Strategies for gazetteer improvement and enrichment

by

Sanket Kumar Singh

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Many applications that use geographical databases (a.k.a. gazetteers) rely on the accuracy of the information in the database. However, poor data quality is an issue in gazetteers; often data is integrated from multiple sources with different quality constraints and there may not be much detail on the sources and the quality of the data. One major consequence of this is that the geographical scope of a location and/or its position may not be known or accurate.

In this thesis, we develop novel strategies to accurately derive the geographical scope of places. Our strategies use the spatial hierarchy of a gazetteer as well as other public information (such as area) to construct a bounding box for each place. We present a probabilistic model of our approach and demonstrate the effectiveness of the bounding boxes in refining the spatial hierarchy of a gazetteer and augmenting it with other public data. Experimental evaluation on two public-domain gazetteers show that the proposed approaches significantly outperform, in terms of the accuracy of the bounding boxes, a baseline that is based on the parent-child relationship of a gazetteer. More specifically, our approaches outperform the baseline by 19-33% in terms of accuracy in a wide range of settings. Among applications, we show how these bounding boxes provide a generic way to improve the accuracy and usability of a gazetteer.

Preface

A part of subsection 4.2 of Chapter 4 of this thesis has been previously submitted and accepted in the *MediaEval 2016 Workshop* [43]. This is a collaborative work between myself and Dr. Davood Rafiei. While I performed the experiments, analyzed the results and made conclusions, Dr. Rafiei provided feedback on improving the approach, design of the experiments and presentation of the paper.

To my family,

for supporting and believing in me at each step.

The proper method for inquiring after the properties of things is to deduce them from experiments.

– Isaac Newton.

Acknowledgements

I would like to thank my supervisor, Dr. Davood Rafiei for guiding and supporting me through out this work. This journey would not have been possible with out his constructive feedback, which has helped me to enhance my research skills. His efficacious remarks has kept me motivated to think differently and to find generic solution to the problems.

I am thankful to the members of the Database Research group, especially Ehsan Kamalloo, with whom I had several insightful discussions on various topics. I learned immensely from the knowledge sharing sessions and it also helped me to connect my work with other related ongoing research work.

I would also like to thank my friends, Sankalp, Vivek, Gautham, Megha, Soham and Abhimanyu for being a source of moral and technological support. I enjoyed our conversations on different subject varying from technological advancements to graduate life in general.

Most importantly, I would like to thank my mother and father for being a source of inspiration and motivating me to pursue my dreams. Without their support and encouragement, I would not have come here. Furthermore, I am grateful to my elder sister and brother for their love, care and inspiriting me to never give up.

Last but not the least, I want to thank Cybera (www.cybera.ca), NSERC (Natural Sciences and Engineering Research Council of Canada), GeoNames (www.geonames.org) and OSMNames (www.osmnames.org) for their support on this study.

Table of Contents

List of Tables

List of Figures

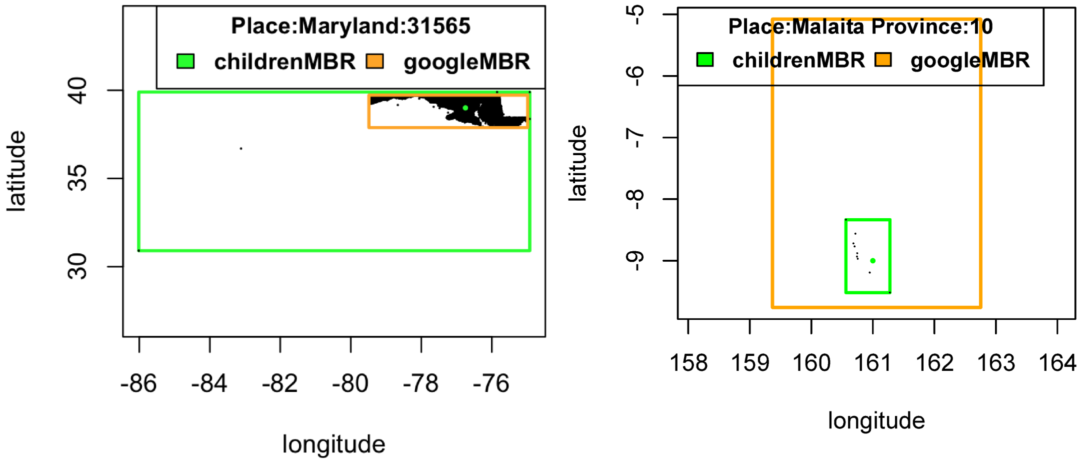
Chapter 1

Introduction

Online world geographical directories (a.k.a. gazetteers) contain information of places such as latitude and longitude, alternative names, country or province to which a place belongs and other geographical features. Gazetteers are extensively used in many different domains and applications because of their wide coverage and detailed information about places. For example, an incoming tweet may have the GPS coordinates of the capturing device, but to detect a populated place the tweet is coming from, one may use a gazetteer such as GeoNames [47] to map those coordinates to an actual location entity. With an increasing number of different capturing devices equipped with GPS (e.g. phones, cars, cameras, etc.), we often have the coordinates of an entity and may want to find an administrative location or a populated place in which the entity is located. Such a query can be important, for example, in dispatching services such as ambulance, police, etc. The literature also reports numerous domains where gazetteers are used, including toponym resolution in text [14], geotagging tweets [51], documents [8] and entities [36], etc. To support many of these applications, one needs to both effectively and efficiently join a gazetteer with other geo-coded data.

However, there are a few challenges that hinder progress in this area: (1) most public gazetteers either do not have bounding boxes for many of their locations (e.g. GeoNames) or their bounding boxes are not accurate (e.g. OSMNames[1], see Section 5.1.3 for details). In the absence of a bounding box,

[1] http://osmnames.org/



(a) Presence of outliers. **(b) Incomplete coverage.**

Figure 1.1: Examples of inconsistencies between gazetteer and other sources (childrenMBR denotes the bounding box formed using the locations in GeoNames based on the containment relationships between places and googleMBR represents the bounding box retrieved using Google Maps for respective location. The number beside the place name is the child count. Green dot represents the given center.)

there is no direct way of checking if an entity falls inside or outside a region boundary[2], and as a result, applications end up implementing their own ad-hoc solutions; (2) data in a gazetteer is prepared by public and is not necessarily accurate especially for under-populated areas; (3) there are inconsistencies within gazetteers and in relationship with other sources (see Fig. 1.1).

Our approach to address those challenges is through creating bounding boxes for places. Attaching a bounding box to each place has a number of benefits, including better support for reverse geo-coding queries, which can be efficiently answered. Also, the consistency of the data can be better monitored and enforced in the form of relationships between bounding boxes; this can reduce inconsistency and improve accuracy. However, creating a bounding box is not a one-time process; boundaries change due to growths, splits and mergers, and as a result, bounding boxes need to be updated on regular basis.

[2] Reverse Geocoding API from Google and other search engines convert the coordinates of a point on the map to a human readable location or address, but the details of their properitary solutions often are not made public.

The problem to be studied in this work is if a bounding box can be accurately constructed for each place based on incomplete and sometimes erroneous information that is available. We take, as a bounding box of a place in a gazetteer, the minimum bounding rectangle (MBR) that can contain MBRs of all locations which are geographically part of it; we refer such relationship between places in the gazetteer as the parent-child constraints of the gazetteer. Despite their imprecisions in some cases compared to, for example, polygons, MBRs provide a simple abstraction of the boundaries of a location, and they have been frequently used to represent sets of multi-dimensional points and objects in spatial querying and indexing [3]; also, constraints can be easily specified and enforced using MBRs [34]. In the remainder of this thesis, the terms "MBR", "minimum bounding rectangle" and "bounding box" are often used interchangeably.

Sometimes, the stated constraints cannot all be satisfied when creating MBRs. We formalize the search for an MBR as a probabilistic optimization, which tries to find the most likely MBR by dropping the least likely constraints. Our contributions can be summarized as follows:

(1) We provide a systematic study of the problem of enriching and improving a gazetteer using bounding boxes of places. To the best of our knowledge, this is the first time such a study is conducted by approximating bounding boxes of the places.

(2) We propose strategies for detecting and resolving inconsistencies in the gazetteer.

(3) We evaluate our strategies and report their accuracy in detecting the boundaries of places and in improving the hierarchy of places.

(4) We report on the effectiveness of our bounding boxes in refining the places hierarchy and in augmenting the gazetteer with other data sources including YFCC100M [46].

Chapter 2

Background And Related Work

2.1 Background

In this section, we present some background material to provide context for the topics we are presenting in this thesis. In particular, we present different spatial footprints that are used to represent spatial objects and some rationals on using a Minimum Bounding Rectangle to represent a place or set of places in a gazetteer. We also review gazetteers and some of their applications as well as the presence of uncertainty and inconsistencies in them.

Spatial abstractions of locations Applications use different spatial representations for the geographical entities which are usually multi-dimensional. Spatial footprints that are used in modelling spatial relationships between such entities are mainly point, polygon, minimum bounding rectangle [16], minimum bounding ellipse [28], minimum bounding circle [5] and convex hull [25].

Chen et al. [12] provides a survey of qualitative spatial representations where they discuss about convex hull as a powerful primitive in establishing the topological relationship between spatial objects. However, construction of an accurate convex hull (with smooth curves) in the absence of enough points can be difficult. Furthermore, Frontiera et al. [16] have shown that a logistic regression model to compute similarity between query-document pairs using a minimum bounding box produces better results than a non-probabilistic model using convex hull. The authors use minimum bounding rectangle and convex hull as geometric approximations to spatially index the queries and the documents.

Cobb et al. [13] uses a graph model to represent the relationships between spatial objects; the authors divide the MBR of a place to multiple smaller rectangles and operate over these smaller regions to obtain a better representational accuracy.

