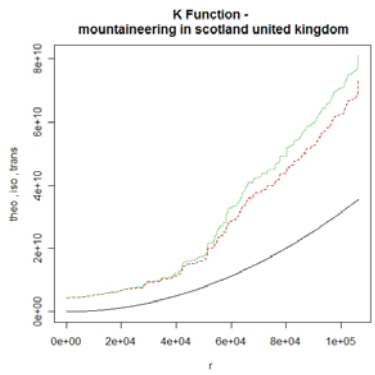
The highest point density is reached for the query "canals near stroud". The Clark Evans values for "mountaineering in scotland" and "beaches in cornwall" indicate clustering tendencies, on one hand, and a random distribution for "cycling south london", on the other hand. For "beaches in cornwall" the distribution of the Y-coordinates is not random, which can be interpreted as a concentration of footprints at the latitude of Cornwall's northern and southern coast lines.

Query	***Footprint distribution***
mountaineering in scotland united kingdom	
beaches in cornwall united kingdom	
walking near beauly united kingdom	
canals near stroud united kingdom	
cycling south london united kingdom	

Tab. 3-8: Illustration of footprint distributions

3.3.2. G, F and K Function

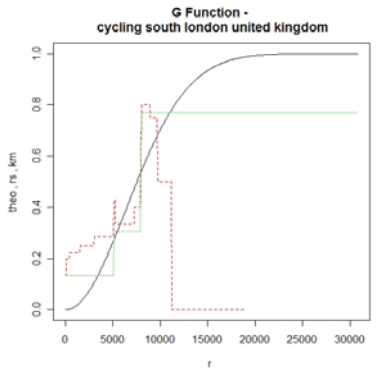
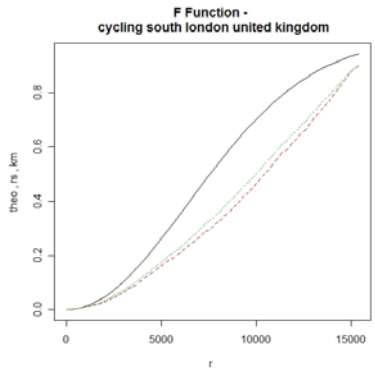
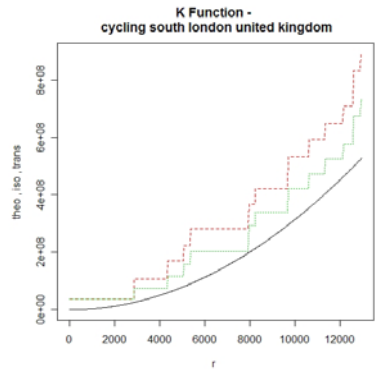
With the footprint distribution as point pattern and its minimum bounding rectangle as study area, the G, F and K functions, introduced in chapter 2.2.3, are calculated. The resulting graphs can be interpreted by using the corresponding curves of a random process as benchmark. In this chapter, the analysis is limited to the queries “mountaineering in scotland united kingdom” and “cycling near london united kingdom”, since they represent the encountered range of results, from clustered to almost random. Tab. 3-9 illustrates the three function graphs for the query “mountaineering in scotland united kingdom”. The calculations were performed in R using the spatstat functions *gest*, *fest* and *kest*. The plots contain the theoretical IRP/CSR curve in black as well as two result graphs in green and red, which represent two different edge correction methods.

mountaineering in scotland united kingdom		
G function	*F function*	*K function*
		

Tab. 3-9: Mountaineering in Scotland; G, F and K function

All three plots confirm the clustering tendencies of the result. The G function is permanently higher than the theoretical distribution, which means that too many nearest neighbours, compared to a random pattern, are found within small distances. The F function flattens after the 40000 mark, which means that only a few points are found within that range from the random points. Note that the parallel offset in the G and K function graph between the theoretical and the measured curves is due to duplicated points whose nearest neighbour distances are zero. The K function graph at low distances remains almost parallel to the theoretical values, but then increases sharply, which indicates the appearance of more points than expected by an IRP/CSR process.

Tab. 3-10 is the equivalent listing of the three graphs for the query "cycling near london united kingdom". Note that the rough behaviour of the G and K function is most probable due to the low number of points. In general, however, one notices that the tendencies towards a random pattern found in Tab. 3-7 are confirmed by the smaller distances between the measured and the expected curves.

cycling south london united kingdom		
G function	*F function*	*K function*
		

Tab. 3-10: Cycling south London; G, F and K function

Note the difference between the F functions of these two example queries. The graph for the London query remains closer to the random distribution, which means that distances from random points to result points are more constant. This confirms the close-to-one value of the Clark Evans aggregation index. The offset between the G and K function's result and expected curves is smaller than in the Scotland query, which indicates a lower number if duplicated footprints. The smaller offset and the similar gradient of the result and the expected graphs, partially in the G but mostly in the K function, are further evidence for a close to random footprint distribution.

3.3.3. Kernel Density Estimation

For the computation of kernel density estimation, introduced in chapter 2.2.2, the spatstat method *density.ppp* is used. The method uses an isotropic Gaussian kernel and the kernel bandwidth is assigned as standard deviation *sigma*. For all KDEs in this work, sigma is defined as

$$sigma = 2x\ mean + standard\ deviation$$

of the nearest neighbour distances of all points in the pattern. Depending on the number of duplicates and triplicates, various nearest neighbour distances are zero.

Another parameter that needs to be taken into account is the number of points used for KDE. According to Seaman & Powell (1996: 2084), kernel estimations from small samples will perform poorly in identifying fine structures. The sample size in their work ranged from 50 to 150 observations. As seen in Tab. 3-2, the largest sample's size in the SPIRIT result set is 50 and contains various duplicated points. The determining criterion whether KDE is performed is the number of unique points. If a result offers more than twenty unique points, the point pattern is considered large enough for KDE. Only four patterns fulfil this condition and are transformed into a density surface. Again, the study region is defined as the minimum bounding rectangle of the result point pattern.

Sigma for the query "Canals near Stroud United Kingdom" is too large for its actual minimum bounding rectangle, which causes an error in the *density.ppp* method. This problem is resolved by enlarging the extent by 11 kilometres in each direction. KDE is then performed within the distended boundaries. The obtained value matrix is converted to an ASCII grid, which is imported to ArcMap and clipped with the actual bounding rectangle. The resulting area, however, differs slightly from the actual bounding rectangle. The clipping procedure for rasters uses the minimum bounding boxes of the input rasters. Since the actual bounding rectangle for this query is inclined, the clipping procedure computes another bounding rectangle around it and uses it as clipping mask. The result of that workaround is a KDE in a window that is slightly larger than the actual resulting extent.

KDE delivers useful and easily interpretable visualisations of a pattern's point density. In order to effectively use the density data for the description of the footprint distribution's characteristics, the range of KDE values is divided into 10% steps and plotted against the area they occupy. The obtained histogram illustrated in Fig. 3-11 quantifies how much area is covered by which density values, which is useful in order to understand the clustering behaviour of a point pattern.

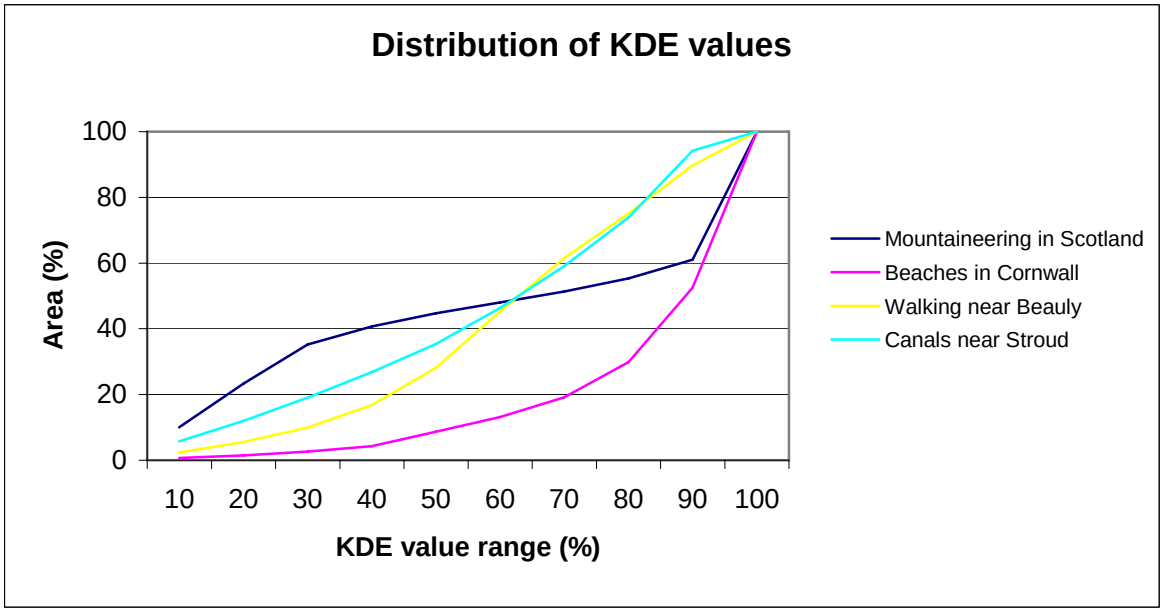


Fig. 3-11: Distribution of KDE values

The area in the Scotland query covered by high KDE-values is much larger than for the Cornwall query, which results in steeper gradient within the high value area. This can be interpreted as one large cluster that covers about 60% of the result extent. 40% of the KDE area is covered by the lowest 10% values of the KDE. This area is the counterpart to the high value cluster. The Cornwall result, on the other hand, is mostly covered by low KDE-values, which causes the steep gradient in the low value area. The cluster is defined by the sharp increasing of the graph at about 70% of the KDE value range. Only 20% of the KDE area is covered by higher values; the remaining 80% area, which contain lower KDE complement the high value clusters. The graphs of the two remaining queries "walking near beauly" and "canals near stroud" are similar and more evenly than the other two graphs. They are generally flatter, which indicates a smaller gradient of the density surface.

Finally, after the point patterns' extents and their distributions' characteristics have been investigated, all the information can be illustrated in one geographic map representation. The following plots illustrate the actual extent of the SPIRIT result point patterns and their density surfaces. The contour lines outline the KDE-value slope in 10% steps. Additionally, the blue rectangle represents the expected extent as defined in chapter 3.2.2. The information extracted from the KDE value distribution plot is found again in these maps. The partially tight spacing between the contour lines in the Scotland and the Cornwall query is visible as well as the different proportions of high versus low KDE values and the more evenly formed gradients in the Stroud and the Beauly query.

Mountaineering in Scotland United Kingdom
Resulting bounding rectangle: 206 x 537 km
Kernel bandwidth: 67608.5 m

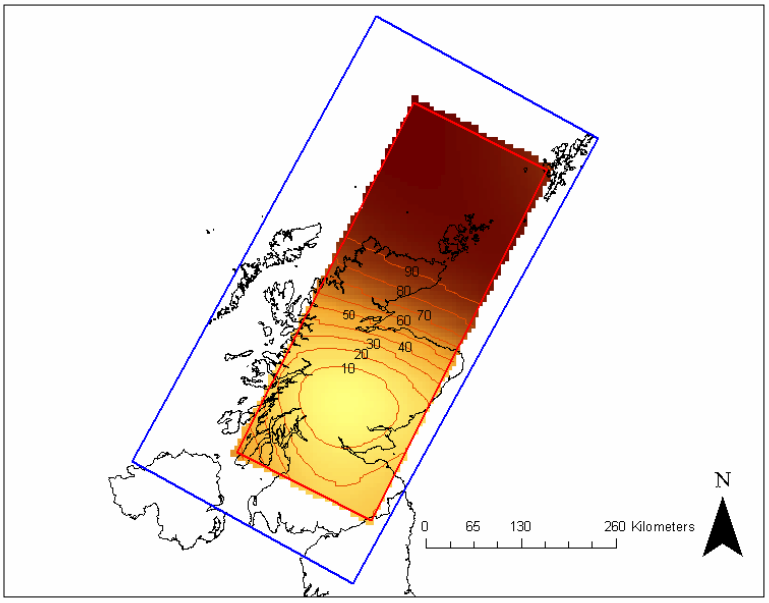


Fig. 3-12: Mountaineering in Scotland; result extent with KDE and expected extent

Beaches in Cornwall United Kingdom
Resulting bounding rectangle: 114 x 41 km
Kernel bandwidth: 6844.895 m

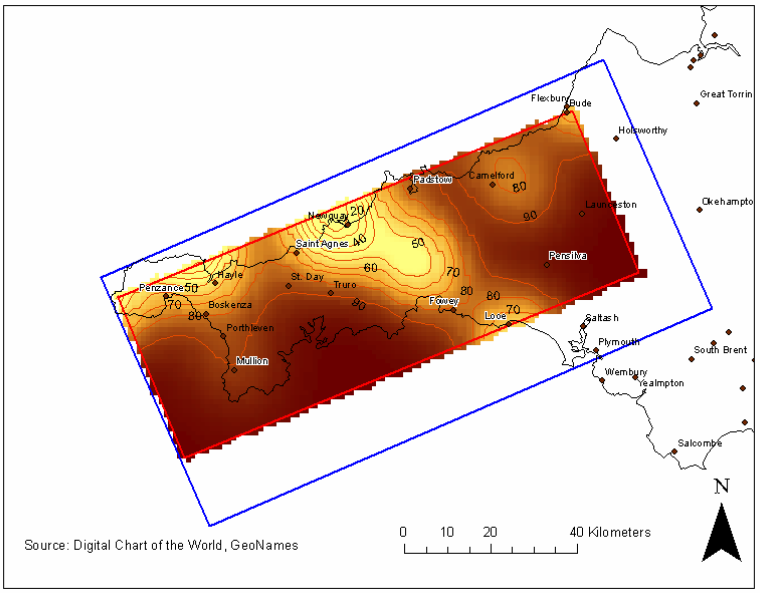


Fig. 3-13: Beaches in Cornwall; result extent with KDE and expected extent

Walking near Beauly United Kingdom
Resulting bounding rectangle: 83 x 110 km
Kernel bandwidth: 14157.97 m

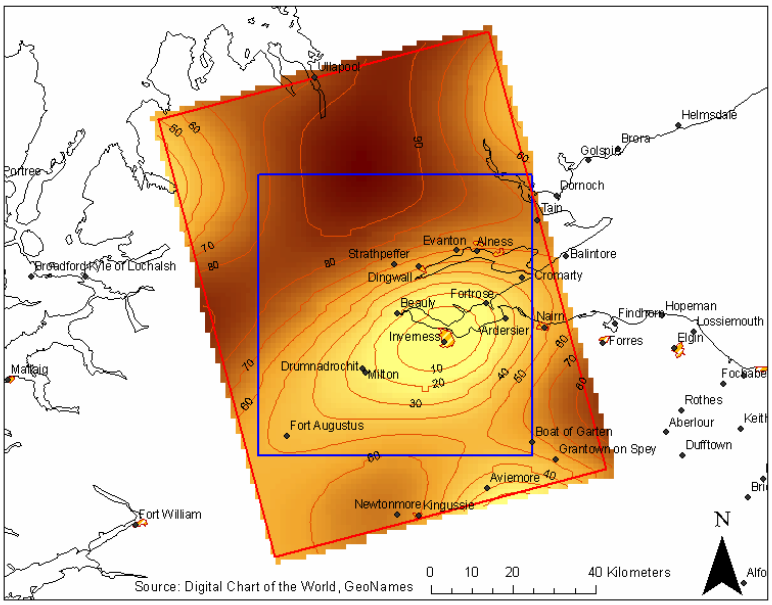


Fig. 3-14: Walking near Beauly; result extent with KDE and expected extent

Canals near Stroud United Kingdom
Resulting bounding rectangle: 56 x 52 km
Enlarged KDE-rectangle: 78 x 74 km
Kernel bandwidth: 14125.21 m

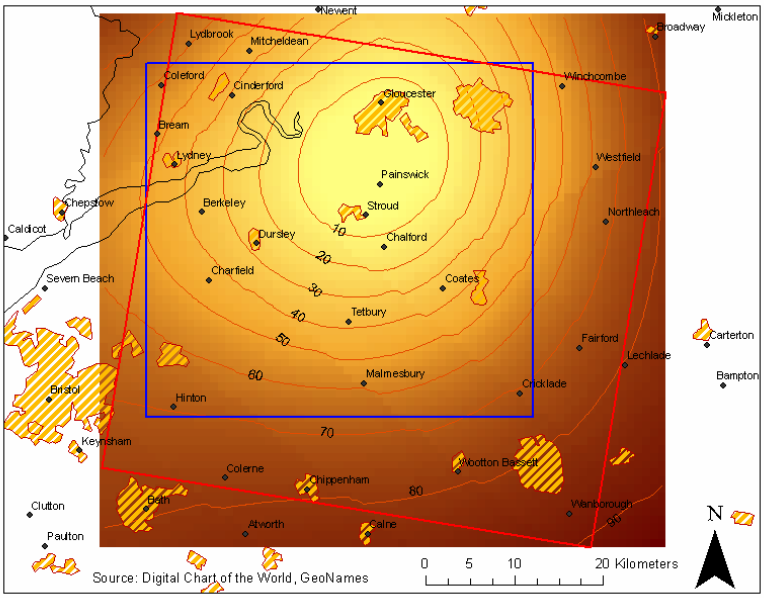


Fig. 3-15: Canals near Stroud; result extent with KDE and expected extent

3.3.4. Conclusions

In the last chapters, several measures were used in order to describe different spatial aspects of the footprint distributions. The Clark and Evans R-statistic returns one number based on the nearest neighbour distances. One needs to keep in mind that this method might be simplifying too much and ignores too much information. For a rough classification and complemented by other measures, however, it might still be a helpful figure.