Importance of Minimum bounding rectangles MBRs are widely used to approximate the spatial scope of entities due to its computational efficiency and ease of storage. Detecting bounding boxes of places have various applications in geo-referencing services such as HERE [1], Esri[2], and in social media websites such as Twitter[3]. For example, the minimum bounding box of a city can be used to answer which administrative buildings are available in the city and can provide precise geo-coordinates for the same. In addition, the geographical scope of a place has been found useful in solving complex problems in geographical information science. Papadias et al. [30] define 169 topological relations between MBRs of places to answer spatial queries on whether one place contains, overlaps, covers or meets another place. MBRs of spatial objects have been used in building spatial indexes using data structures such as R-tree [20] and R$^+$-tree [40]. Brisboa et al. [6] have further used MBRs to answer interval based queries by efficiently encoding MBRs into a rank-grid space.

Gazetteers As an open online geo-resource, gazetteers are often created by merging geographical databases of countries (e.g. GeoBase in Canada, GNIS in USA, etc., which may be maintained by respective government agencies) and using place information from different sources such as travel advisories and blogs. A dataset provided by an official source or governing body may be more accurate especially if the data is collected and maintained by experts. However, other sources including volunteered geographical information and local place gazetteers contributed by public are generally less precise.

Gazetteers can be augmented with other rich knowledge bases. They can

[1] https://here.com/en

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

[3] https://twitter.com/?lang=en

be expanded horizontally, by adding more features or vertically, by adding more places [42]. One general approach used to enrich gazetteers is to detect a novel place name and search information for other features from external sources. The tuple with complete entity information is then added to a gazetteer. There are different methods developed to detect a place name including fuzzy string matching [32], toponym detection from tweets [29] and named entity recognition [17].

Presence of uncertainties and inconsistencies Uncertainty in geographical information can be viewed as a difference in precision, scale and resolution of the feature values with respect to a master data. Some of the reasons for this uncertainty are variations in measuring instruments, data transformations, digitization errors, fuzziness of geographic concepts, etc. (Zhang et al. [50]). Another source of ambiguity that one may notice is the use of different geographical ontologies across different gazetteers. For example, water bodies are classified as peat bog, water course and water body in CORINE land cover while it is classified as canal, lake, bog, pond etc. in WordNet (Laurini et al. [26]).

Different datasets may provide similar information but with varying accuracy as one feature may be more relevant to one source than the other. Thus, one approach to obtain a master record is extracting values of features from different datasets based on a trust score of each source provided by data stewards. One may perhaps build a new dataset by merging data from different sources. For example, the Open Street Map gazetteer contains precise information for London city (≈6 meter distance [18]) but it doesn't contain spatial hierarchy of places as in GeoNames. However, using data from different sources of varying reliability while modelling the task at hand is difficult. Few of the work done to model the uncertainty in spatial datasets are probabilistic skyline [31] and probabilistic spatial queries over existentially uncertain objects [15].

Online free gazetteers such as GeoNames are publicly maintained and therefore the presence of error is inevitable. Dirk [1] has explored different inconsistencies in GeoNames by focusing on the data related to the countries in Central America. The author reports a loss of granularity in the coordinates of places in converting the coordinates from a decimal system to a degree-minute-second (DMS) format, and they find that often the field ‘second’ is missing. Such truncated coordinate is mostly observed for underdeveloped and developing countries. The author also notes that there are places with wrong feature codes. For example, Florencia is assigned feature code of PPL instead of PPLX even though it is in a neighbourhood of Honduras.

Bolstand et al. [4] identify the inaccuracy in the way data is collected, stored, formatted and integrated as the main reason for inconsistency in a spatial data. Generally, spatial data extracted from satellite images, digitization of old maps and crowd sourcing follows different formats. Merging such data from different sources often leads to a loss of precision. Ahlers et al. in [2] find anomalies in online geocoders and propose combining multiple geocoders to determine the coordinate of a textual address.

2.2 Related Work

The literature related to our work can be grouped into (1) estimation of the spatial extent of geographic entities, (2) conflict resolution and data cleansing techniques, and (3) automatic gazetteer expansion and enrichment. This section discusses the work in each area along with the limitations and some of the relationships to our work.

Estimation of the spatial extent of geographic entities The proposed methods estimate the spatial extent of places in a gazetteer and use the bounding boxes to improve and enrich the gazetteer. Hence the literature on obtaining a bounding box of a place is relevant.

Chen et al. [11] develop a method to find the spatial extent of places with vague boundaries. They define the geographical boundary of a place using the density of images mapped to a region, therefore places with less images may have undefined or vague boundaries. In their approach, the authors first determine a set of clean points, which are not isolated and are within a cluster of 95% of images, then they interpolate the boundary for remaining points using Kernel Density Estimation. A limitation of their approach is that if a place is widely spread and have disjoint regions (e.g. places containing islands such as Hawaii), then each part forms its own boundary instead of forming a boundary at a particular level such as country, province or district. While the authors of paper used images from Flickr to estimate the boundaries, we use a set of spatial points for a given location to construct the MBR with no limitation on the landscape of each place.

Somodevilla et al. [44] design a fuzzy set approach to model the spatial extent of a geographical location using spatial, thematic and temporal reasoning. They use point locations and build an inscribed rectangle, which is the maximum rectangle inside the location, and a fuzzy minimum bounding rectangle that includes all points for a place. The region between fuzzy minimum bounding rectangle and inscribed rectangle is considered as a fuzzy region and a final MBR is approximated based on membership value of points in the fuzzy region. While the authors use Euclidean distance between points and the nearest edge of fuzzy MBR to include a point and expand MBR, we present an efficient algorithm to include/exclude points based on probability measures.

The geographic boundary of a place can also be estimated using different geotagged entities such as location specific tags [22], geotagged images, and online documents. Although there are some work to determine an MBR using above approaches, little attention has been given to model the spatial extent using a probabilistic approach. The bottleneck that makes this task challenging is the uncertainty and incompleteness of the spatial information for locations in geographical databases, especially when it is created and maintained publicly.

Conflict resolution and data cleansing techniques This line of research is relevant since it can benefit from the bounding boxes of places (as shown in Section 5.2.1), hence we briefly review it here.

Conflict resolution can be viewed as making a decision between different versions of data to determine a golden record. Different techniques have been developed to cleanse redundant information including rule-based approaches, heuristics and quantitative analysis. Prokoshyna et al. [35] combine logical reasoning and quantitative method to develop a novel data cleansing approach. This quantitative method involves setting some constraints based on the statistical properties of attribute values and flagging inconsistencies if such constraints are violated. The authors propose a minimal-set repair algorithm to find attribute values that minimize a statistical distortion. Their work inspired us in defining the topological constraints (see Section 4.3), more precisely soft constraints, by determining the statistical property of overlap of MBRs.

Volha et al. [7] propose a framework for resolving conflicts by incorporating a learning-based algorithm, which learns a fusion function from a set of functions for each attribute type. In particular, if attributes are numeric, then the authors select a function which (1) minimizes the error between the given and the true values of the attributes and (2) maximizes the count of attributes such that their value does not deviate from the gold standard value by more than a threshold. Similarly, for text fields, the authors maximize the number of matches with a gold standard value. While the authors of the paper use a supervised learning approach to resolve conflicts, we rely on the statistical property of data in the gazetteer to build a probabilistic model which selects an optimal MBR of a place from different possible MBRs.

Another line of work to filter objects which do not follow the general underlying distribution of a dataset are outlier detection methods. A commonly used technique in spatial domains is Boxplot [21], in which all points which lie outside a boundary range are considered as outliers. This boundary is determined based on statistical properties of the dataset. More precisely, one can find the first and the third quartile of the distribution and determine the boundaries on either side by multiplying the first and third quartile by some factor (usually 1.5). Rousseeuw et al. [39] further propose a bivariate generalization of Boxplot, known as Bagplot to determine outliers in a set of 2-D points. In one of the proposed heuristic approaches, we also experiment with Boxplot and Bagplot to remove outlier locations while constructing an MBR of a place.

Automatic gazetteer expansion and enrichment This line of research can also benefit from the bounding boxes of places; for the same reason, it is briefly reviewed here.

Automatic gazetteer creation or enrichment involves adding places to a gazetteer to make it complete or adding new features to make it more useful. It is a challenging task as it requires one to efficiently merge data from heterogeneous sources, each with its own different storage format and representation. Popescu et al. [32] devise an approach to automatically create a gazetteer using diverse information sources. Their paper provides different algorithms for entity extraction, categorization, coordinate discovery and ranking. Entity names and types are extracted from Wikipedia pages as well as any latitude and longitude information when present. For entities with no latitude and longitude in Wikipedia, geotagged images from Panoramia are used to find a geo-coordinate. The authors use the AllTheWeb search engine to determine the rank of each entity based on the count of pages returned for queries formed using the entity name. This research provides an effective way to combine different data sources and their evaluation shows high precision (more than 90%) for entity extraction and categorization.

Recently, Oliveira et al. [29] also attempt to enrich the gazetteer included in GeoSEn [10] system using its geo-parser that utilizes the volunteered geographical information features. The authors augment the spatial hierarchy of the gazetteer by adding places at the level of granularity which is similar to district and streets. In our work, we refine the hierarchy of a gazetteer by moving places deep in the hierarchy based on the containment relationships between MBRs.

One application in which we demonstrate the usage of MBR is geotagging photos and videos from Flickr, which can then be used to enrich places in gazetteers with location specific photos. Related work in the literature includes the work of Serdyukov et al. [41] which proposes a language model to map photos from Flickr to places on earth using user tags attached to images. The authors use a grid based approach to divide the earth surface into cells of equal sizes before predicting a particular cell for each photo. They apply several smoothing techniques based on neighbouring places and location specific tags to refine the cell prediction. This line of work may benefit from a more descriptive geometric structure such as the MBR of places instead of using cells with arbitrary boundaries.