Kolmogorov-Smirnoff test for CSR can be performed for x and y coordinates. Thus, it is quite useful to discover spatial dependencies and biases in x or y direction, as seen in the distribution of Beaches in Cornwall. Cluster tendencies can be located and assigned in a spatial context. However, it is not possible to make assumptions based on the outcome of one cluster measure, about the outcome of the other. Similar Clark Evans values do not imply an corresponding behaviour of the respective Kolmogorov-Smirnoff result. The Clark Evans value for the Stroud query of 0.89, for instance, suggests a distribution close to random. Kolmogorov-Smirnoff, however, is significant in both directions. Since they rely on similar properties of the point pattern, a better correspondence is found between the Clark and Evans aggregation index and the G, F and K functions.

The KDE-value versus area histogram might be a more meaningful representation of the distribution's characteristics than the Clark and Evans R-value and Kolmogorov-Smirnoff test. Clusters can be suggested where the graph has a steep gradient. It is an indication for a strong accession of area, which defines the higher KDE value area as a cluster. The KDE value range included by the cluster differs with the query, but can be revealed by using this methodology.

The Scotland and Cornwall queries have a similarly low Clark Evans value which indicates clustering tendencies. The difference between the two results is the area covered by high KDE values and the moment the KDE area versus value range graph increases. This can bee seen in Fig. 3-11 in the two graphs of the Scotland and the Cornwall query and their corresponding maps, Fig. 3-12 and Fig. 3-13. The weight of the cluster is also represented by the spacing between the percent-value contour lines. The smaller the spacing is, the heavier the concentration and the steeper the gradient of the density surface. We saw that contour lines lie close together in the Scotland and the Cornwall results, which confirms the clustering tendencies. In the Beauly and the Stroud result on the other hand, a more even spacing between the contour lines is encountered due to the flatter density surface.

The combination of extent analysis and KDE value versus area graph delivers the most promising evaluation possibilities and will be used in the next chapters in order to analyse GeoCLEF results. The results of the extent intersection and clustering behaviour are kept in percent, which makes them easily interpretable and allows a quick reception and more importantly, a good comparability of the important figures.

By relating the result extent to an expected extent, the dimensions of the result extent become intuitively comprehensible. By quantifying the result extent's density surface, one can assess what amount of area is covered by what density values. If half of the area is covered by the lowest ten percent of the value range, it is clear that most events are clustered in one site and that the result extent is enlarged by outliers. This combination of information is considered more useful and more easily interpretable than graphs and figures concerned with point distances statistics. Moreover, the statistical proof of a cluster's existence, as introduced in chapter 2.2.4, is not a necessary requirement for a useful evaluation method. Instead, it should provide an intuitive and complete description of the patterns' spatial characteristics. This position is supported by Openshaw et al. (1991: 392), who comment that "*There is an argument, therefore, to abandon the traditional geographical applications of statistical inference in favour of a more descriptive approach in which significance tests are used mainly as a results filtering mechanism*".

This methodology includes a number of weaknesses. As soon as KDE comes into play, the problem of bandwidth definition arises. In this chapter, it was done by considering nearest neighbour distances. These as well as all other methods that were used, are strongly affected by the appearance and the number of duplicated points. Finally, due to the number of unique footprints, only four queries could be evaluated by using KDE, which is a serious limitation in the evaluation of the methodology.

In general, the SPIRIT data have a fine granularity. Many footprints are found in areas with low population density and in places with few inhabitants. At the granularity stage below the city level, however, in queries such as "Pubs in Edinburgh" or "Museums in Cardiff", not more than four unique footprints were found, which makes them uninteresting for spatial evaluation. For those queries, where a reasonable number of footprints are found, the declustering and spreading algorithms used for the relevance ranking appear to work well. Even in patterns with stronger clustering tendencies, many footprints are found outside the cluster areas and the result extents still represent a reasonable fraction of the expected extents.

4. Evaluating GeoCLEF Data

4.1. Data

In chapter 3, a number of point pattern analysis methods were tested and evaluated in a GIR environment, by using data produced by the SPIRIT system. The idea of this chapter is to test the applicability of the proposed methodology and to embed it into a possible process of relevance judgement and result evaluation. Therefore, a different set of GIR data and results are needed. GeoCLEF, which is a complete evaluation campaign, is the solution of choice in order to get a different set of GIR data. The application of a similar methodology to the GeoCLEF data as before to the SPIRIT data allows for a partial comparison of the two datasets' characteristics and further statements about the different interpretations of spatial awareness. A reflection on the differences in "spatialness" of the two systems is given in the discussion chapter.

The new set of data is provided by the Imperial College London (ICL) group of Simon Overell. The delivered data includes the results of the ICL group at GeoCLEF 2005, 2006 and 2007 as well as the complete query set and the qrels file for these years. This group has concentrated on the field of toponym disambiguation and introduced a geographic co-occurrence model based on Wikipedia[28] articles (Overell & Rüger, 2007). Geographical coordinates are assigned after mapping identified place names to the Getty Thesaurus of Geographic Names (TGN[29]).

The delivered data contains a variety of runs, processed through the use of different disambiguation methods. Due to the complicated reconstruction of the result raw data to single relevant footprints per document, the results of only one disambiguation method and one year were processed. All the following results are from GeoCLEF 2005 and are based on the "most referred to" disambiguation method.

4.2. Extended ranking-based spatial Evaluation

A number of changes are encountered when the proposed methodology is applied to the GeoCLEF data. More data is available in more detail and access to the entire ranking as well as to the binary relevance judgement in form of the qrels file is possible. These circumstances

[28] http://www.wikipedia.org/
[29] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

enable a method which combines the proposed spatial measures with standard IR benchmarks, called extended ranking-based spatial evaluation. This method is a combination of the single document relevance judgement and the result ranking evaluation. The spatial measures proposed in chapter 3 are inherited and used for the spatial evaluation of the result ranking. The document relevance judgement is kept binary, based on the information in the qrels file.

As proposed in chapter 3, the spatial measures are divided into extent comparison, and point density and distribution analysis. The use of an expected extent as well as the KDE area versus value range graph are adopted in the new methodology. However, a number of changes in the application mode occur between the SPIRIT and the GeoCLEF experiments. In the GeoCLEF data experiments, outcomes are normalised with the ranking positions, which means that extent and density measurements are performed at different ranking position levels. KDE for the analysis of the clustering behaviour is performed only once, at the highest considered ranking position level, since its significance increases with the number of points and area. Furthermore, the outputs are presented in a condensed way and all figures are merged into one single plot. An additional graphical or map output is still possible.

For the GeoCLEF data, the binary document relevance judgement is given in form of the qrels file. For the SPIRIT data, all used document footprints (ranks 1 to 10) were assumed to be relevant. The idea for future application, however, is to define an expected extent first (3.2.2) and then to use it for the spatial binary document relevance judgement. Documents containing footprints inside the expected extent would be judged as spatially relevant.

In this present work, the term *extended ranking-based spatial evaluation* is proposed as a name for this new methodology. *Extended* because it includes the binary document relevance judgement and the spatial evaluation of the result ranking. *Ranking-based* because not single documents are evaluated, but the first *n* positions of the entire ranking. Finally, *spatial* because an emphasis is put on measuring spatial properties of the ranking. A graphical overview of the method's principles is given in Fig. 4-1.

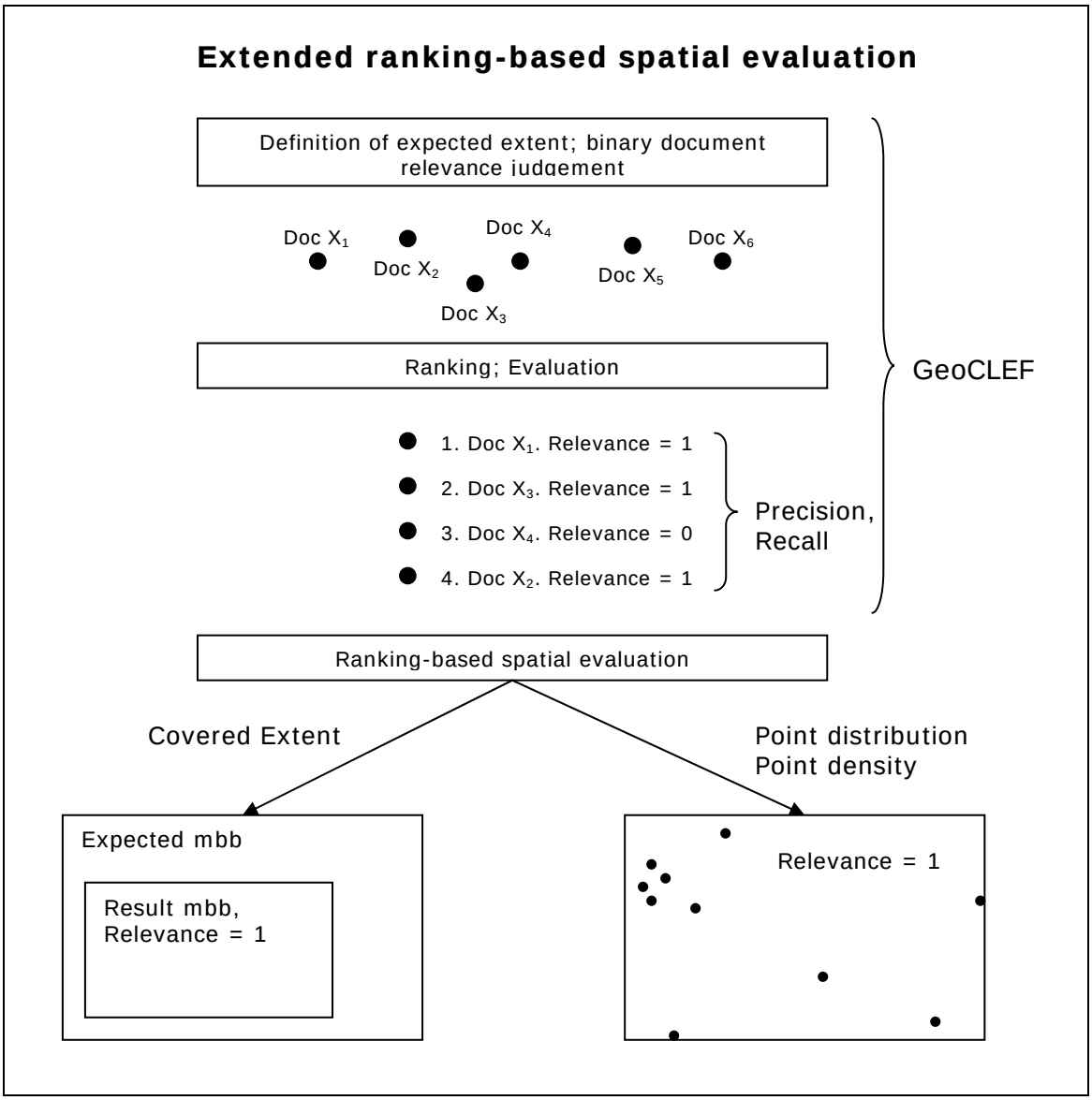


Fig. 4-1: Basic functionality of extended ranking-based spatial evaluation

4.3. Applying extended ranking-based spatial Evaluation

For the GeoCLEF data experiments, the first 100 ranks are used. Precision and recall are calculated by using all ranked documents that were judged relevant in the qrels file. For all spatial calculations, footprints that are equal to the query location or of equal granularity need to be removed, since they are not interpretable spatially and geometrically. Given a query "Trade unions in Europe", documents containing "Europe" are considered relevant for precision and recall calculations, but are removed for extent comparison and for KDE. Occurrences of footprints equal to the query term might be treated as points, by using centroid coordinates. In most cases, this would not influence the extent comparison result, but have a strongly negative impact on point distribution and on KDE, especially if there are numerous

instances of the query location in the documents. A cluster would be created at a location, where no footprints would have been assigned. As a consequence of this, two different input files are needed: The first one contains all relevant footprints; the second one is adjusted and contains footprints of lower granularity than the query region only.

The result extent and its point density within are computed ten times per query, within the first 100 ranked documents, at each tenth position (D_{10}, …, D_{100}). According to that, precision and recall are measured at the same ranking levels, by using the precision/recall at n concept introduced in chapter 2.1.1.5. The KDE area versus value range graph is performed at the D_{100} level only, since its results are not reducible to one number and thus not directly applicable for the precision/recall at n concept.

The definition of the result and expected extents is kept almost the same as in the SPIRIT data analysis. One difference is that bounding rectangles are processed using ET Geo Wizards' function *features to envelope* that produces rectangular, non inclined features. The underlying idea is to illustrate the growth of the resulting extent due the inclusion of new footprints, as the respected ranking level increases. Therefore, the orientation of the rectangles should stay the same and a larger extent's bounding rectangle should include a smaller extent's bounding rectangle entirely. These requirements are not fulfilled by the *features to bounding rectangle* function. Since this method produces features that are aligned with the longest side of the original feature, similar footprint distributions produce differently aligned and orientated bounding rectangles, if a footprint lying outside the current ranking level's bounding rectangle is included and creates a new longest side.

Finally, a last difference between the two experiments is that the DCW and NUTS layers background data are replaced by the ESRI[30] Data & Maps 2006 data series. A graphical overview of the workflow as it was applied to the GeoCLEF results is given in the following figure.

[30] http://www.esri.com/

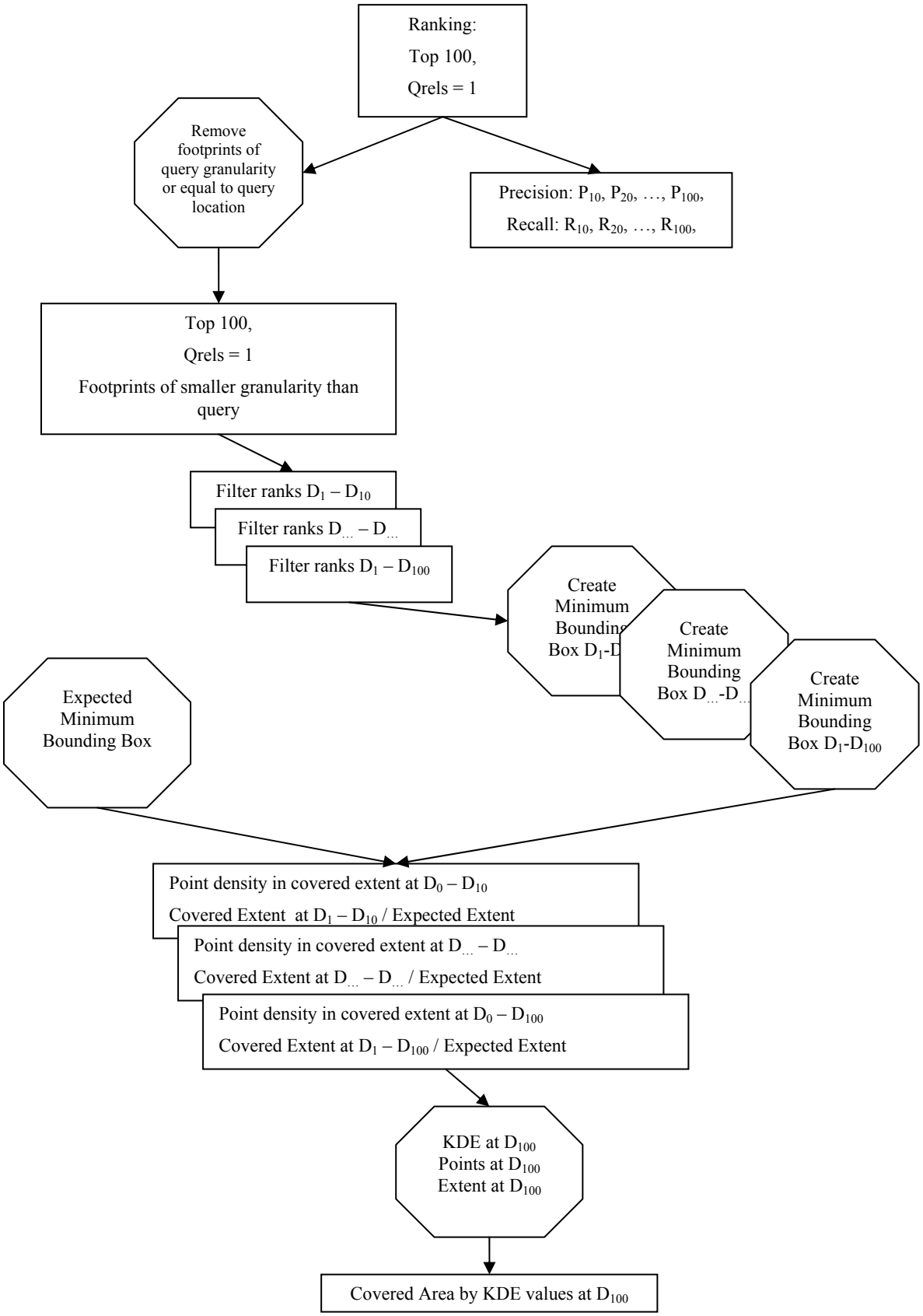


Fig. 4-2: Flowchart of extended ranking-based spatial evaluation

4.3.1. Composed Queries

A difference between the SPIRIT and the GeoCLEF queries are composed queries such as "Shark Attacks off Australia and California" or "Actions against the fur industry in Europe and the U.S.A", which occur in the GeoCLEF search tasks. The proposed methodology needs to be adapted to this kind of query, but is basically able to handle it.