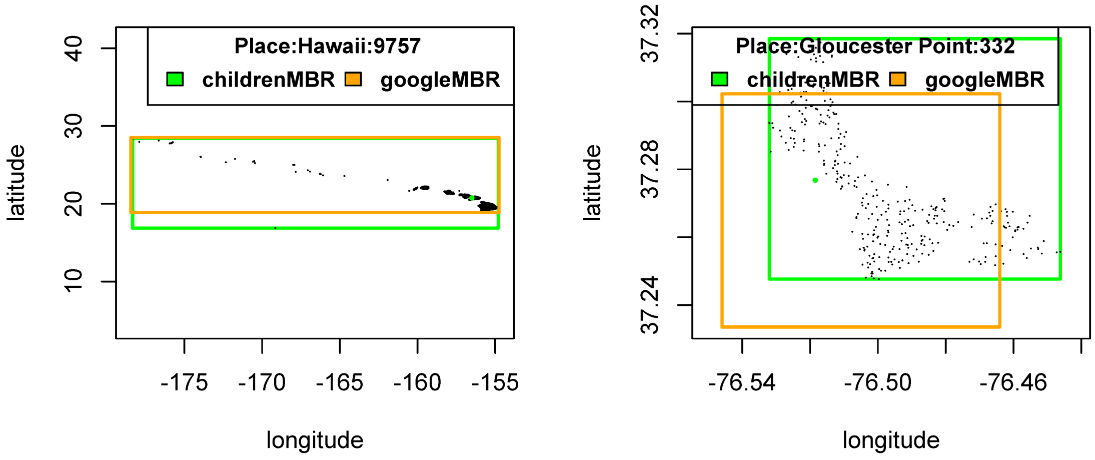
Kordopatis et al. [24] propose a bag of tags approach where they determine the probability of a tag being used by users to describe a region. They weigh the tags in each cell based on spatial entropy which gives less weight to tags that are either user specific or very general. We show that mapping images from Flickr to locations in GeoNames, based on MBRs of places instead of cells, can be an effective alternative.

Chapter 3

Bounding Box Construction

Given a set of containment relationships between places in a gazetteer, our objective is to construct a bounding box for every place such that ideally all stated or known constraints are satisfied. We take as the bounding box of a place, any minimum bounding rectangle (MBR) that is parallel to latitude and longitude axes and satisfy the constraints. This does not always give the most accurate bounding box especially if the true bounding box is not convex; however, it is easier to work with MBRs (rather than arbitrary polygons) for checking containment relationships. A problem here is that there are often relationships or constraints in a gazetteer that are conflicting or contradictory and they cannot be all satisfied. For example, in Figure 3.1(a), it can be observed from the bounding box of Hawaii, it is difficult to minimize the distance between center of the MBR and the given center as the place consists of disjoint islands. Also, in Figure 3.1(b), one can notice from a bounding box returned by Google Maps (googleMBR) that there are no locations in southwest region and it is difficult to conflate the bounding box based on locations (childrenMBR) without using any external information.

We present two basic strategies to construct a bounding box: (1) a Hierarchical approach or MBR of the children, which uses the set of containment relationships expressed in the spatial hierarchy of a gazetteer, and (2) a Geometric approach or MBR of the center point, which constructs an MBR using both the center point and the area information about each place. Furthermore, we present a probabilistic approach that fuses two strategies and



(a) Disjoint group of child locations.

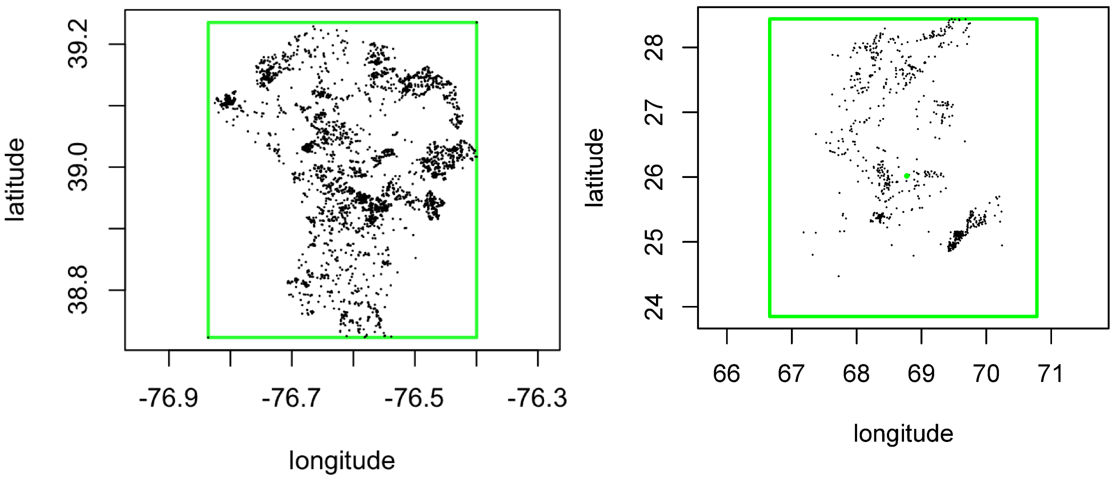
(b) No data in the south-west.

Figure 3.1: Examples of places with uncertainty in spatial data. Green dot represents the given center of the place.

models the construction of a bounding box as a constraint optimization problem and a heuristic approach that employs outlier detection technique and scaling/enlargement operations.

3.1 Hierarchical Approach

One strategy to build an MBR is based on the containment relationship, enforcing that each parent MBR must contain the MBRs of its children. Gazetteers are good at describing the containment relationships between places. For example, GeoNames places each location into an administrative level such as country and state and allows queries to retrieve the children within an administrative level. A major problem with this strategy is that the number of children can vary greatly for different places. For example, more populated places tend to have more children than rural towns, water bodies, natural regions, etc. Another problem is the distribution of the children and that the children are not always spread over the whole region boundary (Fig. 1.1 (b) and 3.1 (b)).



(a) Anne Arundel County (child count = 3161

(b) Sindh Province (child count = 793)

Figure 3.2: ChildrenMBRs for places at different administrative level.

We refer to the MBR of the children as childrenMBR; this is formed using the children locations of a given place in a gazetteer. It is also our baseline approach. The bounds of a childrenMBR for a place is calculated in a bottom-up approach by taking the minimum of south-west coordinates and the maximum of north-east coordinates of all children of the place in the spatial hierarchy. ChildrenMBR for two locations, Anne Arundel County (Maryland, US) and Sindh (a province in Pakistan), are plotted in Fig. 3.2.

3.2 Geometric Approach

A major drawback of obtaining a bounding box using spatial points is that for many places, there are no geographical extent if the gazetteer does not contain any children. Such places are represented by a point or a small rectangle. For example, the area of childrenMBR for Malaita Province constructed from points in GeoNames (as shown in Fig. 1.1(b)) is approximately 19 times smaller than the area of the bounding box obtained from Google Maps. Our next strategy, referred to as centerMBR, handles this by constructing an MBR for each place that is centered at the center point of the place and with an area close to the known area of the MBR of the place.

3.2.1 Center MBR

Given a center point c and area a, one can construct an infinite number of rectangles centered at c with an area a. Without additional information, we don't know which rectangle is more likely. However, our next statement gives some evidence that maybe a square is a better choice.

Conjecture*: Let R be the set of all rectangles with a center point c and area a and $r \in R$. Assuming that all rectangles are equally likely, the expected area of overlap between R and r is maximized when r is a square.*

Therefore, we construct a square of given area A from the given center C of the place. The bounds of a MBR can be obtained by shifting the latitude and the longitude of the center in either directions. The latitudes of north-east (NE$_{lat}$) and south-west (SW$_{lat}$) points are obtained by shifting the latitude of the center in north and south by a factor F given as

$$\mathbf{NE_{lat}} = C_{lat} + F \quad \text{and} \quad \mathbf{SW_{lat}} = C_{lat} - F$$

where $F = (\sqrt{A})/(2 * L)$, L is the distance between two consecutive latitude ($\approx$ 111 km) and C_{lat} and C_{long} are the latitude and longitude of center C.

To calculate the longitude of the two end points, we first obtain the distance between two longitudes at a given latitude (D_{lat}). This is required because the distance between consecutive longitudes shrinks as we move toward the poles. The longitudes of the end points is then obtained by shifting the longitude of the center point by shift factors F_{ne} and F_{sw} which are calculated as below:

$$F_{ne} = (\sqrt{A})/(2 * D_{nelat})$$

$$\text{where} \quad D_{nelat} = L * cos(nelat_{radian}) \text{ (from [45]) and}$$

$$nelat_{radian} = (NE_{lat} * \pi)/180,$$

$$F_{sw} = (\sqrt{A})/(2 * D_{swlat})$$

$$\text{where} \quad D_{swlat} = L * cos(swlat_{radian}) \text{ (from [45]) and}$$

$$swlat_{radian} = (SW_{lat} * \pi)/180.$$

$$\therefore \quad \mathbf{NE_{long}} = C_{long} + F_{ne} \quad \text{and} \quad \mathbf{SW_{long}} = C_{long} - F_{sw}.$$

Note that the precision of coordinates of endpoints depends on the precise value of L and the given center.

The bounding box estimated using this geometric approach, also referred to as centerMBR in later sections, is expected to be accurate in cases when the child locations are distributed uniformly around the given center. However, this approach may not perform well when child locations include outliers or the center of the place is away from the mass of child locations.