Instead of building one expected minimum bounding box, two (or more) expected minimum bounding boxes need to be built with respect to query regions. A minimum bounding box covering Australia and California, for instance, would cover half the world and make any further calculations useless. The two (or more) regions are then treated like two different queries and extent as well as distribution properties are calculated separately. In order to merge the two sets of figures to the usual output form, the single regions are interpreted as equally important and results are calculated by using the mean values of all single region calculations.

4.3.2. Projections

The Lambert Conformal Conic projection that was used for analysing SPIRIT footprint distributions in the United Kingdom is not applicable for different continents. For the extended ranking-based spatial evaluation with GeoCLEF data, the Europe Lambert Conformal Conic projection is kept for all topics with a European query region. For query regions in different continents, appropriate Lambert Conformal Conic projections are applied, which are listed in Tab. 4-1. These projections and metric coordinates are necessary for this evaluation methodology, since the calculation of the kernel bandwidth is based on nearest neighbour distances.

Continent	***Projection***
Europe	Europe_Lambert_Conformal_Conic_spirit_1984
Australia	GDA_1994_Geoscience_Australia_Lambert
North America	North_America_Lambert_Conformal_Conic

Tab. 4-1: Used map projections for Europe, North America and Australia

5. Results and Interpretation

5.1. GeoCLEF Data Basis

The first part of this result chapter is dedicated to the evaluation of the GeoCLEF data basis. The next two chapters contain the results for three ICL GeoCLEF 2005 result sets, which are evaluated by using the methodology described in chapter 4.

Before the new methodology is applied to the ICL GeoCLEF data, the raw data is processed and basic properties, such as the number of relevant documents or the number of relevant footprints, are presented in the following statistics. This is an important step, because many of the limitations for the spatial evaluation can already be identified in the properties of the queries and the data basis. Fig. 5-1 summarises the number of humanly judged relevant documents per query, according to the qrels file, for GeoCLEF 2005, 2006 and 2007.

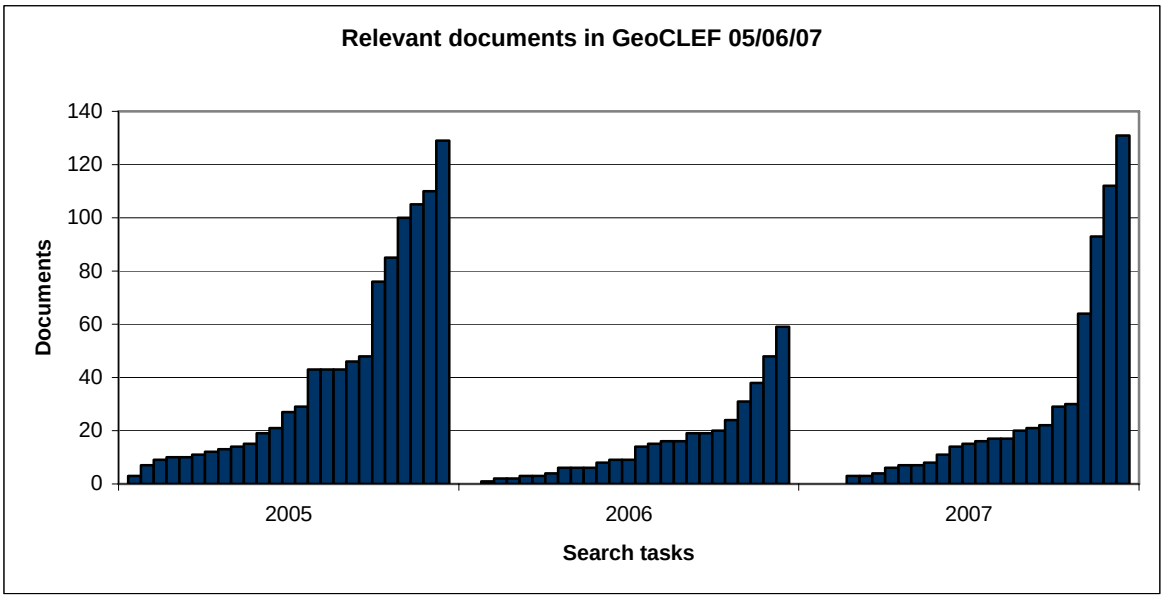


Fig. 5-1: Number of relevant documents in GeoCLEF 05/06/07

In all three years, for about half of the queries, less than 20 documents were considered relevant. In 2006, the highest number of relevant documents for a query was smaller than 60; for the other two years, maximum values were more than twice as much. The most relevant documents in 2005 were found for the topic "American Troops in Sarajevo, Bosnia-Herzegovina", for "Combats and embargo in the northern part of Iraq" in 2006, and for "Rivers with floods" in 2007. A reason for the high number of relevant documents might be that the combination of the thematic and the spatial concept in these queries does not narrow

the number of relevant articles down; instead, their occurrence in a newspaper article is highly correlated. Articles about floods are very likely to contain some information about a river; articles about Iraq refer in most cases to combats. Articles about Bosnia-Herzegovina dating from 1994 and 1995 probably report on NATO or U.S. troops. In all these examples, one part of the query hardly adds any further information to it and a text search for “floods”, “Iraq” or “Sarajevo” is likely to deliver an equal result as the full query.

After having looked at the number of relevant documents, it is interesting to look at the number of footprints and most of all at the number of unique footprints that were found within the relevant documents. This number obviously differs between participant groups, since the toponym detection and the toponym resolution will produce different results depending on the underlying system. The unique constraint for footprints is important for spatial measurements, since it provides a first indication of the spatial footprint distribution. Fig. 5-2 gives an overview of the number of unique footprints found by the ICL group for all queries at GeoCLEF 2005. The lack of results for the 2006 and 2007 queries is due to the complicated processing of the raw data.

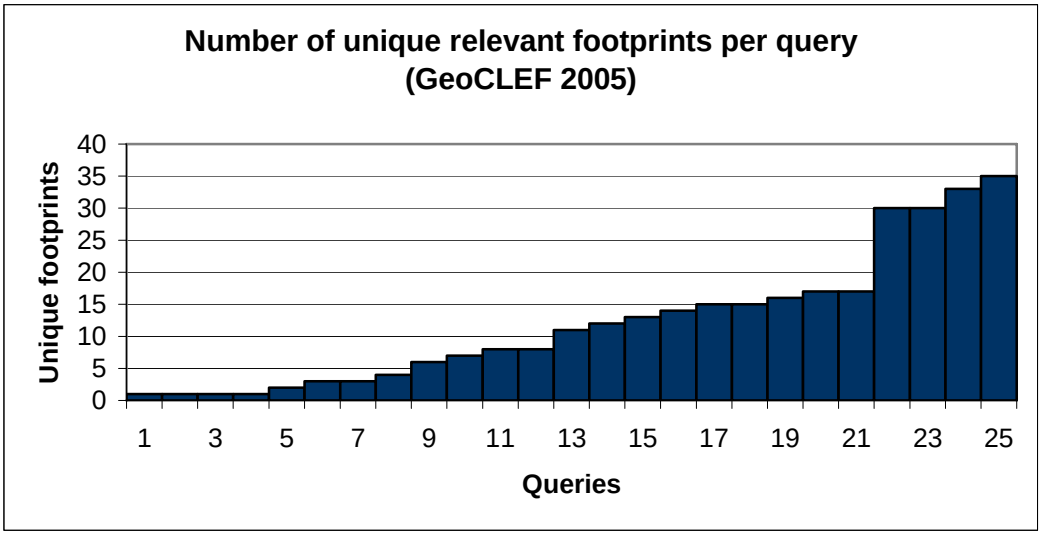


Fig. 5-2: Unique relevant footprints per query in GeoCLEF 2005

Compared to the number of relevant documents and, most of all, to the number of duplicated footprints, the number of unique footprints for many queries is very low. Four queries achieve not more than one single unique relevant footprint. For these, most spatial measures are meaningless. Moreover, not more than four queries reach a number of unique relevant footprints higher than 20, which was considered the minimum number of points for KDE in chapter 3.3.3.

A further criterion whether a footprint is usable for the extended ranking-based spatial evaluation or not, is the inequality of the footprint to the query region. As seen in chapter 4.3, footprints equal to the query region or of equal granularity need to be removed for spatial analysis, since they produce incorrect results in terms of extent and distribution. Fig. 5-3 illustrates the number of relevant footprints before and after removing footprints equal to the query region or of equal granularity to query region, for three ICL GeoCLEF 2005 topics. Furthermore, it reveals the proportion of all relevant versus unique relevant footprints for these three topics.

In the first example, half of the relevant footprints are assigned to "Scotland". These are removed for further spatial measurements, since the query is "Walking holidays in Scotland". In the second example, about one third of the relevant footprints refers to "Europe" and is therefore removed. In the third example, only a few footprints refer to the two query regions, Europe and the United States.

Since each footprint occurs only once, the difference between before and after removing for the unique relevant footprints is always one. More interesting is the fraction of unique versus all relevant footprints, which shows the tremendous amount of duplicated footprints that is found in the newspaper article collection. Note that for the query "Walking holidays in Scotland", where the *Glasgow Harold* is located, the initial set of footprints is almost 300. After the two filtering steps, however, even less footprints are available than for the topic "Trade Unions in Europe", with an initial footprint set of about 150. The disambiguation method might as well be a reason for the strong concentration of footprints. By using the "most referred to" approach, ambiguous toponyms are always assigned to the same instance.

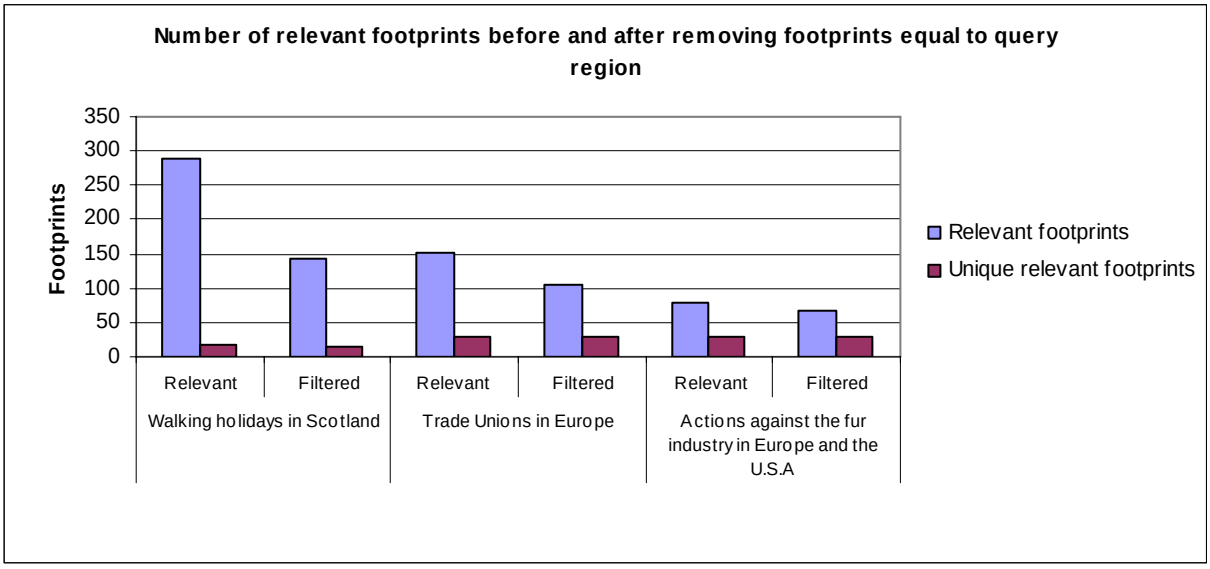


Fig. 5-3: Number of relevant footprints after hierarchical filtering

5.2. Graph Visualisations

In this section, the ICL GeoCLEF result sets for the queries "Trade unions in Europe", "Walking holidays in Scotland" and "Actions against the fur industry in Europe and the U.S.A" are evaluated, by using the extended ranking-based spatial evaluation. The results are presented in graph form.

The x-axis represents the ranking position D_1 - D_{100} divided in classes of ten ranks. At the same time, it stands for the percent groups of the KDE value range, according to Fig. 3-11. Thus, the 10 on the x-axis stands for the ranking class D_1 - D_{10} as well as for the highest 10% of the KDE value range.

Fig. 5-4 illustrates the result for the query "Trade unions in Europe". Precision reaches its maximum of 40% at D_{10} then decreases to 20% at D_{30} and remains at this level until D_{100}. The highest recall level is reached at D_{100}, which means that about 20% percent of all relevant documents were found and are ranked within the first 100 ranking positions.

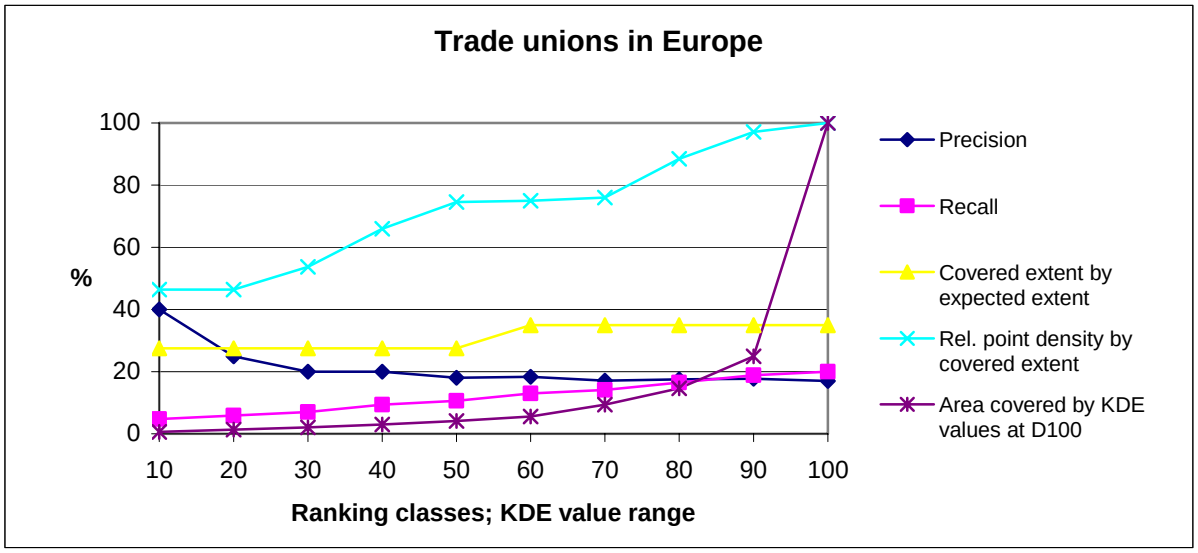


Fig. 5-4: Trade unions in Europe; extended ranking based relevance judgement

The proportion of covered versus expected extent is rather stable and shows only one small rise from D_{50} to D_{60}. This means that no point enlarging the minimum bounding rectangle of the result footprint pattern is found between D_{10} and D_{50}. The relative point density, however, increases in the mean time, which means that additional points were found inside the current result extent. As the covered extent is enlarged by including footprints lying outside of the current covered extent, the relative point density remains stable and finally reaches its maximum at D_{100}.