3.3 Probabilistic Approach

Geographic information for a place in a gazetteer can be inconsistent, inaccurate, incomplete, obsolete or duplicate. One can apply traditional rule based methods to filter the redundant information. However, it requires experts and is time consuming. Moreover, the main challenge which still remains is to handle the uncertainty in spatial data.

In this approach, we assume a probability can be assigned to each constraint (expressed in a gazetteer) or piece of information to indicate its degree of certainty or accuracy, and model our task as a constraint optimization problem. Later, we present an efficient algorithm which uses the model and generates a bounding box of a place.

3.3.1 Model

A bounding box obtained from our baseline approach suffers from both inaccuracies and inconsistencies. For example, inaccurate center information may lead to many neighbouring places to be included or excluded in an MBR. A childrenMBR can be smaller than expected if there are either not enough points or the points are not widespread enough to form a larger bounding box.

Also, a childrenMBR can be much larger than expected if a few locations are wrongly placed as children.

Locations in a gazetteer are described by a latitude and a longitude; we assume that a given coordinate of a place represents the center of its bounding box with high certainty. Public gazetteers often do not provide much detail on how the coordinates are obtained and if the given coordinate is actually the center point of the MBR of the place. To test this, we randomly selected 1000 places each from GeoNames and OSMNames. The coordinates of these places were checked against the center point of the bounding box obtained from a different source (in our case MBR obtained from Google Maps referred to as googleMBR). We found that only 63% of places in GeoNames and about 97% of the places in OSMNames had a coordinate within 10 km of the center obtained from googleMBR.

Modeling center point: Let d_c denote the distance between a given center c of a place and its true MBR center. If we assume d_c follows the normal distribution with parameters μ and σ, then we can write

$$Pr(d_c|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(d_c-\mu)^2/2\sigma^2} \tag{3.1}$$

where μ and σ are respectively the mean and the standard deviation of d_c. Let P_{center} denote this probability for fixed values of μ and σ. The parameters of the distribution can be easily estimated from the data. In our random sample of 1000 places, μ and σ are 88.894 and 408.760 for GeoNames and 2.057 and 7.09 for OSMNames, in km, respectively.

Modeling children: Let q be the probability that an arbitrary location is placed under a correct parent in the gazetteer. The value of q can be estimated by checking for each place in a sample whether its children are assigned a correct parent node. In our sample of 1000 places, the value of q is calculated as 0.968 for GeoNames and 0.882 for OSMNames respectively. One may observe that the probability that an arbitrary location is placed under a correct parent is relatively high; hence based on this empirical result and without much additional knowledge of which places may be correct or incorrect children, it is more desirable to include rather than exclude a child inside the MBR of its listed parent. For a parent place with n children and an MBR that includes i of its children, the probability that a random child location of the parent is enclosed in the MBR ($P_{children}$) can be written as

$$P_{children} = \frac{i}{n}. \tag{3.2}$$

Putting it together: Assuming independence of the center point and the children listing, we can put together the two probabilities into an objective function. Given a place with center c, MBR area A and a set of children locations S, we want to find a set $S' \subseteq S$ of children such that

$$\begin{aligned} \underset{S' \subseteq S}{\text{argmax}} \quad & (P_{children} \cdot P_{center}) \\ \text{subject to} \quad & area(MBR(S')) \leqslant A \end{aligned} \tag{3.3}$$

where area(MBR(S')) refers to the area of MBR formed from places in S'. Everything equal, the model selects an MBR with a center point closest to the given center c. Equation 3.3 provides a way to model the inclusion and exclusion of locations under an MBR and to estimate the center of the MBR with high certainty. We refer to this approach as Probabilistic Optimization Model (POM) in our experiments.

3.3.2 Optimization

Optimizing Eq. 3.3 can be computationally intensive since one needs to consider all possible solution MBRs. The problem may be broken down into two cases: (1) there is an MBR that includes all points and rectangles in S and the area of the MBR is not larger than A, (2) there is an MBR that includes all points and rectangles in S and has an area larger than A. For (1), the MBR that includes all points in S with an area not exceeding A will maximize the first term $P_{children}$. It can be noted (from our sample of 1000 places and their given center) that there is more uncertainty in finding the center of an MBR than that in including a correct child inside an MBR, hence one may maximize the first term ($P_{children}$) before maximizing the second term (P_{center}). This means the MBR that includes all points in S can simply be expanded (if needed), moving the center of the MBR to the given center and maximizing the objective function, without violating the area constraint.

Algorithm 1 Find an optimal MBR (as per Equation 3.3) when child locations are all points.

```
procedure Maximum–Enclosing–Points
    Inputs:
    A ← area of MBR of the place P
    C ← center of the place P
    S ← set of n unique locations {$p_1 \ldots p_n$}, under P
    Initialize:
    bestMBR ← Nil
    maxProbability ← 0
    ($p_1, p_2 \ldots p_m$)← InitialSolution(S)                          ▷ Alg. 2
    ptsOutsideMBR ← ($p_1, p_2 \ldots p_m$)
    for i = 0 to m do
        for j = 0 to m-i do
            for k = 0 to m-i-j do
                for l = 0 to m-i-j-k do
                    currentMBR ← FormCandidateMBR(i,j,k,l, S) ▷ Alg. 3
                    if area(currentMBR) > A then
                        continue
                    end if
                    x ← (i+j+k+l)                      ▷ # of excluded places.
                    currCenter ← center of currentMBR
                    Calculate $P_{center}$ using C and currCenter in Eq. 3.1
                    $P_{children}$ = (n-x)/n (as in Eq. 3.2)
                    currProb ← $P_{children} \cdot P_{center}$
                    if currProb > maxProbability then
                        maxProbability ← currProb
                        bestMBR ← currentMBR
                    end if
                end for
            end for
        end for
    end for
    return bestMBR
end procedure
```

Algorithm 2 Find Initial Solution

```
procedure INITIAL–SOLUTION
    Inputs:
    S ← set of n unique locations $\{p_1 \ldots p_n\}$, under P

    Start dropping one point at a time from furthest end in each direction, north, south, east and west.
    At each exclusion, form MBR with remaining points and compare its area with A.
    Stop when area of the MBR is less than or equal to A.
    Find all the points which are strictly outside the current MBR and return it.
end procedure
```

Algorithm 3 Form Candidate MBR

```
procedure FORM–CANDIDATE–MBR
    Inputs:
    i, j, k, l ← number of points to be dropped from north, east, west, south direction.
    S ← set of n unique locations $\{p_1 \ldots p_n\}$, under P

    Using S, drop i, j, k, l points from north, east, west, south direction respectively as shown in Fig. 3.3.
    Form MBR with remaining point and return it.
end procedure
```

Now consider the case where there is no MBR that includes all data points with an area less than or equal to A. Again with a high certainty in including a correct child as compared to that in selecting a correct center, we can optimize the first term before plugging in the second term.

Naive algorithm: A naive approach to perform this optimization is to enumerate all possible MBRs and select the one that maximizes the objective function in Equation 3.3. This is equivalent to sweeping the complete search space, selecting a point at a time from each direction and forming an MBR that has the selected points on its sides. With n data points, there are n

possible choices for each side of an MBR; hence there are $\mathcal{O}(n^4)$ MBRs to choose from. Among those MBRs, the algorithm selects the MBR that maximizes the objective function. For large values of n, this is still an expensive process. Our next algorithm further prunes the search space without affecting the correctness of the result.

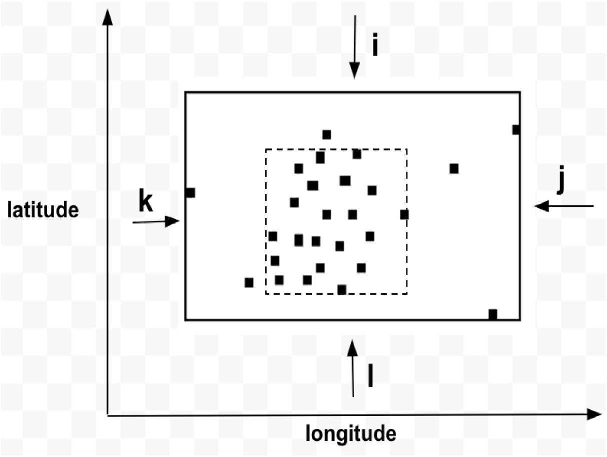


Figure 3.3: Demonstration of Algorithm 1. ChildrenMBR is shown by the outer bounding box while the reduced MBR is shown by dotted lines.

Improved algorithm: The main idea to further prune the search space is to reduce the number of points that can form an MBR and only consider cases where the constraint on the area is not violated. The algorithm first finds an initial solution by dropping extreme points in each direction until the area constraint is met. It can be noted that we only use child locations while finding an initial solution i.e. P_{center} is assumed as constant. This is based on the empirical results from our sample of 1000 places and their given center. The algorithm then tries to improve upon the initial solution while making sure the area constraint is satisfied. Let m be the number of points that are dropped to find the initial solution. This sets a limit on the number of points in each side that an MBR can pass through. There are $\mathcal{O}(m^4)$ such MBRs. The number of wrong entries in a gazetteer is expected to be a small fraction, hence m is expected to be much smaller than n. In our experiments, maximum value of m for GeoNames and OSMNames is found to be 62 and 78 while maximum value of n is 74765 and 708 respectively. One can also do a binary search when selecting the last side of the MBR, reducing the complexity of the naive algorithm to $\mathcal{O}(n^3 log(n))$ and that of the improved algorithm to $\mathcal{O}(m^3 log(m))$. The details of the improved algorithm are given in Algorithm 1.

A limitation of this probabilistic approach is that it generates an optimal MBR only for places whose area of childrenMBR is greater than the known area of the place (hence there is room to cut childrenMBR). In the next section, we present some heuristic solutions which can be applied to all places and may be easier to compute.

3.4 Heuristic Approach

Heuristics approaches may be used to construct the MBR of a place. Consider our hierarchical approach where the MBR of a place encloses the MBRs of all of its children. A problem with this approach is that a child that is placed under a wrong parent can dramatically change both the size and the shape of the MBR. Additionally, gazetteers are often formed by merging data from different sources, and some sources are not as reliable as the others. Our first heuristic is based on detecting outlier children.

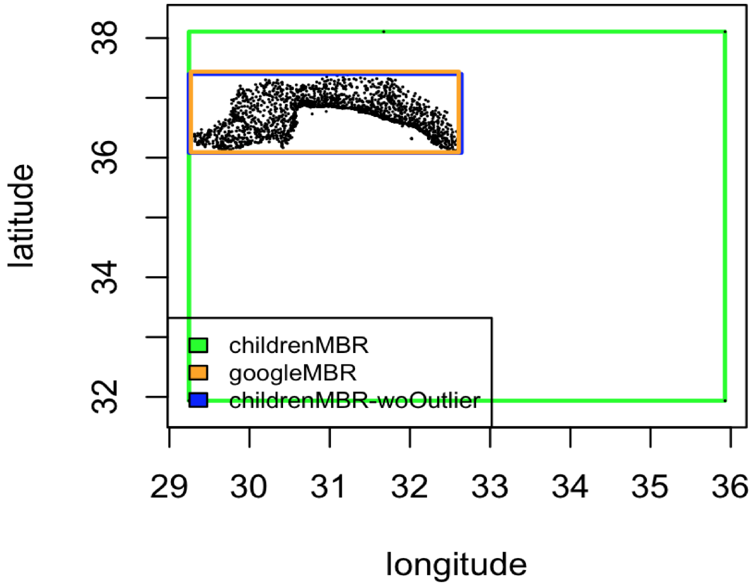


Figure 3.4: Outlier removal using Bagplot: Antalya, a province in Turkey.

3.4.1 Outliers Removal

Gazetteers sometimes have locations that are wrongly placed. For instance, Kemer is placed under the Antalya province of Turkey in GeoNames although its geocoordinates falls under the Konya province of Turkey[1]. Such wrong placements may show as an outlier especially if the wrong child is quite far from other children listed under the same parent. Hence, an outlier detection method may be used to detect such places before constructing an MBR. Outlier detection is extensively studied in the literature (e.g. [21, 38]). Two approaches that are used in geographical contexts are Boxplot [48] and Bagplot [39].

Boxplot is a univariate method that can be applied across both latitude and longitude dimensions. A point may be deemed an outlier if it is classified as an outlier in any one of the two dimensions; such outliers can be removed before constructing an MBR. Whereas, Bagplot is a bivariate extension of Boxplot, which generates a convex hull with 50% of the points (called a ‘bag’) and an outer loop (known as ‘fence’), which can vary in size depending on the number of points one wants to include. We expand the outer loop till the area of the MBR formed by enclosed points is closest or equal to the given area of the place. All the points which are outside the outer loop are considered as outliers and are excluded in the MBR construction. Figure 3.4 shows the application of Bagplot to obtain accurate bounding box of Antalya.

Similar to our approach, removing outliers before constructing an MBR is expected to generate an accurate MBR when the area of the children MBR is greater than the given area. A difference is that our probabilistic approach uses the center of the MBR, in addition to the set of child locations and the given area, to select an MBR whose center is closest to the given center. This may give a better MBR as compared to those generated from outlier removal methods especially when the majority of child locations are away from the given center.

We refer to MBR formed after removing outliers from childrenMBR as

[1] This example is from GeoNames and the geoname id for Kemer is 308213.

‘childrenMBR_woOutlier’ in later sections.

3.4.2 Hybrid MBR

An MBR of a place may be obtained using its center point and area; an MBR of a place may also be obtained based on the children listed. If we treat each MBR as a random variable which is 1 for points inside the MBR and 0 for points that fall outside, the region where the two MBRs overlap is where both random variables are taking the value of 1. The rectangle marked by the intersection of the two MBRs is expected to give a more reliable description of the boundary. However, the region of overlap can be much smaller than the actual MBR. We next discuss how this intersection region can be expanded such that its area matches the given area.

(1) Hybrid MBR with uniform enlargement (H-enlarge): Let l and w denote the length and the width of an MBR. To enlarge the MBR, we may enlarge both l and w by a constant s. Given an area a, we want $(l+s)(w+s)$ to be close to a. In other words, the value of s can be obtained by solving the following quadratic equation:

$$s^2 + (l+w)s + (lw - a) = 0 \quad \text{where } s > 0.$$

The coordinates of the expanded MBR are obtained by shifting the latitude and longitude of north-east and south-west region by degree equivalent to $s/2$ length in each direction. If C is the center of intersection region with latitude C_{lat} and longitude C_{long} respectively, the latitude of endpoints (NE_{lat}, SW_{lat}) are given as

$$\mathbf{NE_{lat}} = C_{lat} + F \quad \text{and} \quad \mathbf{SW_{lat}} = C_{lat} - F$$

where $F = s/(2 * L)$ and L is the distance between two consecutive latitude ($\approx$ 111 km). The longitude of endpoints (NE_{long}, SW_{long}) are calculated in similar way as in Section 3.2.1 with few changes. Modified equations are given below:

$$F_{ne} = s/(2 * D_{nelat})$$

$$\text{where} \quad D_{nelat} = L * cos(nelat_{radian}) \text{ (from [45]) and}$$

$$nelat_{radian} = (NE_{lat} * \pi)/180,$$

$$F_{sw} = s/(2 * D_{swlat})$$

$$\text{where} \quad D_{swlat} = L * cos(swlat_{radian}) \text{ (from [45]) and}$$

$$swlat_{radian} = (SW_{lat} * \pi)/180.$$

$$\therefore \quad \mathbf{NE_{long}} = C_{long} + F_{ne} \quad \text{and} \quad \mathbf{SW_{long}} = C_{long} - F_{sw}.$$

(2) Hybrid MBR with scaling (H-scale): Under this strategy, sides of the intersection region are scaled by a factor s such that the area of the expanded MBR becomes a. Hence, the value of s can be obtained as

$$s = \sqrt{}(a/lw).$$

The endpoints of the expanded MBR are obtained in the same way as in the previous case. In our experiments, we find the intersection of childrenMBR_woOutlier with centerMBR and apply the above conflation technique. In a case of no intersection region, we use childrenMBR as the default MBR.

Table 3.1 lists different approaches that are discussed in this chapter and the MBRs created using a given input.

Methods	Input	Output MBR
Hierarchical	child locations	childrenMBR
Geometric	center point and area of true MBR	centerMBR
Probabilistic	center point, area of true MBR, child locations	POM-based MBR
Heuristic	child locations and area of true MBR	childrenMBR_woOutlier
	center point, area of true MBR and child locations	hybridMBR

Table 3.1: List of different methods, input required and output MBRs.

Chapter 4

Applications

A bounding box of points in a spatial dataset represents a level of abstraction for grouping and to describe the common features of the points that fall inside the box. Including such an abstraction in a gazetteer has a number of advantages. In this chapter, we present a few application areas where the knowledge of the geographical boundaries of places can be useful; these applications can clearly benefit from maintaining such information in the spatial hierarchy of a gazetteer.

4.1 Gazetteer Refinement

Gazetteers provide a spatial hierarchy of places based on their containment or part-of relationships. However, such relationships are often described in a coarser granularity than desired. For example, at the time of writing this thesis, GeoNames places "University of Alberta" as a child under Alberta instead of Edmonton. This is partly because places such as cities and towns are represented by a point rather than with some geographical extent, which in turn prevents locations such as localities, parks and hospitals to be placed under it. This can adversely impact applications that query a gazetteer to determine the next administrative division in the spatial hierarchy that can contain a given place; an example query is "In which city a particular hospital or library is located".

One approach to refine the spatial relationships in a gazetteer is to use the bounding boxes of places for restructuring the spatial hierarchy. This

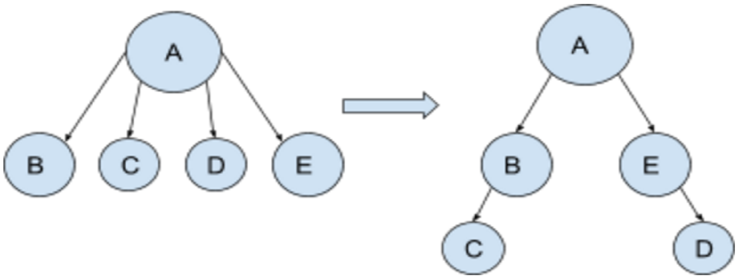


Figure 4.1: Gazetteer Refinement: Places C and D are pushed deeper in the hierarchy after a restructuring.