The last graph in the above diagram corresponds to the KDE area versus value range measurement according to Fig. 3-11. We see an abrupt increase between 90% and 100% of the KDE value range, which indicates that about 70% of the total area is covered by the lowest 10% of the KDE value range. This means that a cluster including the range values up to 90% can be expected. The gradient of the graph before the 90% value range mark is steadily rising and reaches about 30% of the total area that are included by the cluster. The constant rising of the graph up to the 90% value range mark means that there is a visible spacing between the contour lines and that the slope of the density surface can be expected to be reasonably smooth.

The result plot given in Fig. 5-5 was produced for the GeoCLEF 2005 topic "Walking holidays in Scotland". Precision is very similar to the last example; recall, however, reaches almost 50%, which is 30% higher than for the last query. Again, the covered extent is enlarged only at one ranking level, but the increase between the two extents is larger than in Fig. 5-4. The relative point density increases quickly between D_{20} and D_{30}. A difference to the last example is that the relative point density increases at the same time as the covered extent increases from 20% to 60%. The slope of the KDE area versus value graph increases sharply from D_{90} to D_{100}, which indicates a distinct gradient and a tight spacing between the KDE value range contour lines.

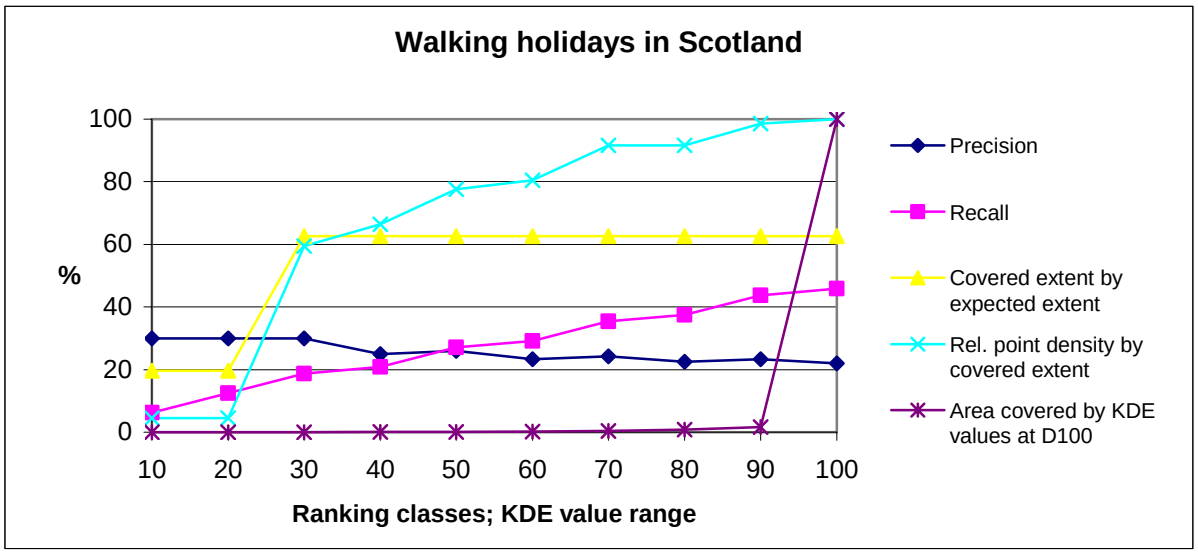


Fig. 5-5: Walking holidays in Scotland; extended ranking based relevance judgement

The last example, "Actions against the fur industry in Europe and the U.S.A" is a composed query as introduced in chapter 4.3.1. The final result represents the mean values of the calculations that were made separately for both query regions. The result plot is given in Fig. 5-6.

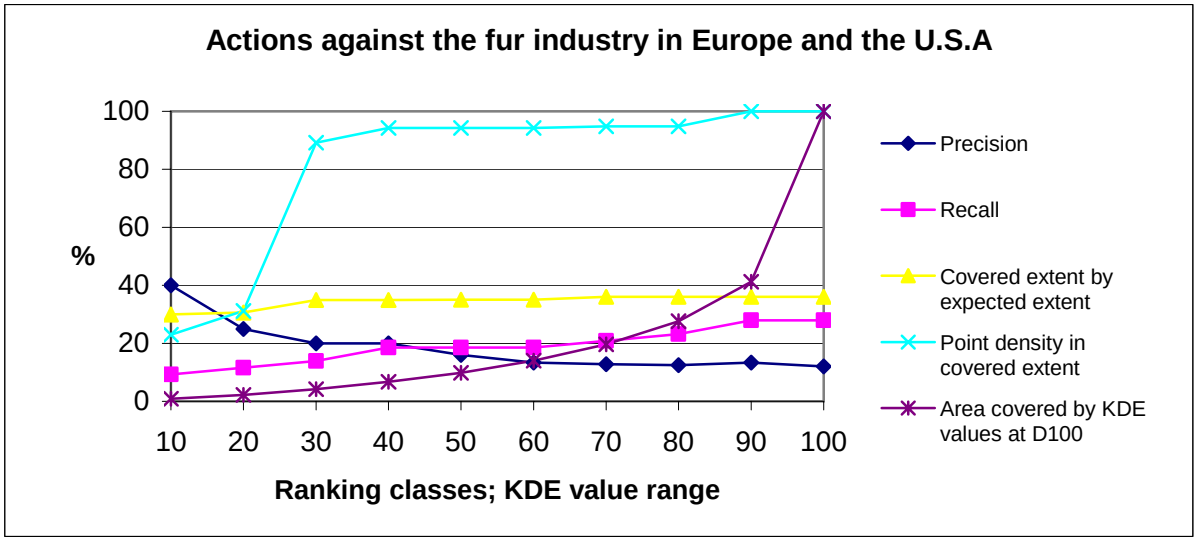


Fig. 5-6: Actions against the fur industry in Europe and the U.S.A; extended ranking based relevance judgement

Precision and recall are similar to the previous examples and the covered extent remains nearly constant, which is mostly due the covered extent in the United States which hardly changes at all. The coverage in general, however, appears to be rather poor; at the D_{100} level, less than 40% of the expected extent is covered. The point density shows a steep increase from D_{20} to $D_{30,}$ which means that documents containing many relevant footprints were ranked within that range. Corresponding to this footprint density increase, the covered extent is enlarged almost to its final level. The KDE area versus value range graph is more similar to the first example, suggesting weaker clustering than in Fig. 5-5. But still, more than half of the result area is covered by the lowest 10% of the KDE values. For composed queries, the map visualisations presented in the next chapter are especially interesting, since they illustrate both results and reveal the differences between the two query regions.

5.3. Map Visualisations

A different output form to complement the extended ranking-based spatial evaluation results are geographic map representations. This representation is one with which most people are familiar and is helpful to extract important properties of the ranking in an intuitive way. Moreover, geographical maps automatically put an emphasis on the results' spatial component. Fig. 5-7 is the map illustration of the topic "Trade unions in Europe". According to chapter 3.3.3, the blue rectangle represents the expected extent. The red rectangle represents the result extent at D_{100} and includes the KDE density surface contoured by its 10% value range contour lines.

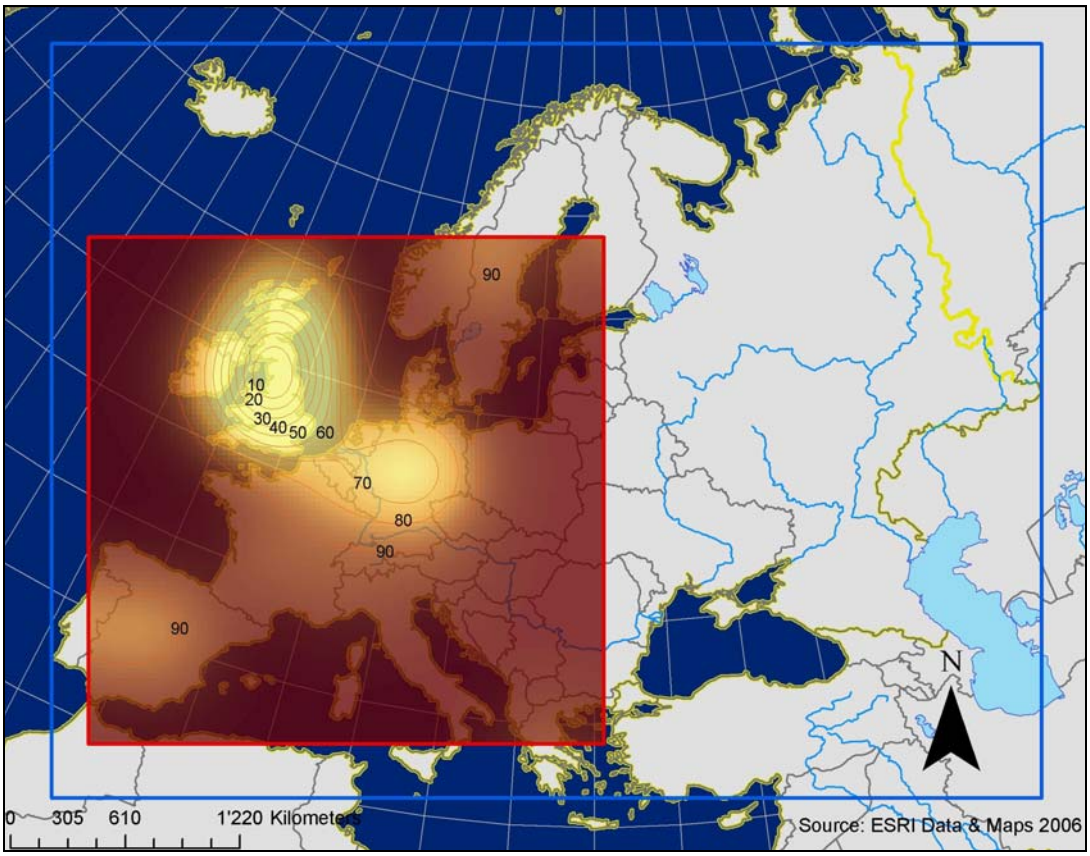


Fig. 5-7: Trade unions in Europe; result extent with KDE at D_{100} and expected extent

Western Europe is nicely covered by the footprints distribution's minimum bounding box at D_{100}. It includes Spain and the United Kingdom in the west, Finland in the northeast and Greece in the southeast. Further in the east, towards Russia and further in the north no points were assigned and the extent appears to be very similar to the European Union. Within the result extent, the point distribution is illustrated by the density surface, which reveals the cluster that was expected in the previous chapter. The Footprints are concentrated in the United Kingdom and Germany, the remaining area's point density can be considered marginal. Furthermore, the influence of the document collection is visible in the high concentration of Footprints in the United Kingdom and in Germany, due to the toponym occurrences in the *Glasgow Herald*, the *Frankfurter Allgemeine Zeitung* and *Der Spiegel*. The contour line spacing is tight in the United Kingdom and slightly wider towards Germany and central Europe, which indicates a steep slope towards the density surfaces peak in the United Kingdom.

Fig. 5-8 is the map representation of the result for "Walking holidays in Scotland". The covered extent is similar to the expected extent. A larger uncovered area is found mainly in the south. In the KDE area versus value range graph we saw that a strong clustering can be expected. Indeed, a heavily clustered footprint distribution with a steep slope towards the

footprint density peaks is found. The footprints are concentrated at a couple of prominent places such as Glasgow, Edinburgh or, less prominent Kirkwall in Scotland's north.

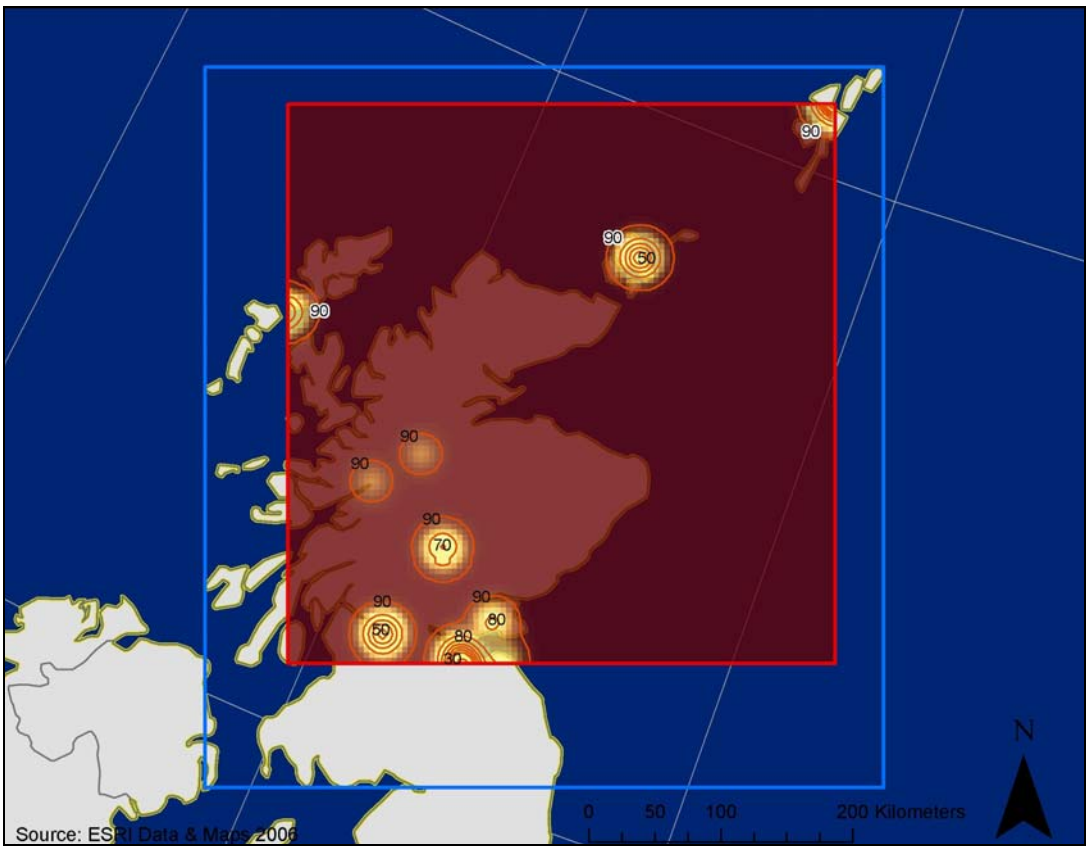


Fig. 5-8: Walking holidays in Scotland; result extent with KDE at D_{100} and expected extent

Another advantage of map visualisations is that the number of clusters can be investigated. Until now, by using the KDE area versus value range graph, conclusions could be derived about the slope of the density surface, the value range included by the cluster and its heaviness, but none about the number of clusters. In this map we see that the little area that reaches more than 90% of the KDE value range is concentrated at eight to ten locations.

For the last query, "Actions against the fur industry in Europe and the U.S.A", two maps corresponding to the two query regions were produced. Fig. 5-9 illustrates the extent overlap and KDE for the query region Europe. We see that the poor coverage found in Fig. 5-6 is mostly due to the small result extent in the European part, which reaches only about 20% of the expected extent. The distribution of the KDE values is similar to the one in the result for "Trade unions in Europe", with a density maximum in the United Kingdom, a wider spacing between the contour lines than in the result for "Walking holidays in Scotland", and a constant spacing until the 90% KDE contour level. Fig. 5-10 represents the result of the second query region, the U.S.A. The covered extent is higher than in the Europe part and the largest difference between the expected and the result extent is found in the south.