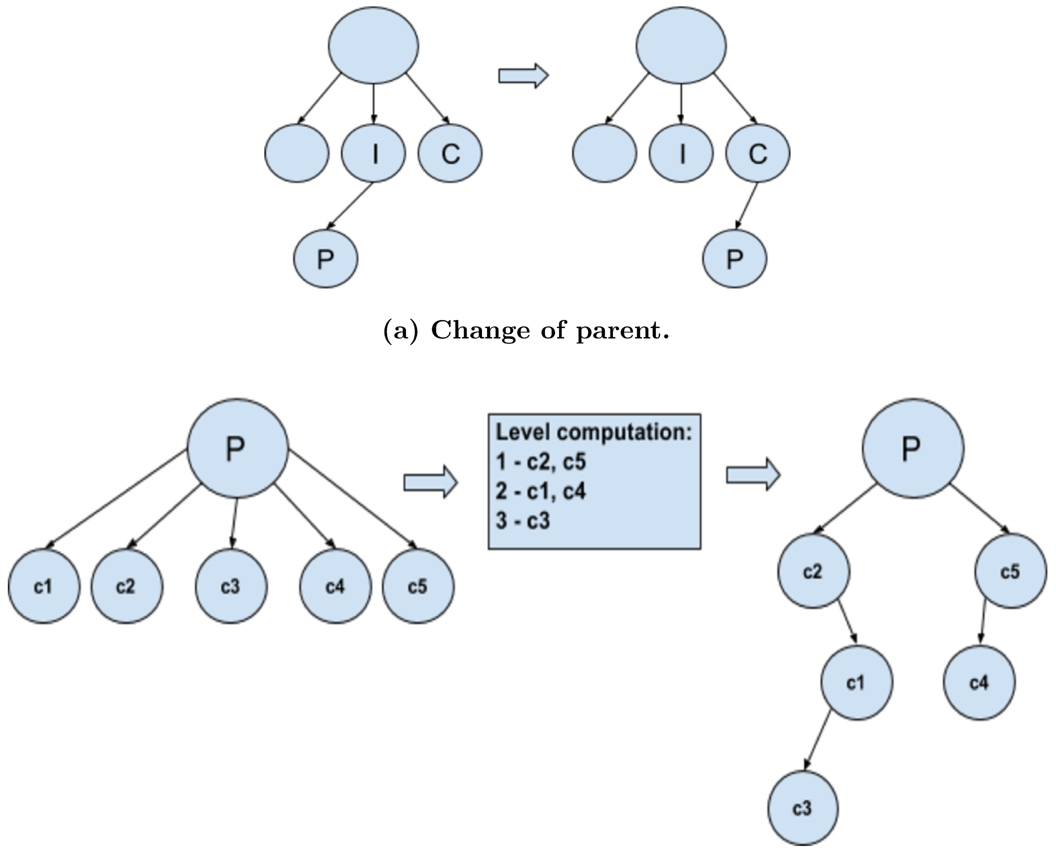
allows locations at lower level divisions to be moved deeper in the hierarchy for accurately looking up higher administrative division places in a gazetteer. Figure 4.1 shows an example of a hierarchy refinement.

Given a gazetteer with both parent-child relationships between places (as provided by the gazetteer hierarchy) and their spatial footprints (as obtained in Chapter 3), a gazetteer refinement may perform the following two operations (as shown in Figure 4.2).

(1) Change of Parent For each child node C_i placed directly under a parent node P in the spatial hierarchy of a gazetteer, we expect the MBR of node C_i to be fully contained in the MBR of P, where the MBR of places may be obtained using one of our approaches discussed earlier. A child location that does not satisfy this containment relationship is a candidate for parent change. There can be multiple locations that can contain the MBR of C_i and they may or may not be true parent places. One approach to reduce the likelihood of selecting a wrong parent location is through setting some constraints. One such constraint that can be employed when MBR of a node C_i can be contained in the MBR of node P' is to change the parent from P to P' if P' is the only place that can contain C_i.

(2) Restructuring Children Our aim is to correctly position each child node in the lowest level of the hierarchy, hence refining the hierarchy.

Our restructuring algorithm checks for each place c in level l if its MBR is fully contained in the MBR of another place p at the same level. If one such relationship holds and p is only node that has this relationship with c, then c



(a) Change of parent.

(b) Moving children to the lowest possible level.

Figure 4.2: Different operations for refining the spatial hierarchy.

becomes a child of p. This process is continued until no more move is possible.

4.2 Gazetteer Enrichment via Geotagging

Sometimes we want to enrich a gazetteer with data that may be available from other sources. In a geographical context, gazetteer enrichment can be viewed as including or integrating information about geographic locations. Such information may add more value to a gazetteer and may include data such as tags, tweets, news, documents, photos, videos etc.. To augment a gazetteer with such information, one needs to know the geographic extent of the locations in a gazetteer and the geocoordinate of the entities to be augmented. However, often the resources on the Web are not geotagged. A related question is if using the geographical extent of places can improve geotagging; we study this in the context of geotagging photos and videos from Flickr.

Geotagging digital resources in general is a useful operation in scientific projects [36] and commercial applications such as Instagram and Facebook that rely on location specific content. Textual annotations may be used in geotagging resources on the Web [24]. Such annotations can sometimes be machine generated or may be explicitly provided by users. Geotagging based on annotations is a challenging task since tags attached to an entity such as a photo or a video may not be location-specific; even location-specific tags can be vague and may refer to multiple locations. Also, some entities may have no tags or have only tags that have not been seen before (e.g. in the training phase).

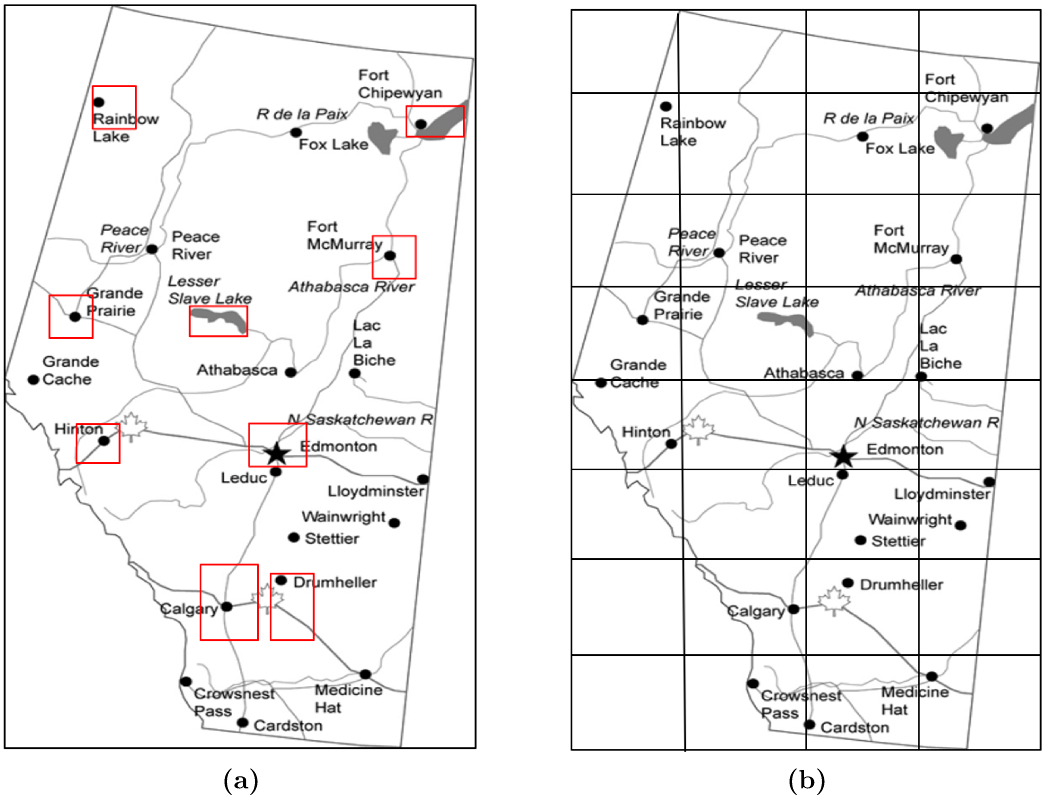


Figure 4.3: Map of Alberta with (a) the geometric boundary of places, and (b) a fixed size grid cell is shown.

Geotagging using MBRs: One approach for geotagging is to divide the earth surface into a grid of equal size cells [41, 24, 43] and predict the most probable cell for a given photo or video. A problem with cells is that it is possible that tags are spread over many neighbouring cells if the region it covers is a province or a country. In other words, spatial scope of a place can be distributed over different cells. On the other hand, MBRs of places are expected to maintain the locality relationships and place boundaries as shown in Fig. 4.3[1]. Thus, a natural question is if using MBRs instead of grid cells leads to a better geotagging.

Our approach to geotag photos and videos from Flickr using MBRs is divided into two steps (1) MBR prediction in which we predict an MBR for a photo or a video, and (2) coordinate estimation in which we determine the coordinate of the photo or the video within the predicted MBR. The details of the approach is discussed below:

(1) MBR prediction The relevance of a tag t to an MBR can be defined as the probability that a user inside the MBR m_j uses t to tag his/her photos. This probability can be estimated for a single tag t_i as

$$p(t_i|M_j) = \frac{\text{\# of users who use tag } t_i \text{ in MBR } m_j}{\text{\# of users in mbr } m_j}$$

where M_j is the model of MBR m_j. This formulation is based on the hypothesis that different users in an MBR may use similar tags to describe a place. The more a tag is used by users from the same location, the more location specific the tag is. To avoid zeroing the score in case a tag is not seen in a training phase, we apply the Jelinek-Mercer smoothing [49]. This changes the relevance of a tag t to an MBR as

$$p(t_i|m_j) = \alpha p(t_i|M_j) + (1 - \alpha)p(t_i|M_{mbrs})$$

where α is the smoothing factor with a value in the range (0, 1) and $p(t_i|M_{mbrs})$ is the model for all MBRs, defined as

$$p(t_i|M_{mbrs}) = \frac{\text{\# of users who use tag } t_i \text{ over all MBRs}}{\text{\# of users over all MBRs}}.$$

[1] The map of Alberta is obtained from www.canadah.com. Map and MBRs may not be to the scale and accurate.

In our experiment, we set the value of α at 0.8. Assuming independence between the tags inside an MBR, the relevance score of a test instance T with tags $t_1, \ldots, t_n$ is given as

$$p(T|m_j) = \prod_{i=1}^{n} p(t_i|m_j).$$

In our experiments, we use the log of $p(T|m_j)$ as our scoring function for numerical stability; this changes the product on the right side to a sum as given below

$$\log(p(T|m_j)) = \sum_{i=1}^{n} \log(p(t_i|m_j)).$$

One may note that the relevance score can be biased toward user-specific tags, which often do not carry any location information (e.g. person name). To avoid such ambiguity, we remove all tags which are used just by a single user. Further, to allow a locality of the tags, the same user in different cells or MBRs is considered as a new user. Finally, the MBR with the maximum score is considered as a best MBR ($m_{predicted}$) for a given test instance i.e.

$$m_{predicted} = \underset{m_j \in M}{\operatorname{argmax}}(\log\left(p(T|m_j)\right)).$$

One limitation of this model is that it is only applicable for test instances that contain at least one tag seen during the training phase. For test instances with all unseen tags, one can employ different heuristics to predict an MBR. One heuristic that we use is to predict the most popular MBR i.e. the MBR which has the maximum number of photos or videos assigned.

(2) Geocoordinates estimation The strategy to estimate the geocoordinates of a photo or a video is dependent on whether the MBR is predicted using the probabilistic language model discussed above or using the most popular MBR. In the former case, one approach is to find the Jaccard Similarity [23] between the tag set of the test instance and each of the training photo or video in the predicted MBR. The geocoordinates of a training photo or video, which gives the maximum value of the Jaccard Similarity, can be considered as the geocoordinates of the test instance. In the other case when an MBR is predicted based on the most popular MBR the geocoordinates of the training instance, which has the minimum Haversine distance [37] to other instances in the MBR, can be considered as the geocoordinates of the test instance.

Our experiments in Section 5.2 show how the idea of using an MBR instead of a grid cell can improve the performance of a geotagging application.

4.3 Topological Constraints

Topological constraints, when defined, between the objects in a spatial database can ensure to some degree the accuracy or the correctness of the database. Such constraints between places in a gazetteer may be defined using spatial structures such as convex hull or polygon. However, enforcing the constraints is computationally the simplest if the structure to represent a place is an MBR [9]. Relationships such as containment, disjointness and overlap between MBRs can be harnessed to define a set of topological constraints. Such constraints can further be classified into (1) soft constraints and (2) hard constraints. *Soft constraints* may be violated but the violations are expected to be rare and may lead to a warning. *Hard constraints* are those that cannot be violated; They are akin to referential integrity or foreign key constraint in relational database. The following topological constraints may be defined to maintain the data integrity of a gazetteer.

(1) ***Hierarchical***: The following constraint may enforce the hierarchical relationships between MBRs:

If a location A contains another location B, the MBR of location A must contain the MBR of location B.

One example of it is shown in Fig. 4.4[2]. This constraint is not true vice-versa; for example, the MBR of Saudi Arabia contains the MBR of Qatar even though both are different countries. This is of type hard constraint since a violation can cause incorrect modification of a parent or a child MBR.

(2) ***Disjoint***: The bounding boxes of places are expected to be disjoint

[2] The map of India is obtained from www.d-maps.com. Map and MBRs may not be to the scale and accurate.

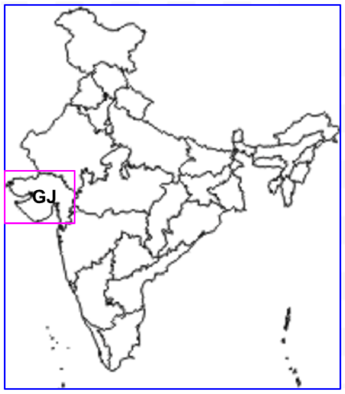


Figure 4.4: An example of a Hierarchical constraint where the state Gujarat (GJ) is part of India and the same relationship holds between their MBRs.

unless they are in some form of ancestor-descendent relationships. This may be expressed as:

If the MBR of place A does not contain the MBR of place B, then location A is not part-of or doesn't contain location B.

This is also a type of hard constraint and may allow pruning of the search space based on the containment relationship between MBRs of the places.

(3) ***Overlap***: The MBR of a place can partially overlap with that of its neighbouring locations (other than its parent and child locations) and the area of the overlap region may depend on shape, size and the position of the places in the spatial hierarchy. One may estimate the area of overlap for neighbouring locations at different administrative levels, and such estimates may be used to define a type of constraint that is expected but not guaranteed to hold. We classify this constraint as a soft constraint; a database administrator may employ one such constraint and raise exceptions in the form of warnings when an update to the gazetteer violates the constraint.

Hard constraints are parameter free and can be applied directly. However, soft constraints require the estimation of some parameters. The parameters of a soft constraint can be specific to the type of constraint to be created. We provide examples of a few soft constraints based on the area of overlap between places and the statistical properties of the overlap. Our aim is to estimate the overlap area based on places for which we think we have accurate MBRs and

Level	# of places	μ	σ	Threshold	
				$\alpha = 0.05$	$\alpha = 0.01$
ADM1	2000	0.2756	0.1451	0.514	0.613
ADM2	2000	0.1819	0.1655	0.454	0.566
ADM3	2000	0.1634	0.1462	0.403	0.503

Table 4.1: Expected area of overlap between MBRs at each level (ADM1 = province, ADM2 = district or large city, ADM3 = locality or small town) and the respective thresholds at each level of significance.

then, calculate parameters to define a constraint. However, often we do not have such information about places and their MBRs. To address this, (i) we only consider the overlap between places at the same administrative levels, and (ii) we ignore places when their MBR is fully contained in another MBR; such overlaps can be (a) due to a parent-child relationship that may or may not be correctly represented in GeoNames, or (b) because of an inaccuracy in data.

Example 1 In three different random sample of 2000 locations obtained from different levels (ADM1, ADM2 and ADM3) of the spatial hierarchy of GeoNames, we find the overlap area of each place P with other places (except those which can contain P or those which can be contained by P) at the same administrative level. The overlap area of a place P in a sample is further normalized with the area of P. We can calculate parameters such as mean and standard deviation of the normalized overlap area of these locations in each sample as shown in Table 4.1. Based on the distribution parameters, different basic constraints or rules can be defined. For example, if for any location at ADM2 level, its average normalized overlap area with all other places is much below or higher than the mean normalized overlap area (0.1819), then an assertion may be triggered. This example, however does not provide a threshold with which one can compare mean normalized overlap area of a place and can decide with some certainty that whether an update is actually unusual.

Example 2 Consider the setting of the first example, but suppose we want to calculate an upper bound on the mean normalized overlap area of a place based on the mean normalized overlap area of a sample. In the absence of information for complete population, we assume the given sample as our population and its mean and standard deviation as mean and standard deviation of the population. Given the mean μ and standard deviation σ of the population, our null hypothesis (H_0) may be defined as: the difference between the population mean and the mean normalized overlap area of a place is zero. The alternative hypothesis (H_1) can be that the mean normalized overlap area of a place is greater than the population mean. As our alternative hypothesis is one-tailed, we calculate a threshold or upper bound on mean normalized overlap area of a place (T) based on one-tailed critical values of z-score (Z = 1.645 and 2.326 corresponding to $\alpha = 0.05$ and 0.01 respectively) at different levels of significance and setting the sample size 'n' at one. Formally, it is given as

$$Z_\alpha = \frac{(T - \mu)}{\sigma/\sqrt{n}}.$$

In our case, $n = 1$ as we want to find a mean normalized overlap area for a place and therefore, T is given as

$$T = (Z_\alpha * \sigma) + \mu.$$

Table 4.1 presents for our sample an expected upper bound on the area of overlap at different administrative levels and different levels of significance. A soft constraint on this can be that if an update for a place P at ADM3 level makes its mean normalized overlap area greater than 0.403, then the null hypothesis can be rejected at $\alpha = 0.05$.

Chapter 5

Experimental Evaluation

In this chapter, we evaluate the proposed approaches in terms of both the accuracy and the effectiveness of the bounding boxes that are constructed.

5.1 Accuracy of an MBR

The accuracy of an MBR of a place may be measured against published data from authoritative sources such as government agencies and international organization bodies. We are not aware of any such source providing a comprehensive list of boundary regions, though there are sources that provide data on a best-effort basis. As one such source, we use Google Reverse Geocoding API[1] to fetch the true bounding boxes for a location. Google Maps has been used in similar context in the literature [19, 27].

5.1.1 Dataset and Evaluation Measures

Two gazetteers, namely GeoNames[2] and OSMNames[3], are used to evaluate the proposed methods.

GeoNames: GeoNames contains 9 feature classes, 667 categories or feature types and a total 11,031,666 geographical entities as of May, 2016. For each place, it provides name, alternative names, latitude and longitude, feature class, feature type, country code and administrative division code at different

[1] https://developers.google.com/maps/documentation/geocoding/intro

[2] http://www.geonames.org/

[3] http://osmnames.org/

levels. In this work, we use latitude and longitude, feature classes and feature types to construct our spatial hierarchy.

The testset for our experiments with GeoNames consists of three independent datasets: 50 USA states, named as 'Geo-50', 540 randomly selected places each with at least one child place, named as 'Geo-540' and another 140 randomly selected places, named as 'Geo-140', such that each place has at least one child and the area of the MBR of the children (childrenMBR) is larger than that of the true MBR. We treat US states separately to study the behaviour of the methods on places which are well-covered by GeoNames, even at lower levels. The ground truth for evaluating Geo-50 is obtained by querying Google Maps. Geo-140 is intended to test our probabilistic approach which is applied for instances whose area of childrenMBR is larger than expected. The testsets are created by randomly picking 2500 places in GeoNames with the constraint that each place must have at least one child and then retrieving the ground truth using Reverse Geo-coding API[4] of Google Maps; this process retrieves the bounding boxes for 680 places. Out of these 680 places, we obtain a set of places whose area of childrenMBR is larger than that of the true MBR which gives 140 places, with which we form Geo-140 testset, while the remaining places are used to form Geo-540.