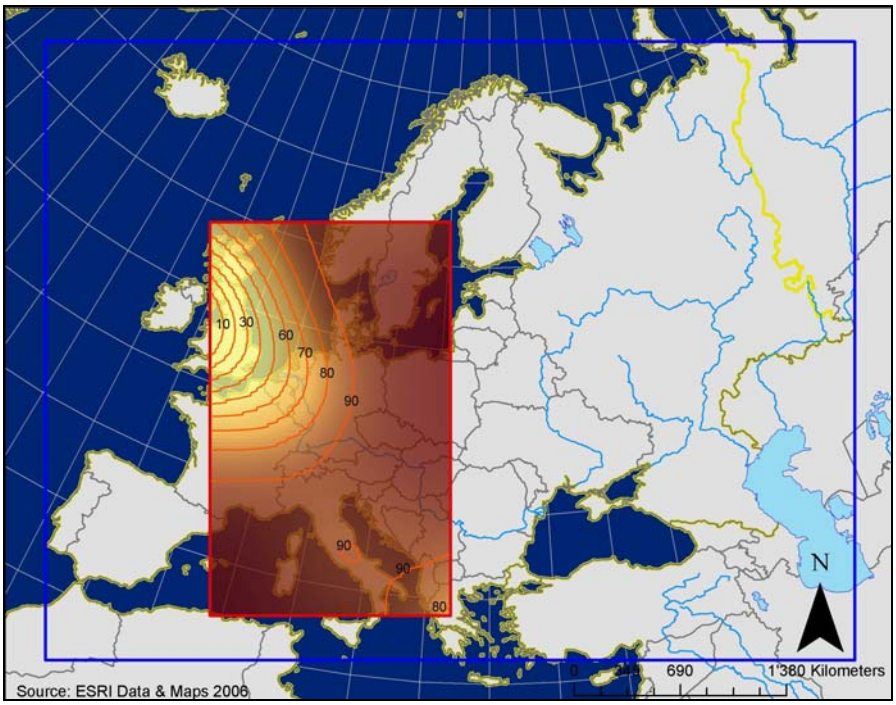


Fig. 5-9: Actions against the fur industry, Europe; result extent with KDE at D_{100} and expected extent

Most of the footprints, however, are located along the two coast lines; the footprint density in the central states is low, as the KDE reveals. Most of the relevant footprints were found on the east coast and some of them are assigned to ambiguous place names, such as Essex, Dover or Plymouth, which could also refer to Great Britain. The expected extent does not include Alaska, Hawaii and Puerto Rico, since they might not be too interesting for this query. They could be included, however, by splitting the expected extent further up as done before for this topic, since a single encompassing extent for the U.S. including these states is too inaccurate.

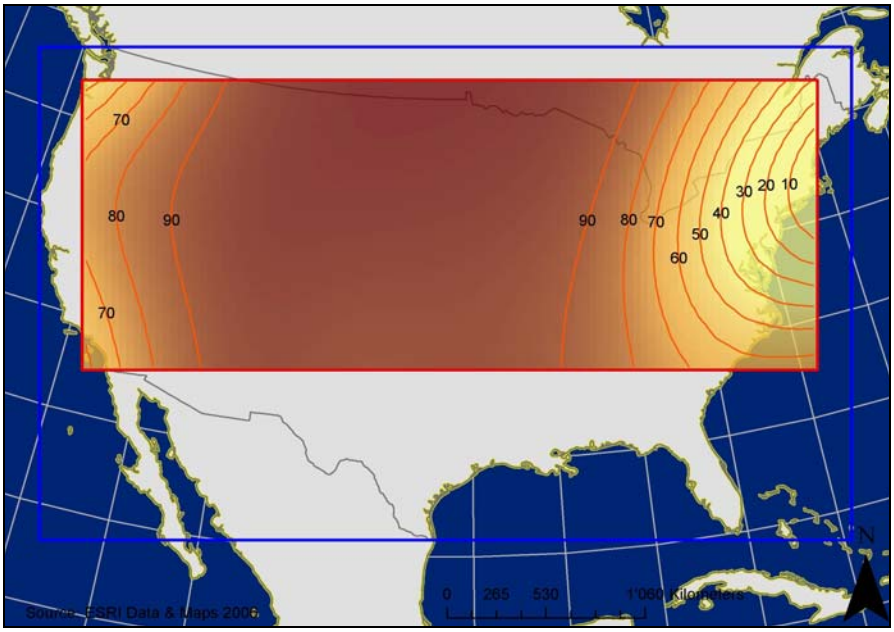


Fig. 5-10: Actions against the fur industry, U.S.A; result extent with KDE at D100 and expected extent

6. Discussion

6.1. Findings

6.1.1. Possibilities and Limitations of extended ranking-based spatial Evaluation

The methodology and result discussion concerning the SPIRIT evaluation part was already done in the two conclusions chapters 3.2.4 and 3.3.4. In chapter 3.2.4 we saw that footprint distributions can be quantified by using the spatial extent they cover and set in relation to a preliminary expected spatial extent. Further conclusions were that "in queries" have a significantly higher percentage of intersecting result area than the other queries and that the concept of minimum bounding boxes used for extent analysis works well enough and would not need to be replaced by convex hulls. In chapter 3.3.4, the decision was taken to use a combination of extent comparison and KDE area versus value range graph in order to evaluate the GeoCLEF results. It was considered the most promising approach, since it describes both spatial aspects, extent and distribution, and further allows the combination with standard IR evaluation.

The rest of this chapter will focus on the GeoCLEF evaluation part as well as on overall discussion issues. The most important finding and main objective of this work is a proposed methodology that measures the spatial quality and allows for the comparison between different GIR rankings' spatial properties. The extended ranking-based spatial evaluation is based on the binary document relevance judgement and uses the ranked relevant footprints in order to assess the covered extent and the characteristics of the footprint distribution in space. This new approach addresses and combines a number of previous efforts and approaches, and can answer some of the problems occurring when working with spatial relevance.

Bucher et al. (2005), Clough et al. (2006) and Purves & Clough (2006) introduced a ternary scheme as the binary scheme was too simplistic and tried to give "better" documents a "better" judgement. In this work, the document judgement is still binary, but the idea that "better" documents result in a higher score is taken over. A document is "good", if it is both spatially and thematically relevant. Binary spatial relevance can be defined as footprints lying inside of the expected extent. According to Kreveld et al. (2005), a document or a ranking gets more interesting, if its footprint spreading within the relevant region increases. By using

the proposed methodology, a document containing many spatially relevant footprints will result in a large result extent and cover a high percentage of the expected extent. Since all results are plotted against the ranking positions, it is possible to see what extent is covered by which ranking position group and conceive, for instance, the best ranking as the one that covers the largest extent earliest in the ranking. Clustering tendencies, areas with low footprint density as well as footprint hotspots can be detected by using the KDE value distribution graph. The more area is covered by high KDE values, the more even the footprint distribution and the less redundant footprints found in the pattern. Again, the best result could be defined as the most evenly distributed one, which indicates a high spreading within the result extent.

If we look back to chapter 2.1.4.3, we remember that queries became geographically more challenging over the years, including difficulties like (1) ambiguity, (2) vague geographic regions (Near East), (3) geographical relations beyond IN, (4) granularity below the country level and (5) complex region shapes (along the rivers Danube and Rhine) (Mandl et al. 2008). Some of those problems do not only affect the searching and ranking of documents, but also the relevance judgement and the evaluation. The proposed methodology, however, can answer most of these challenges. Problem (3) has been discussed extensively in 3.2.2.2; problems (2) and (5) demand an additional effort in defining an expected extent, since boundaries of imprecise regions are harder to define and rely on people's cognition. Purves et al. (2005) introduces an approach which uses web mining and produces a density surface from web occurrences. An expected extent and its bounding box could be derived from these density surfaces and used for the further spatial analysis.

Problem (4), that is low granularity, is a fundamental problem of the geoCLEF setup and was encountered at almost every query. Many of the results were not interesting for the extended ranking-based spatial evaluation because of the poor spatial footprint distribution and massive footprint clusters at a few of prominent places, which mostly were of a high granularity level. The problem that the data basis does not provide a sufficiently fine granularity level and has an inherent predefined footprint distribution is further discussed in chapter 6.2.

Due to the extended ranking-based spatial evaluation methodology and its ability to measure and map spatial characteristics of a GIR result, the first hypothesis H_1 is rejected.

H_1: All relevant characteristics of a GIR ranking result can be revealed using precision and recall measurement and spatial considerations do not benefit the process of GIR result evaluation.

The SPIRIT data showed a finer granularity than the GeoCLEF data. One reason for that is surely that queries were in general at a finer granularity stage too. Other reasons might be the different data basis, web pages instead of newspaper articles and most of all the resources used for toponym identification. Furthermore, the SPIRIT data appeared to be distributed more evenly and maximum occurrence of the same footprint was smaller than in the GeoCLEF data, where the density surface slope was much steeper in some cases. However, this might have something to do with the fact that, for the SPIRIT queries, only the first ten ranks were evaluated. If the first 100 ranks had been evaluated, the density slope might have got steeper too due to the accumulation of heavier clusters in more prominent places. On the other hand, the SPIRIT data was not checked in terms of binary relevance and no qrels file is available for these data. This means that all documents are assumed to be relevant and the number of footprints might thus be too high.

In general, the comparisons between the SPIRIT and GeoCLEF results are problematic, since the queries and the document collections are not equal and different ranks are considered. Nevertheless, some interesting statements can be derived from a confrontation of the two analysed datasets, which concern most of all the preconditions for geographic searching created by the used data basis. Fig. 6-1 illustrates the assigned number of unique footprints of the 20 SPIRIT queries and the 25 2005 ICL GeoCLEF queries. The first 100 ranks of the GeoCLEF data provided just a few more unique footprints than the first ten SPIRIT ranks. Since the query scope of the GeoCLEF queries is usually higher than that of the SPIRIT queries (in 2005, only six out of 25 query scopes were below the country level), the number of unique footprints found by the SPIRIT systems becomes even more remarkable. This confirms the impression that SPIRIT results usually achieve a good spatial spreading within the first ranking positions due to the underlying scattered ranking algorithm (Kreveld et al. 2005) and the gazetteer's fine granularity.

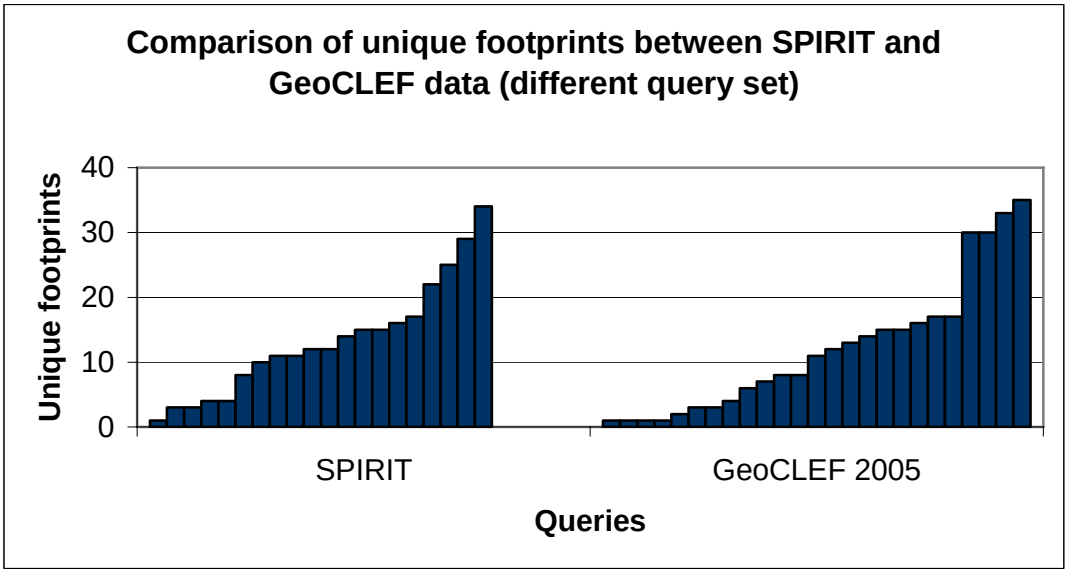


Fig. 6-1: Number of unique footprints in SPIRIT versus GeoCLEF 2005 data

6.1.2. Geographical Limitations of GeoCLEF

In four 2005 GeoCLEF data examples, only one unique footprint, which is equal to the query region, is found. Geographical search is rather pointless under these circumstances; a keyword matching based system would deliver equal or better results. According to Andrade & Silva (2006), spatial ranking and relevance judgement are query-dependent. If queries like "Shipwrecks in the Atlantic Ocean" or "Sea rescue in North Sea" sound geographically tricky, they are in fact not very geographical, since the geographical granularity of the Atlantic Ocean is unlikely to be represented within the used gazetteers and a spatially distributed result can not be obtained. Another indication for a lack of consideration of geographical granularity and hierarchical relationships is the large number of footprints that is equal to the query region, as documented in Fig. 5-3.

These observations are confirmed by a number of publications concerned with the performance at GeoCLEF and post hoc experiments. Bischoff et al. (2006a) lamented the lack of relevant documents in the 2005 German document collection and the resulting low MAP values. In Clough et al. (2005), many participants commented on the first GeoCLEF topics as being too similar to normal adhoc queries, with purely text based retrieval performing better than spatial methods. Text based IR methods that outperform spatially aware approaches still appear to be an issue in more recent publications. Cardoso et al. (2007) were "*intrigued with the constant better results obtained with the Terms Only experiment*", although the Lisbon group realised a number of changes in their system. For 2007, they renewed query processing with a special focus on geographical feature types and spatial relationships, renewed the text

mining by assigning documents a geographic signature, which is a list of geographic concepts, and adapted the geographic ranking to these. At the same time, their ontology is not comprehensive on coordinates in order to serve the geographical heuristics that calculate a spatial similarity.

Post hoc experiments by the ICL group revealed similar results on the topic of text versus geo-aware searching. Overell et al. (2007), "*attribute the relatively poor results of the Text+Geo method to the way the textual and geographic relevance were combined*". Another example confirming the superiority of textual search is given in Li et al. (2007), who note that "*MSRAText run achieves the best precision in our results*". The sometimes un-geographical queries in combination with a data basis of very asymmetric granularity and distribution of place names appear to create a testing environment in which textual search performs better than geographically aware search. Therefore, H_2 can be rejected too, which means that text based retrieval performs better if queries and document collection do not enable spatially aware search.

H_2: As soon as a query is built following the concept@location form, geographical aware information retrieval outperforms text based retrieval and independently allows for spatial relevance judgement of documents and evaluation of ranking results.

A fine granularity in the document collection is necessary for geographical intelligence to actually operate. The queries should be formulated for both parts, the thematic and the spatial concept, to contain relevant information for the retrieval. Classic IR is likely to perform better, if the query is a combination of key words, whose appearance in newspaper articles is highly correlated, such as "combats" and "Iraq". Moreover, GIR relevance judgement and evaluation are dependent on the query, the spatial operator and the document collection. In the SPIRIT part, many results were not evaluated, since the footprint distributions contained too little unique points. In the GeoCLEF examples, the clustering tendencies were even stronger and the footprint distributions provided by the data basis appeared to be poor with respect to the wide query regions. Therefore, spatial evaluation needs to be thought of as a dynamic process, which means that a result referring to "Atlantic Ocean" might be treated differently than one referring to "Europe". The same applies for different spatial operators. Queries containing "near" are harder to handle than queries containing "in" and might therefore need adjustments in the evaluation methodology.

So far, the best strategy for achieving a high MAP was found by the Berkeley group (Gey & Petras, 2005), who use a fined tuned blind relevance feedback. Note that improvements in

Gey & Petras (2005) for English monolingual runs were moderate compared to English – German bilingual runs. Bischoff et al. (2005b) agree that query expansion by adding geographic names from the top-ranked documents improved retrieval performance substantially. A different opinion is represented by Li et al. (2007) who comment that "*automatic query expansion by pseudo feedback weakens the performance because the topics are too hard to be handled and many unrelated locations are added to new topics*". Most of the participant groups, however, see query expansion and relevance feedback as one key parameter for their system performance and do research in order to improve it. The selection of the right terms as well as their weighting appear to be the crucial points for successful query expansion. An example for failure of query expansion is also found in Gey & Petras (2005), where the expansion of the German term "Europa" to all European country names resulted in a zero precision for the corresponding topics.

6.2. Problems

A major problem of this work is that direct comparisons between the different result sets are problematic and have a limited validity. The SPIRIT and the GeoCLEF data used in this work have a number of important differences:

- Different queries were performed with a different underlying data basis.
- SPIRIT queries are on a finer granularity level then GeoCLEF queries.
- GeoCLEF: Top 100 ranks, SPIRIT: top 10 ranks.
- GeoCLEF data are checked for relevance in a binary scheme and irrelevant documents are excluded, whereas all SPIRIT documents are assumed to be relevant.
- GeoCLEF results are checked for footprints equal or of equal granularity to the query region, which will be excluded for spatial calculations. SPIRIT data might contain points standing for the entire query region at its centroid coordinate, but a description of the document footprints was not available.

Since the extended ranking-based spatial evaluation seems to work, as chapter 5.2 shows, it would have been interesting to compare the results of two equivalent datasets. A direct comparison between two different GeoCLEF result sets of the same queries is missing as well, since no second result set was available. The result interpretation is therefore limited to the comparison of the results for different queries. This is especially true for the covered versus expected extent measure. The comparison between the results of different systems for the same query would be more interesting than the comparison of the result extent to the expected extent only or the comparison between the covered extents for two different queries.

The expected extent is understood as a unit for normalisation and might be too large or otherwise inappropriate in some cases. Therefore, the behaviour of the resulting extents produced by different systems for the same query, with the expected extent used as benchmark, would be more interesting than the comparisons that were possible with the available data in this work.

Finally, GeoCLEF might not be an optimal testing environment for the development and testing of a spatially aware evaluation methodology. According to Fig. 5-1, not more than 20 documents are relevant for most of the topics. If those documents are found by the retrieving algorithms, the footprint distribution is mostly predefined and will be equal for all runs. Furthermore, we saw in Fig. 5-3 that the number of unique footprints and hierarchically filtered footprints, which are crucial for the application of spatial evaluation, is smaller than twenty in most cases. However, one needs to keep in mind that these results are of one participant group only and that other participant groups might have identified more spatially interesting footprints due to a more detailed gazetteer or a higher reached recall level. An optimal testing environment for the extended ranking-based spatial evaluation should deliver more relevant documents with more spatially unique footprints, so that different extent and distribution types may actually arise and the method's behaviour and reaction to a wider range of results might be explored.

6.3. Recommendations

How useful are newspaper article archives as data basis? Brunner (2008) claims that place names mentioned in articles are generally clustered and that toponyms in local newspapers tend to be concentrated in the region in which the newspaper is located. Further conclusions from that work are that a fine granularity is unlikely to be achieved in articles that have a large scope, and that articles with references to country names cover a larger scope and are of higher granularity than articles with references to local toponyms. Fig. 6-2 illustrates the spatial distribution of the relevant footprints assigned by the ICL group at GeoCLEF 2005. Note that the distribution is also biased by the query regions, which means that the high density values in the United Kingdom are not only caused by the toponym distribution in the *Glasgow Herald*, but also by the frequent appearances of terms like "Scotland", "United Kingdom" or "Highlands" within the search topics. A further reason for the strong clustering might also be the data's underlying "most referred to" toponym disambiguation approach. Ambiguous toponyms are always assigned to the same instance, which increases clustering at these locations.

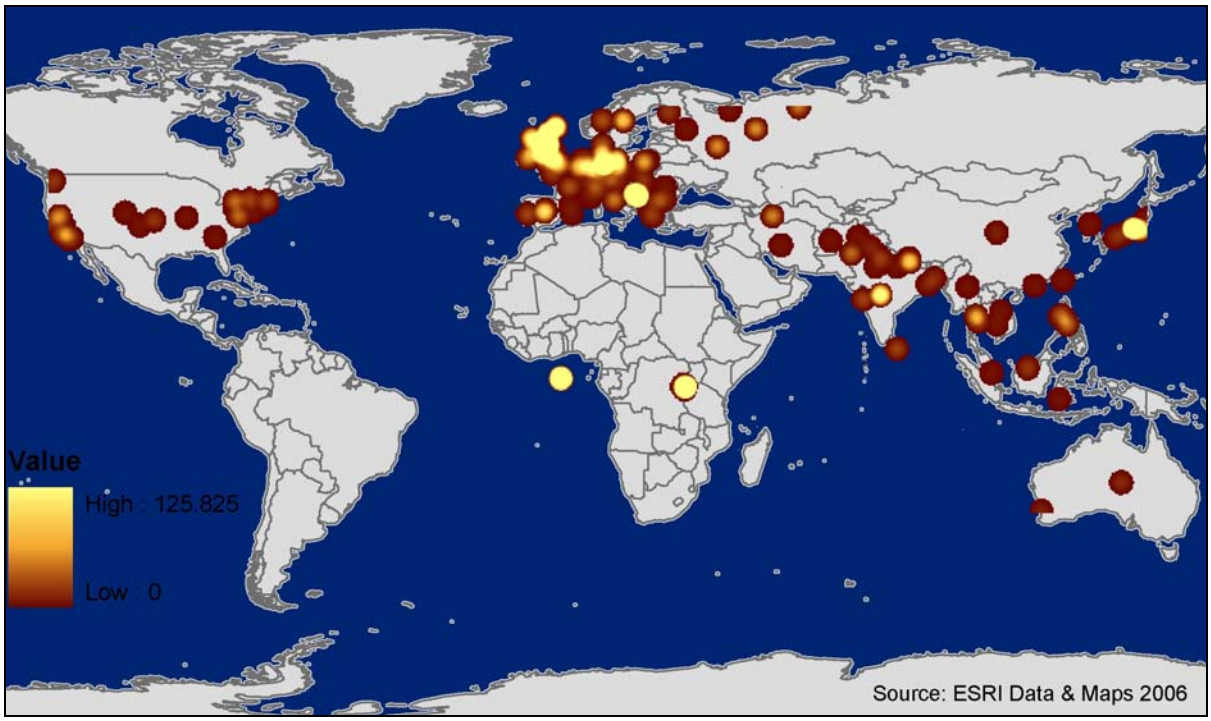


Fig. 6-2: Distribution of relevant footprints in GeoCLEF 2005

Gey et al. (2007a) noticed the same problem within their remarks to the 2006 GeoCLEF, stating that "*the snow conditions or danger of avalanches in Grisons (canton in Switzerland) may be reported frequently by the Swiss news agency SDA or even German newspapers, whereas the British or American newspapers may not see the relevance for their audience*". Their approach to resolve this problem is that "*well known places and larger regions as well as international relevant or dramatic concepts had to be focused on, although this may not reflect all user needs for GIR systems*".

A first recommendation resulting from this work would be to diversify the document collection rather than find topics that deliver a few badly distributed and high granularity results within the current document collection. It is true that the *L.A. Times* would probably not report on the danger of avalanches in Grisons, while the local newspaper *Die Südostschweiz* probably would. A more promising approach towards a GIR evaluation setup could thus be to adapt the document collection to the query region and to use more local data sources, which provide more relevant documents and a finer granularity. Moreover, the thematic part and the spatial part of the queries should be more independent from one another.

Another interesting issue would be to investigate those queries, in which text based IR performs more strongly than GIR and put these into a context with the granularity of the query region, the granularity of the document collection as well as with the footprint frequencies and distributions within the relevant documents. This might deliver a

classification of constellations, in which text based IR performs better than GIR and vice versa, and systems could be trained to use either one of them, depending on the constellation. For such a quantitative analysis, however, more results from a wider range of GIR systems, also including text based runs, are required.

Bucher et al. (2005) noticed that spatial relevance judgement is best done by assessors with local knowledge of the area under query. This work confirmed that local knowledge or at least additional geographic knowledge is necessary. For the extended ranking-based spatial evaluation methodology, this knowledge is required in order to define a reasonable expected extent, on which the binary spatial document relevance judgment is based and to which the result extents are compared. Hence, if a lack of geographical local knowledge is encountered, it has to be gained by considering geographic information of any kind.

The binary document relevant judgement might be criticised for being too simple and not revealing the differences between “better” and “worse” fitting documents. All spatial measures and key figures presented in this work could also be used for single document relevance judgement and thus improve the current binary system. Instead of measuring the covered extent and the clustering behaviour of different ranking groups, the spatial methodology can be applied to single documents. However, one needs to keep in mind that most of the documents will contain only little relevant footprints, which decreases the significance of most spatial measures.

A fundamentally different approach to relevance judgement could be derived from the ideas examined in this work. Instead of assigning binary values, the relevance judgement could be understood as comparing and ranking documents relatively to each other, which produces a preliminary document ranking following the quality of thematic and spatial responsiveness in decreasing order. Again, a document, which is thematically relevant and contains a wide range of geographically relevant footprints as well as an even footprint distribution, would be considered “better” than a document that offers little relevant footprints or shows footprint clusters. The ranking evaluation would then need to consider both the quality of the document from the relevance judgement ranking and the position at which it occurs in the result ranking. The earlier a highly important document occurs in the ranking, the better the result of the ranking evaluation.

7. Conclusions

7.1. Contributions

The chapter "current state of research" gave an overview about different GIR processes, systems and the GeoCLEF evaluation campaign, and presented a compilation of pattern and cluster analysis techniques. In the first practical part, a methodology for the spatial description of GIR results was outlined by using SPIRIT data. The most promising approach was found to be a combination of extent overlap and Kernel Density Estimation value distribution measurement which represents an explorative pattern analysis approach. Furthermore, Mann Whitney U-Test revealed a significant relationship between extent intersection behaviour and queries containing the spatial operator "in". In the second methodology part, the proposed extended ranking-based spatial evaluation was used for the evaluation of GeoCLEF 2005 results, provided by the Imperial College London group. Extended ranking-based spatial evaluation can be summarised into the following steps:

- Definition of an expected extent
- Binary document relevance judgement: thematic and spatial based on the expected extent
- Ranking based evaluation: precision, recall, point density, covered extent, Kernel Density Estimation value distribution
- Graph and map visualisations

This approach is a combination of the present single document relevance judgement and the ranking evaluation based on standard IR benchmarks. This new evaluation approach, however, includes a spatial dimension allowing for the explicit description of the results' spatial characteristics. The results of the proposed methodology come in compact form and in percentage, which makes them easy to interpret. Moreover, a map output with a footprint density surface, which illustrates the spatial properties of the ranking result in an intuitive way, can be created. The hypothesis that spatial considerations and spatial measures do not additionally benefit the result evaluation was therefore rejected.

Many spatially aware runs at GeoCLEF were outperformed by purely text based IR methods. Possible reasons for the partial failure of methods that respect geographical space were given in the result and in the discussion chapter. The combination of queries and data basis was found to be insufficiently spatially considerate. This means that the use of a high scale

gazetteer or of spatially biased newspaper articles creates a poor geographic granularity, in which the geographic intelligence might be outperformed by purely text based retrieval. A further conclusion concerning GIR evaluation setups is that the spatial relevance judgement and the usefulness of the spatial ranking evaluation are somehow dependent on the query, its geographical granularity, its spatial operator as well as on the geographical granularity provided by the underlying document collection. The hypothesis that GIR outperforms IR, if a query has a spatial component and relevance judgement and evaluation can be done independently of the query, was therefore rejected too.

7.2. Implications

One of the aims of this work was to propose a GIR result evaluation methodology which respects the results' spatial characteristics and which is objective and reproducible. The proposed relevance judgement and evaluation method fulfils these requirements and gives a more spatial perspective to the result evaluation, which should be of interest for GIR. A straightforward implication would thus be to apply a methodology based on this work's suggestions in a GIR evaluation campaign.

Given the fact that text based IR still outperforms GIR, many publications focus on possible reasons for the superiority of text based IR. In this work, a critical look was taken at the GeoCLEF setup in order to find out if the queries and the data basis are spatially considerate enough or if they are actually better suited to text based IR. A recommendation derived from these observations would be the following: For geographically and spatially aware searching to be beneficial, the data basis needs to be adapted to the query regions in order to ensure a fine granularity. In this way, geographical awareness and intelligence can actually operate. The data basis should be enlarged in order to include articles of local newspapers of all queried regions.

Another interesting research question concerning the performance differences between IR and GIR would be to see if purely text based IR still outperforms GIR, when space is explicitly respected in the result evaluation.

7.3. Outlook

This masters thesis is meant to be a first attempt to include point pattern and cluster analysis methods into GIR result evaluation. If this approach is accepted by the GIR community, more research is necessary in order to create more robust methods which are able to evaluate poorly distributed results and include spatial measures that allow for statistical inherence, if required.

Another important issue is the degree of automatisation. In this work, some of the calculation steps were automatised but isolated from one another, others were repeated manually. The first step of the procedure, that is the definition of an expected extent and that of an appropriate projection, would still rely on human judgement as well as the thematic relevance judgement. The spatial relevance judgement based on the defined spatial extent could be done automatically by checking the documents for footprints inside the specified area. All remaining steps could be automatised and integrated into one single procedure which would allow for the production of more evaluation results as well as for more quantitative analysis and direct comparisons between different systems.

The aim of these efforts is an improvement of GIR result's spatial quality through feedback of the spatially aware relevance judgement and evaluation. This should help to emphasize space more at all GIR stages and to create a real benefit compared to purely text based IR.

8. References

Agosti, M., Di Nunzio, G.M., Ferro, N., Peters, C. (2007): CLEF: Ongoing Activities and Plans for the Future. Proceedings of NTCIR-6 Workshop Meeting, 15-18 May 2007, Tokyo, Japan.

Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2004): Web-a-Where: Geotagging Web content. In: Sanderson, M., Jàrvelin, K., Allan, J. & Bruza, P. (Eds), (2004): SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25-29 July 2004, 273-280, (New York: ACM Press).

Andrade, L., Silva, M. J. (2006): Relevance Ranking for Geographic IR. In: Proceedings of the workshop on Geographic Information Retrieval, SIGIR 06, Seattle, USA.

Bischoff, K., Mandl, T., Womser-Hacker, C. (2006a): Cross-linguales geographisches Information Retrieval. In: FGIR 2006: Workshop Information Retrieval 2006 of the Special Interest Group Information Retrieval (FGIR): (Hildesheim): 9-11 October 2006, LWA 2006 : Lernen - Wissensentdeckung – Adaptivität, Martin Schaaf, Klaus-Dieter Althoff [Eds].

Bischoff, K., Mandl, T., Womser-Hacker, C. (2006b): Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006. In: Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006, Alicante, Spain, http://www.clef-campaign.org/2006/working_notes/.

Brunner, T. (2008): Geographic Information Retrieval: Identifikation der geographischen Lage von Zeitungsartikeln. Masterarbeit, Geographisches Institut, Universität Zürich.

Bucher, B., Clough, P., Joho, H., Purves, R.S., Syed, A.K. (2005): Geographic IR systems: requirements and evaluation. In: Proceedings of the 22nd International Cartographic Conference, A Coruña, Spain, CD-ROM.

Buckley, C., Voorhees, E.M. (2000): Evaluation measure stability. In: Belkin, N. J., Ingwersen, P. & Leong, M.-K. (Eds), (2000): Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 24-28 July 2000, 33-40 (New York: ACM Press).

Bush, V. (1945): As We May Think. Atlantic Monthly, 176,101–108, July 1945.

Cai, G. (2002): GeoVSM: An Integrated Retrieval Model for Geographic Information. LNCS 2489, 70-85.

Cardoso, N., Cruz, D., Chaves, M., Silva, M.J. (2007): The University of Lisbon at GeoCLEF 2007. . In: Working Notes for CLEF 2007, Budapest, Hungary, http://www.clef-campaign.org/2007/working_notes/.

Cardoso, N., Silva, M.J. (2007): Query Expansion through Geographical Feature Types. In: 4th Workshop on Geographic Information Retrieval, GIR 07 (held at CIKM'07), Lisbon, Portugal, 9th November 2007.

Chen, A., Gey, F. (2004): Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Decompounding. Information Retrieval 7 (1-2), 149-182.

Clark, P.J., Evans, F.C. (1954): Distance to nearest neighbour as a measure of spatial relationships in populations. Ecology, 35, 445-453.

Clough, P.D., Gey, F., Joho, H., Larson, R.R., Petras, V., Sanderson, M. (2005): GeoCLEF: the CLEF 2005 Cross-language Geographic Information Retrieval Track. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria, September 2005, http://www.clef-campaign.org/2005/working_notes/.

Clough, P.D., Joho, H., Purves, R.S. (2006): Judging the spatial relevance of documents for GIR. In: Proceedings of the 28th European Conference on IR Research (ECIR'06), London, UK, April 2006 LNCS Volume 3936, 548-552 (Berlin: Springer).

Cooper, W.S., Gey, F., Dabney, D.P. (1992): Probabilistic retrieval based on staged logistig regression. In: 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 21–24 June 1992, 198-210 (New York: ACM Press).

Cunningham, H., Maynard, D., Bontcheva, K., 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.

Diggle, P. J. (1990): A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. In: Journal of the Royal Statistical Society, Series A, 153, 349-62.

Di Nunzio, G.M., Ferro, N. (2005): DIRECT: A System for Evaluating Information Access Components of Digital Libraries. In A. Rauber, S. Christodoulakis, and A. Min Tjoa, (Eds), Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005). Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 483-484.

Egenhofer, M., Mark, D. (1995): Naive Geography. In: Spatial information theory. A theoretical basis for GIS (COSIT95), A. Frank and W. Kuhn (Eds). Lecture Notes in Computer Science, 988, 1-15, (Berlin: Springer).

Ferrés, D. D. (2007): Geographical Information Resolution and its Application to the Question Answering Systems, Universitat Politècnica de Catalunya.

Frontiera, P., Larson, R.R., Radke, J. (2008): A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science, 22(3), 337-360.

Fu, G., Jones, C.B., Abdelmoty A.I. (2005): Ontology-Based Spatial Query Expansion in Information Retrieval. In: OTM Conferences (2), volume 3761 of LNCS, 1466–1482, Springer.

Gattrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S., (1996): Spatial point pattern analysis and its application in geographical epidemiology. Transactions of the Institute of British Geographers 21, 256–274.

Gey, F., Larson, R.R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P. (2007a): GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Evaluation of Multilingual and Multi-modal Information Retrieval, (Berlin: Springer).

Gey, F., Larson, R.R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Montoyo, A. (2007b): Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF. The First International Workshop on Evaluating Information Access (EVIA), May 15, 2007, Tokyo, Japan.

Hauptmann, A.G., Olligschlaeger, A.M. (1999): Using Location Information from Speech Recognition of Television News Broadcasts. In: Robinson, T. & Renals, S. (Eds), (1999): Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio. Cambridge, U.K., 19-20 April 1999, 102-106.

Joachims, T., Granka, L., Pan, B. (2005): Accurately Interpreting Clickthrough Data as Implicit Feedback. In: Annual ACM Conference on Research and Development in Information Retrieval, SIGIR'05, 15-19 August 2005, Salvador, Brazil (New York: ACM Press).

Joho, H., Sanderson, M. (2004): The SPIRIT collection: an overview of a large web collection. SIGIR Forum, 38(2), 57-61.

Jones, C.B., Purves, R.S (2005): Proceedings of the ACM Workshop on Geographic Information Retrieval (GIR) held at the Conference on Information and Knowledge Management (CIKM), Bremen, Deutschland, November 4, 2005 (New York: ACM Press).

Jones, C.B., Purves, R.S., Ruas, A., Sanderson, M., Sester, M., Kreveld, M. van., Weibel, R. (2002): Spatial information retrieval and geographical ontologies: an overview of the SPIRIT project. In: Proceedings of SIGIR-02, the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 387-388 (New York: ACM Press).

Jones, K.S. (1972): A Statistical Interpretation of Term Specificity and its Application to Retrieval. Journal of Documentation, 28(1), 111-121.

Kluck, M., Mandl, T., Womser-Hacker, C. (2002): Cross-Language Evaluation Forum (CLEF): Europäische Initiative zur Bewertung sprachübergreifender Retrievalverfahren. In: Information – Wissenschaft und Praxis, 53 (2), 82-89.

Kreveld, M. van., Reinbacher, I., Arampatzis, A., Zwol, R. van, (2005): Multidimensional scattered ranking methods for geographic information retrieval, Geoinformatica, 9(1), 61-84.

Krupka, G. R., Hausman, K. (1998): Isoquest, Inc: Description of the NetOwl extractor system as used for MUC-7. In: 7th Message Understanding Conference (MUC 7): Proceedings of a Conference held in Fairfax, Virginia, USA, April 19 - May 1.

Kulldorff, M. (1997): A spatial scan statistic. In: Communication in Statistics – Theorv and Methods, 26, 6, 1481-1496.

Kulldorff, M., Feuer, E.J., Miller, B.A., Freedman, L.S. (1997): Breast Cancer Clusters in the Northeast United States: A Geographic Analysis. In: American Journal of Epidemiology, 146, 2, 161-170.

Lakoff, G., Johnson, M. (1980): Metaphors we live by. Chicago University Press.

Lancaster, F.W. (1968): Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York, 1968.

Larson, R.R. (1996): Geographic Information Retrieval and Spatial Browsing. In: Smith, L. & Gluck, M. (Eds), (1996): GIS and Libraries: Patrons, Maps and Spatial Information, University of Illinois, 81-124.

Larson, R.R., Frontiera, P. (2004a): Ranking and Representation for Geographic Information Retrieval. In: Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004, 25-29 July 2004, Sheffield, United Kingdom.

Larson, R.R., Frontiera, P. (2004b): Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. In: Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL 2004, R. Heery and L. Lyon (Eds). September 2004, Bath, UK, Lecture Notes in Computer Science 3232, 45-57 (Berlin: Springer).

Leidner, J. (2008): Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Dissertation.com.

Leidner, J., Sinclair, G., Webber, B. (2003): Grounding spatial named entities for information extraction and question answering. In: Kornai, A. & Sundheim, B. (2003): HLT-NAACL 2003 Workshop: Analysis of Geographic References, 31-38. Association for Computational Linguistics, Edmonton, Alberta, Kanada.

Li, H., Srihari, R.K., Niu, C., Li, W. (2003): InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Kornai, A. & Sundheim, B. (Eds), (2003): HLT-NAACL 2003 Workshop: Analysis of Geographic References, S. 39-44. Association for Computational Linguistics, Edmonton, Alberta, Kanada.

Li, Z., Wang, C., Xie, X., Ma, W-Y. (2007): MSRA Columbus at GeoCLEF2007. In: Working Notes for CLEF 2007, Budapest, Hungary, http://www.clef-campaign.org/2007/working_notes/.

Luhn, H.P. (1957): A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1957.

Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R.R., Sanderson, M., Santos, D., Womser-Hacker, C., Xing, X. (2008): GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol et al. (Eds.). 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, (Berlin: Springer).

Markert, K., Nissim, M. (2002): Towards a corpus annotated for metonymies: the case of location names. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), 1385-1392. ELRA, Paris, France.

Maron, M.E., Kuhns, J.L (1960): On relevance, probabilistic indexing and information retrieval. In: Journal of the ACM, 7, 216-244.

Martins, B., Silva, M.J. (2004): A statistical study of the WPT-03 corpus. DI/FCUL TR 04–04, Department of Informatics, University of Lisbon

McCurley, S.K. (2001): Geospatial mapping and navigation of the web. In: Proceedings of the 10th International WWW Conference, Hong Kong, 1–5 May 2001, 221-229 (New York: ACM Press).

Openshaw, S., Charlton, M., Wymer, C., Craft, A. (1987): A mark 1 geographical analysis machine for the automated analysis of point data sets. In: International Journal of Geographical Information Systems, 1, 335–358.

Openshaw, S., Maguire, D.J., Goodchild, M.F., Rhind, D.W. (1991): Developing appropriate spatial analysis methods for GIS. In: Geographical Information Systems, 1, 389-401, Wiley + Sons Inc., New York.

Openshaw, S., Taylor, P.J. (1981): The modifiable areal unit problem. In: Quantitative Geography; A British View, N, Wrigley and R.J. Bennett (eds), Routledge and Kegan Paul, London, 60-70.

Overell, S.E., Magalhães, J., Rüger, S. (2007): GIR experiments with Forostar at GeoCLEF 2007. In: Working Notes for CLEF 2007, Budapest, Hungary, http://www.clef-campaign.org/2007/working_notes/.

Overell, S.E., Rüger, S. (2006): Identifying and grounding descriptions of places. In: SIGIR Workshop on Geographic Information Retrieval, S. 14-16.

Overell, S.E., Rüger, S. (2007): Geographic Co-occurrence as a tool for GIR. In: CIKM Workshop on Geographic Information Retrieval, Jones, C., Purves, R.S. (Eds), November 2007, Lisbon, Portugal, 71-76.

O'Sullivan, D., Unwin, D. (2003): Geographic Information Analysis. John Wiley & Sons, Inc., New Jersey.

Robinson, A.H., Morrison, J.L., Muehrcke, P.C., Kimmerling, A.J., Guptil, S.C. (1995): Elements of Cartography. 6th Edition, John Wiley & Sons Inc.

Purves, R. S., Jones, C.B. (2004): Workshop on geographic information retrieval at SIGIR 2004. SIGIR Forum, 38, 53-56.

Purves, R.S., Clough, P., Joho, H. (2005): Identifying imprecise regions for geographic information retrieval using the web. In: Proceedings of the GIS RESEARCH UK 13th Annual Conference, Glasgow, UK, R. Billen, J. Drummond, D. Forrest and E. João (Eds), 313-318 (Glasgow: University of Glasgow).

Purves, R.S., Clough, P. (2006): Judging spatial relevance and document location for geographic information retrieval. Extended abstract, In: Proceedings of 4th International Conference on Geographic Information Science (GIScience 2006), Münster, Germany, September 2006, 159-164 (Muenster, Germany: IfGI Prints).

Purves, R.S., Jones, C.B. (2006a): Geographic Information Retrieval (GIR). Computers, Environment and Urban Systems, Volume 30, Issue 4, 375-377.

Purves, R.S., Jones, C.B. (2006b): Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006 (New York: ACM Press).

Purves, R.S., Jones, C.B. (2007): Proceedings of the 4th ACM Workshop On Geographic Information Retrieval, GIR 2007, Lisbon, Portugal, November 9, 2007 (New York: ACM Press).

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B. (2007): The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet, International Journal of Geographical Information Science, Vol. 21, No. 7, 717-745.

Pouliquen, B., Steinberger, R., Ignat, C., De Groeve, T. (2004): Geographical information recognition and visualization in texts written in various languages. In: Haddad, H., Omicini, A., Wainwright, R. L., Liebrock & L. M. (Eds) (2004): Proceedings

of the 2004 ACM Symposium on Applied Computing, 1051-1058 (New York: ACM Press).

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fluart, F., Zaghouani, W., Widiger, A., Forslund, A.-C., Best, C. (2006): Geocoding multilingual texts: Recognition, disambiguation and visualisation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 53-58, ELRA.

Rauch, E., Bukatin, M., Baker, K. (2003): A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographical references, 50-54, ACL.

Ripley, B.D (1976): The second-order analysis of stationary point processes. In: Journal of Applied Probability, 13: 255-266.

Robertson, S.E. (1977): The probabilistic ranking principle in IR. In: Journal of Documentation, 33, 294-304.

Robertson, S.E., Walker, S., Beaulieu, M. (1998): Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In: NIST Special Publication 500–242: The 7th Text Retrieval Conference (TREC-7), Gaithersburg, MD E.M. Voorheer and D.K. Harman (Eds) (Gaithersburg, MD: NIST), 253-264.

Seaman, D.E., Powell, R.A. (1996): An evaluation of the accuracy of kernel density estimators for home range analysis. Ecology, 77(7), 2075-2085.

Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N. (2006): Adding geographic scopes to web resources. In: Computers, Environment and Urban Systems, Volume 30, 378-399.

Singhal, A. (2001): Modern information retrieval: a brief overview. IEEE Data Engineering Bulletin, Special Issue on Text and Databases, 24(4), Dec. 2001.

Smith, D. A., Crane G. (2001): Disambiguating geographic names in a historical digital library. In: Research and Advanced Technology for Digital Libraries: 5th European Conference (ECDL 2001), 127-136.

Tobler, W.R. (1970): A computer model simulation of urban growth in the Detroit region. Economic Geography, 46, 234-240.

Van Rijsbergen, C.J. (1979): Information Retrieval, Butterworth-Heinemann, Newton, MA. (http://www.dcs.gla.ac.uk/~iain/keith/index.htm)

Wang, C., Xie, X. et al. (2005): Detecting geographic locations from web resources. In: Proceedings of the 2005 Workshop on Geographic Information Retrieval. Bremen, Germany.

Yu, B., Cai, G. (2007): A Query-Aware Document Ranking Method for Geographic Information Retrieval. In: Proceedings of the 4th ACM workshop on Geographic Information Retrieval, November 2007, Lisbon, Portugal.

World Wide Web

Association for Computing Machinery (ACM): http://www.acm.org/; last accessed: 28.9.2008

CLEF: http://www.clef-campaign.org/; access: 3.8.2008.

Conference on Information and Knowledge Management (CIKM): http://www.cikm.org/; last accessed: 29.9.2008

Digital Chart of the World: http://www.maproom.psu.edu/dcw/ http://www-sul.stanford.edu/depts/gis/DCW.html#desc/; access: 6.5.2008.

ESRI: http://www.esri.com/; access: 13.8.2008

Evaluating Information Access (EVIA) workshop: http://ntcir.nii.ac.jp/index.php/EVIA-2008/; access: 3.8.2008.

Forestry Commission Great Britain: http://www.forestry.gov.uk/forestry/redkite/; access: 30.4.2008.

GeoCLEF 2005-2007: http://ir.shef.ac.uk/geoclef/; last accessed: 24.8.2008

GeoNames: http://www.geonames.org/; access: 6.7.2008.

Getty Thesaurus of Geographic Names: http://www.getty.edu/research/conducting_research/vocabularies/tgn/; access: 28.7.2008

Google Earth: http://earth.google.com/; last accessed: 14.5.2008

Google Maps: http://maps.google.com/; last accessed: 24.5.2008

Lucene: http://lucene.apache.org/java/docs/index.html; access: 3.8.2008

MetaCarta: http://www.metacarta.com/; access: 21.8.2008

NUTS 2003 Regions, 1:3 Million: http://epp.eurostat.ec.europa.eu/portal/page?_pageid=2254,64099847,2254_64185160&_dad=portal&_schema=PORTAL; access 17.4.2008.

NUTS Eurostat: http://ec.europa.eu/comm/eurostat/ramon/nuts/introduction_regions_en.html; access: 6.5.2008.

R 2.6.2; the R Project for Statistical Computing: http://www.r-project.org/; access: 22.2.2008.

Special Interest Group on Information Retrieval (SIGIR): http://www.sigir.org/; access: 29.9.2008

Spatstat; open source software for spatial statistics: http://www.spatstat.org/; access 25.2.2008.

Text retrieval conference: http://trec.nist.gov/; access: 29.9.2008

Text retrieval conference Qrels: http://trec.nist.gov/data/qrels_eng/; access: 5.8.2008

United Kingdom National Parks: http://www.nationalparks.gov.uk/; access: 22.4.2008.

Wikipedia: http://www.wikipedia.org/; last accessed: 30.9.2008

Wikipedia, Red Kite: http://en.wikipedia.org/wiki/Milvus_milvus; access: 30.4.2008.

Wikipedia, Stroud: http://en.wikipedia.org/wiki/Stroud,_Gloucestershire; access: 23.4.2008.

Persönliche Erklärung:

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Raphael Kunz

9. Appendix

SPIRIT Lambert Conformal Conic (Europe)

PROJCS["Europe_Lambert_Conformal_Conic_spirit_1984",
GEOGCS["D_WGS_1984",
DATUM["D_WGS_1984",
SPHEROID["WGS_1984",6378137.0,298.257223563]],
PRIMEM["Greenwich",0.0],
UNIT["Degree",0.0174532925199433]],
PROJECTION["Lambert_Conformal_Conic"],
PARAMETER["False_Easting",0.0],
PARAMETER["False_Northing",0.0],
PARAMETER["Central_Meridian",25.0],
PARAMETER["Standard_Parallel_1",35.0],
PARAMETER["Standard_Parallel_2",65.0],
PARAMETER["Latitude_Of_Origin",52.0],
UNIT["Meter",1.0]]

Australia

Projected Coordinate System: GDA_1994_Geoscience_Australia_Lambert
Projection: Lambert_Conformal_Conic
False_Easting: 0.00000000
False_Northing: 0.00000000
Central_Meridian: 134.00000000
Standard_Parallel_1: -18.00000000
Standard_Parallel_2: -36.00000000
Latitude_Of_Origin: 0.00000000
Linear Unit: Meter
Geographic Coordinate System: GCS_GDA_1994
Datum: D_GDA_1994
Prime Meridian: Greenwich
Angular Unit: Degree

North America

Projected Coordinate System: North_America_Lambert_Conformal_Conic
Projection: Lambert_Conformal_Conic
False_Easting: 0.00000000
False_Northing: 0.00000000
Central_Meridian: -96.00000000
Standard_Parallel_1: 20.00000000
Standard_Parallel_2: 60.00000000
Latitude_Of_Origin: 40.00000000
Linear Unit: Meter
Geographic Coordinate System: GCS_North_American_1983
Datum: D_North_American_1983
Prime Meridian: Greenwich
Angular Unit: Degree

Summary of expected areas of interest

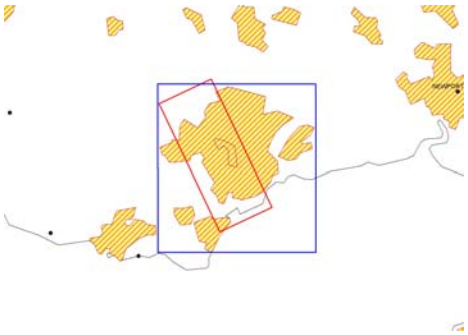
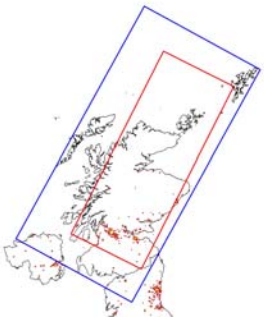
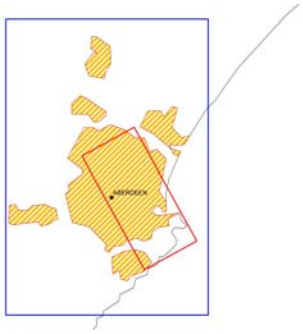
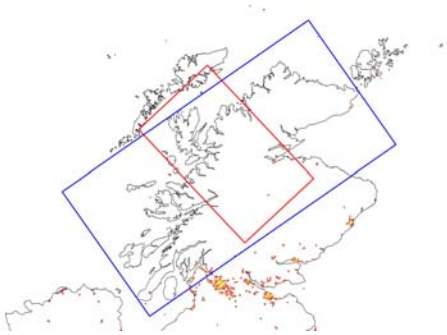
Query	*Centroid Lambert*	*Area (km^2)*	*X-Dist (km)*	*Y-Dist (km)*
museums in cardiff united kingdom	-1845040 / 300710.9	160.6	12.3	13.1
hotels in cardiff united kingdom	-1849277 / 300399.2	284.5	20.8	13.7
mountaineering in scotland united kingdom	-1661637 / 968224.9	241577.5	346.6	696.9
oil industry in aberdeen united kingdom	-1560882 / 845531.5	203.3	11.8	17.3
camping in Highland,Scotland,United Kingdom	-1719975 / 915234.8	78542.4	199.7	393.3
beaches in cornwall united kingdom	-1994806 / 228159.0	8048.6	127.4	63.2
walking in fife united kingdom	-1658319 / 773957.6	2445.7	78.7	31.1
pubs in edinburgh united kingdom	-1666241 / 747165.6	207.6	18.3	11.3
shipping in liverpool united kingdom	-	-	-	-
schools in norwich united kingdom	-1525107 / 315798.7	81.2	9.8	8.3
climbing near aviemore united kingdom	-1656574 / 871939.8	7734.1	100.7	76.8
camping near Lancaster,Lancashire,United Kingdom	-1701105 / 575916.1	6365.8	68.5	93.1
hotels near edale united kingdom	-1686318 / 458771.4	2054.9	39.9	51.7
walking near beauly united kingdom	-1673715 / 928869.9	4552.7	66.7	68.4
canals near stroud united kingdom	-1776418 / 300443.3	1734.5	43.5	39.9
red kites near Cromarty,Upper Ythan,United Kingdom	-1649454 / 938581	2432	59.6	40.8
walking outside edinburgh united kingdom	-1667943 / 741523.2	1812	54.4	37.1
cycling south london united kingdom	-1658948 / 189415.2	4257.5	110.5	38.7
castles east edinburgh united kingdom	-1636078 / 729194.3	2437.3	57	42.8
hotels west fort william united kingdom	-1783865 / 881720.1	7130.5	96.4	73.9

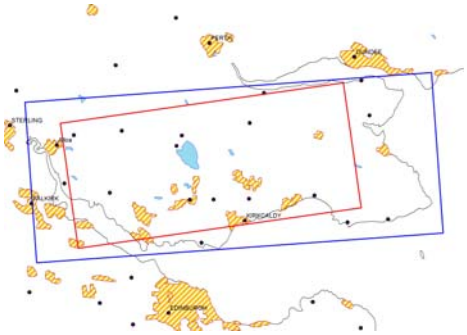
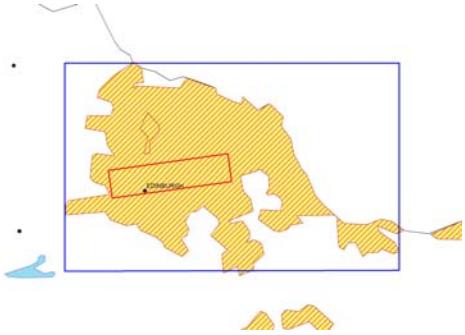
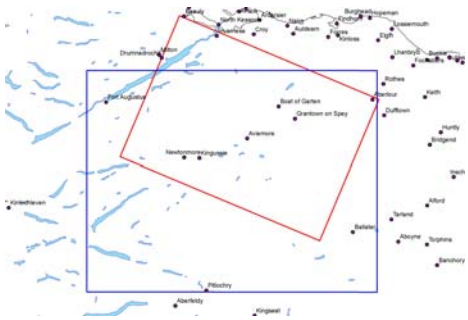
Spatial relations between expected and result areas

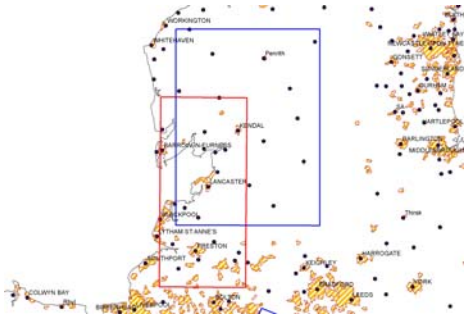
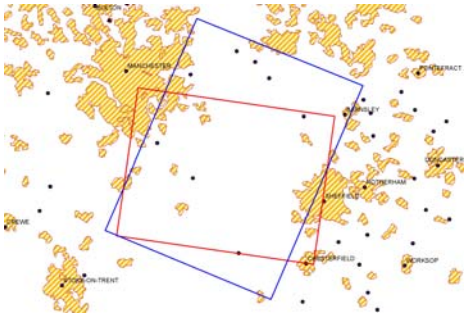
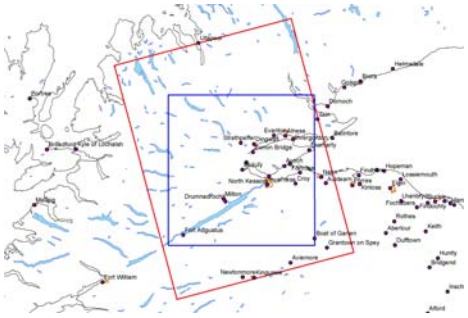
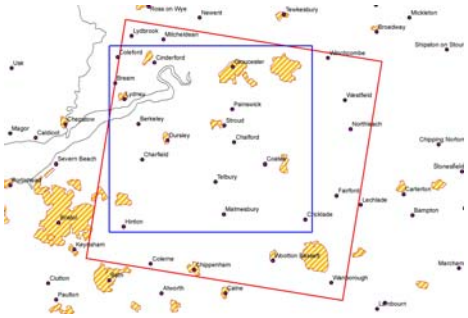
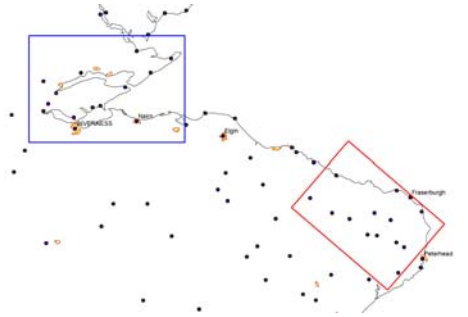
Query	*Result Area / Expected Area (%)*	*Intersection Area (km^2)*	*Intersection Area / Total Covered Area (%)*
museums in cardiff united kingdom	30.75	49.2	30.59
hotels in cardiff united kingdom	34.31	95.9	33.49
mountaineering in scotland united kingdom	45.77	110564.2	45.77
oil industry in aberdeen united kingdom	12.88	26.2	12.88
camping in Highland,Scotland,United Kingdom	32.00	22842.8	28.26
beaches in cornwall united kingdom	57.56	4632.7	57.56
walking in fife united kingdom	54.93	1343.4	54.93
pubs in edinburgh united kingdom	4.78	9.9	4.78
shipping in liverpool united kingdom	-	-	-
schools in norwich united kingdom	37.99	30.8	37.98
climbing near aviemore united kingdom	51.52	3477.5	42.20
camping near Lancaster,Lancashire,United Kingdom	57.92	2034.5	25.37
hotels near edale united kingdom	71.42	1337.1	61.18
walking near beauly united kingdom	201.93	4550.1	49.48
canals near stroud united kingdom	166.80	1712.6	58.75
red kites near Cromarty,Upper Ythan,United Kingdom	66.47	0.00	0.00
walking outside edinburgh united kingdom	79.29	945.3	41.04
cycling south london united kingdom	91.32	3061.3	60.21
castles east edinburgh united kingdom	25.24	317.5	11.61
hotels west fort william united kingdom	2.21	157.4	2.21

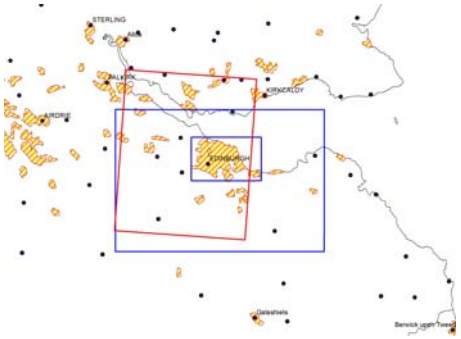
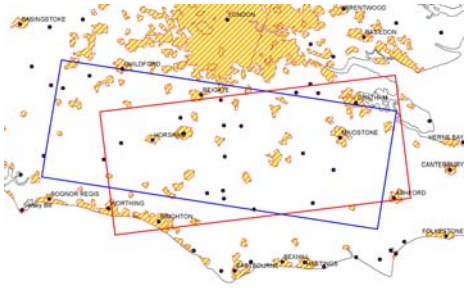
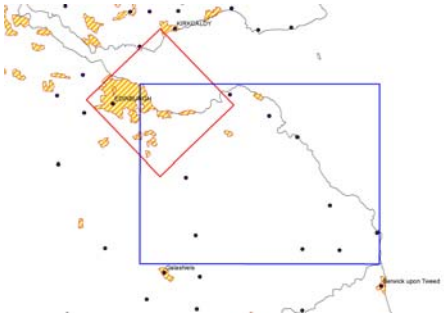
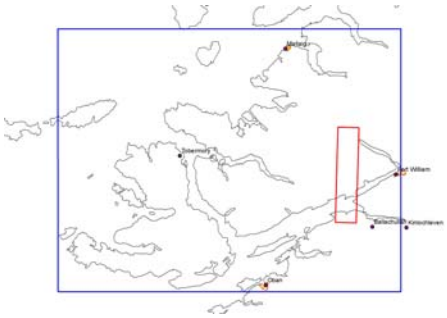
Summary of cluster analysis values

Query	*Average Density (Points per km^2)*	*Clark Evans Agg. Index*	*K-S X*	*K-S Y*
museums in cardiff united kingdom	0.466000	0.18	Sig.	Sig.
hotels in cardiff united kingdom	0.204000	0.99	Not Sig.	Sig.
mountaineering in scotland united kingdom	0.000452	0.48	Sig.	Sig.
oil industry in aberdeen united kingdom	0.458000	1.18	Sig.	Sig.
camping in Highland,Scotland,United Kingdom	0.000836	0.38	Not Sig.	Not Sig.
beaches in cornwall united kingdom	0.011000	0.35	Not Sig.	Sig.
walking in fife united kingdom	0.010400	0.95	Not Sig.	Sig.
pubs in edinburgh united kingdom	1.210000	0.89	Sig.	Sig.
shipping in liverpool united kingdom	-	-	-	-
schools in norwich united kingdom	0.681000	1.12	Sig.	Sig.
climbing near aviemore united kingdom	0.005250	0.78	Not Sig.	Not Sig.
camping near Lancaster,Lancashire,United Kingdom	0.005420	0.81	Sig.	Not Sig.
hotels near edale united kingdom	0.014300	0.94	Sig.	Not Sig.
walking near beauly united kingdom	0.005110	0.67	Not Sig.	Not Sig.
canals near stroud united kingdom	0.010400	0.89	Sig.	Sig.
red kites near Cromarty,Upper Ythan,United Kingdom	0.006800	0.96	Not Sig.	Not Sig.
walking outside edinburgh united kingdom	0.013200	0.54	Sig.	Not Sig.
cycling south london united kingdom	0.003840	1.13	Not Sig.	Not Sig.
castles east edinburgh united kingdom	0.032500	0.34	Not Sig.	Sig.
hotels west fort william united kingdom	0.038100	0.51	Sig.	Not Sig.

Query		Intersecting	Non-intersecting	
Museums in Cardiff United Kingdom	***Result Area***	99.61 %	0.39 %	
	Expected Area	30.63 %	69.37 %	
	Topology	Intersect		
Hotels in Cardiff United Kingdom	***Result Area***	98.22 %	1.78 %	
	Expected Area	33.69 %	66.31 %	
	Topology	Intersect		
Mountaineering in Scotland United Kingdom	***Result Area***	100 %	0 %	
	Expected Area	45.77 %	54.23 %	
	Topology	Inside		
Oil industry in Aberdeen United Kingdom	***Result Area***	100 %	0 %	
	Expected Area	12.88 %	87.12 %	
	Topology	Inside		
Camping in Highland, Scotland, United Kingdom	***Result Area***	90.89 %	9.11 %	
	Expected Area	29.08 %	70.92 %	
	Topology	Intersect		

Query		Intersecting	Non-intersecting	
Beaches in Cornwall United Kingdom	Result Area	100 %	0 %	
	Expected Area	57.56 %	42.44 %	
	Topology	Inside		
Walking in Fife United Kingdom	Result Area	100 %	0 %	
	Expected Area	54.93 %	45.07 %	
	Topology	Inside		
Pubs in Edinburgh United Kingdom	Result Area	100 %	0 %	
	Expected Area	4.78 %	95.22 %	
	Topology	Inside		
Schools in Norwich United Kingdom	Result Area	99.97 %	0.03 %	
	Expected Area	37.98 %	62.02 %	
	Topology	Intersect		
Climbing near Aviemore United Kingdom	Result Area	87.27 %	12.73 %	
	Expected Area	44.96 %	55.04 %	
	Topology	Intersect		

Query		Intersecting	Non-intersecting	
Camping near Lancaster, Lancashire, United Kingdom	*Result Area*	55.18 %	31.96 %	
	Expected Area	44.82 %	68.04 %	
	Topology	Intersect		
Hotels near Edale United Kingdom	*Result Area*	91.11 %	8.89 %	
	Expected Area	65.07 %	34.93 %	
	Topology	Intersect		
Walking near Beauly United Kingdom	*Result Area*	49.49 %	50.51 %	
	Expected Area	99.94 %	0.06 %	
	Topology	Intersect		
Canals near Stroud United Kingdom	*Result Area*	59.20 %	40.80 %	
	Expected Area	98.74 %	1.26 %	
	Topology	Intersect		
Red kites near Cromarty, Upper Ythan, United Kingdom	*Result Area*	0 %	100 %	
	Expected Area	0 %	100 %	
	Topology	Disjoint		

Query		Intersecting	Non-intersecting	
Walking outside Edinburgh United Kingdom	***Result Area***	65.80 %	34.20 %	
	Expected Area	52.17 %	47.83 %	
	Topology	Intersect		
Cycling south London United Kingdom	***Result Area***	78.73 %	21.27 %	
	Expected Area	71.90 %	28.10 %	
	Topology	Intersect		
Castles east Edinburgh United Kingdom	***Result Area***	51.61 %	48.39 %	
	Expected Area	13.02 %	86.98 %	
	Topology	Intersect		
Hotels west Fort William United Kingdom	***Result Area***	100 %	0 %	
	Expected Area	2.21 %	97.79 %	
	Topology	Inside		

Mann-Whitney U-Tests

Result area intersecting with expected area; “in-queries” versus other queries

Ranks

	queryTypeNum	N	Mean Rank	Sum of Ranks
resultIntersecting	0	9	14.17	127.50
	1	10	6.25	62.50
	Total	19		

Test Statistics(b)

	resultIntersecting
Mann-Whitney U	7.500
Wilcoxon W	62.500
Z	-3.110
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.001(a)

a Not corrected for ties.
b Grouping Variable: queryTypeNum

Expected area intersecting with result area; “in-queries” versus other queries

Ranks

	queryTypeNum	N	Mean Rank	Sum of Ranks
expectedIntersecting	0	9	8.56	77.00
	1	10	11.30	113.00
	Total	19		

Test Statistics(b)

	expectedIntersecting
Mann-Whitney U	32.000
Wilcoxon W	77.000
Z	-1.061
Asymp. Sig. (2-tailed)	.288
Exact Sig. [2*(1-tailed Sig.)]	.315(a)

a Not corrected for ties.
b Grouping Variable: queryTypeNum

Intersection Area / Total Covered Area; result topology, intersect versus inside

Ranks

	topologyNum	N	Mean Rank	Sum of Ranks
interTCA	0	12	10.33	124.00
	1	6	7.83	47.00
	Total	18		

Test Statistics(b)

	interTCA
Mann-Whitney U	26.000
Wilcoxon W	47.000
Z	-.937
Asymp. Sig. (2-tailed)	.349
Exact Sig. [2*(1-tailed Sig.)]	.385(a)

a Not corrected for ties.
b Grouping Variable: topologyNum