OSMNames: OSMNames is another open access gazetteer constructed from the Open Street Map data, providing coordinates along with complete address details for 21,055,841 geographical entities as of March 2017. For our evaluation, we extracted two independent testsets, from OSMNames, consisting of 1500 places, named as 'OSM-1500', and another 160 places, named as 'OSM-160'. While OSM-1500 dataset consists of places having at least one child location, OSM-160 is constructed based on the same extraction criteria as for Geo-140. Both datasets are formed by randomly picking 2500 locations which satisfy the stated criteria and invoking the Reverse Geocoding API to find the

[4] The bounding box of a place is obtained by invoking Geocoding API with latitude and longitude of the place (using LatLong class of Geocoding API) as arguments. This returns a list of places with their MBRs in JSON format. The bounding box of a place in the list whose *long_name* attribute value contains the given place name is returned as the googleMBR of the place. To ensure the accuracy of the geocoding API, we further require the distance between a query place and its returned match is less than 30 km.

Gazetteer	Testset	Hierarchical	Probabilistic	Geometric	Heuristic
GeoNames	Geo-140	✓	✓	✓	✓
	Geo-50	✓		✓	✓
	Geo-540	✓		✓	✓
OSMNames	OSM-160	✓	✓	✓	✓
	OSM-1500	✓		✓	✓

Table 5.1: Different testsets and the methods evaluated on each testset.

(ground truth) bounding box for each place. OSMNames also provides the bounding boxes for places it contains. We extracted the bounding boxes for the places in our testsets (OSM-140 and OSM-1500) to evaluate their accuracy with respect to the bounding boxes from Google Maps and also to compare them with those generated by our methods. Table 5.1 lists all our testsets and the strategies evaluated on each testset.

In our experiments, we use the given latitude and longitude of a place as the center of the place. An MBR is represented as a curvilinear rectangle on the spherical surface of earth. For our evaluation, we calculate the approximate area of a curvilinear rectangle by finding the Haversine distance between the endpoints to derive its length and width and use it to determine the area of the minimum bounding rectangle. More precisely, we calculate the area of a curvilinear rectangle as the product of its curvilinear length and width. This approximation is justified here because (1) we use the same formulation to calculate the area of true bounding box using the endpoints obtained from Google Maps, and (2) the area of curvilinear rectangle and the area of an MBR for a place is not expected to differ much unless the spatial extent of the place is stretched along longitude axes or the place is located near the poles; the number of such places is small.

Furthermore, in our calculations of area and the coordinates of an MBR, we consider the radius of earth as 6371.0 km and the distance between consecutive longitudes at equator as 111.0 km. We use an existing implementation of

Bagplot, available as an R package called 'aplpack'[5] for generating an MBR without outliers.

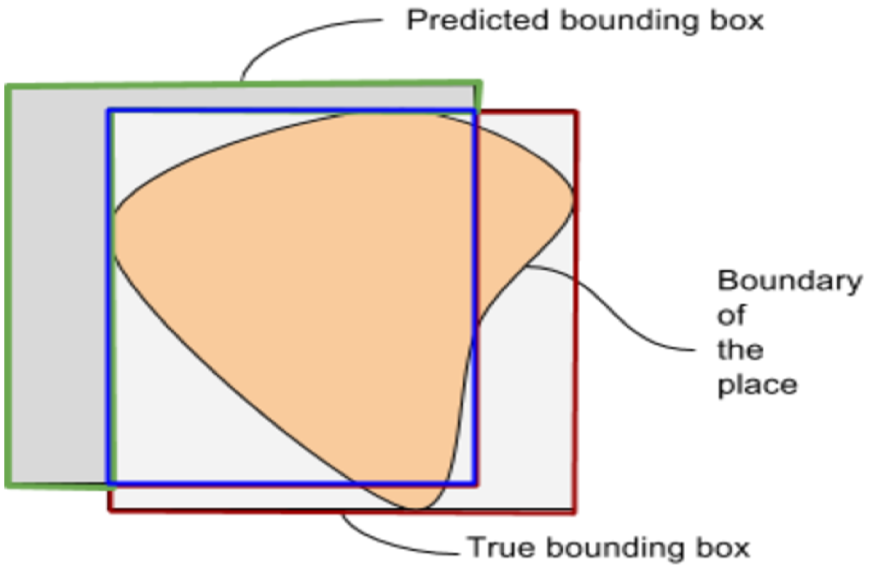


Figure 5.1: Evaluation measures: Region with boundary highlighted in (1) blue = accurate prediction (2) red = false negative (3) green = false positive.

Evaluation Metric: The accuracy of a bounding box of a place is measured in terms of its area and number of places it encloses. Figure 5.1 shows different regions based on which the following measures are developed:

Area Overlap Accuracy (AOA): It is the ratio of the area of intersection region, formed by the overlap of a predicted MBR with the true bounding box, over the area of the region formed by the union of the true and the predicted MBRs.

False Negative for area overlap (FN_{area}): It is the ratio of the area of the true bounding box not covered by a predicted MBR and the area of the true MBR.

False Positive for area overlap (FP_{area}): It is the ratio of the area of a predicted MBR which is not part of the true bounding box and the area of the predicted MBR.

Point Overlap Accuracy (POA): It is the ratio of the number of points

[5] https://CRAN.R-project.org/package=aplpack

in intersection region of the true bounding box and a predicted MBR over the number of points in the region formed by the union of the true and the predicted MBR.

False Negative for point overlap (*FN_{point}*): It is the ratio of the number of points in the true MBR which are not covered by a predicted MBR over the number of points covered by the true MBR.

False Positive for point overlap (*FP_{point}*): It is the ratio of the number of points in a predicted MBR which are not covered by the true MBR over the number of the points in the predicted MBR.

5.1.2 Data preprocessing

One redundancy observed in GeoNames is that it contains places with the same coordinates but different names. This adds overhead to the construction of the bounding box of a place. Therefore, we only use all unique child locations, based on their geo-coordinates, to prepare a baseline MBR (childrenMBR) for a place. Only those locations which follow a containment relationship with their parent locations are considered as children.

Our experiments use the spatial hierarchy expressed in a gazetteer. As an example from GeoNames, suppose we want to construct the relationships that cities Edmonton and Calgary are part of the province of Alberta and that is part of Canada. To build this hierarchy, we query GeoNames with name 'Canada', feature code 'PCLI' and feature class 'A' to obtain the attributes for Canada such as country code (CA). Provinces in Canada can be obtained by querying GeoNames using feature class 'A', feature code 'ADM1' and Country code 'CA'. The returned ADM1 code for the province of Alberta is '01'. Further, populated places such as Edmonton and Calgary under Alberta can be obtained by querying GeoNames using feature class 'P', feature code 'PPL', country code 'CA' and ADM1 code '01'.

In case of OSMNames, the spatial hierarchy of places can be built using 'display name' attribute in the given data file. For example, the display name for the Stanford University is "Stanford University, Santa Clara County, California, United States of America", and this ordering of the place names gives the parent-child relationships. One problem in dealing with place names is that two places at different locations can have the same name and sometimes the same OSM IDs (for example, OSM ID 4807857 is assigned to both Lindsay Place in Canada and Microrregio de Osasco in Brazil). Thus, we first assign a unique ID to places with a duplicate OSM ID, then replace place names in the display name field with their unique IDs and built the hierarchy thereafter.

5.1.3 Experimental Results and Discussions

In this section, we present our experiments on evaluating each of the proposed methods across different testsets. We also study the changes in accuracy as we vary some of the parameters. In our graphs and tables, we use abbreviations such as "Children", "Center", "Children_woOutlier", "POM", "H-enlarge" and "H-scale" to indicate ChildrenMBR, CenterMBR, ChildrenMBR_woOutlier and MBRs generated by probabilistic and heuristic approaches respectively. We also specify the outlier detection technique used by alias 'bag' (i.e. Bagplot) and 'box' (i.e. Boxplot) when we present results for ChildrenMBR_woOutlier in the tables. The experiments are categorized as follows:

(1) Overall accuracy on Geo-140 and OSM-160: The goal of this experiment is to determine the accuracy of each of the proposed methods on Geo-140 and OSM-160 testsets and to understand the factors or underlying data properties that affect the results. The results of the experiment are tabulated in Tables 5.2 and 5.3 respectively.

Strategies	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	44.43	53.66	**7.02**	91.32	8.67	**0.71**
Center	63.78	24.09	23.66	84.87	6.79	12.94
POM	**78.16**	9.42	17.29	**92.89**	6.79	2.12
Children_woOutlier (bag)	67.12	**9.30**	27.95	87.10	**6.44**	7.56
H-enlarge	72.09	17.98	17.87	89.37	6.82	7.89
H-scale	72.16	17.96	17.85	89.98	6.85	7.23

Table 5.2: Evaluation result for **Geo-140** dataset in (%)

Strategies	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	36.33	59.86	**18.85**	60.99	39.0	**3.75**
Center	53.39	**32.60**	32.34	76.47	**15.81**	14.99
POM	43.98	39.57	46.20	64.20	34.15	19.52
Children_woOutlier (box)	39.76	43.79	30.82	64.51	33.96	5.43
H-enlarge	**55.33**	31.03	30.98	**77.41**	17.18	12.18
H-scale	55.04	31.41	31.36	77.40	17.27	11.99

Table 5.3: Evaluation result for **OSM-160** dataset in (%)

Table 5.2 shows that childrenMBR for locations in Geo-140 are inflated due to wrong child locations placed under parent places. This can be seen from the high false positive (53.66%) for this approach. Our probabilistic approach performs the best as compared to other strategies, with around 33% improvement over our baseline. It is because with the probabilistic approach MBRs are shrunk by excluding outliers such that the area of each MBR that is formed is close to a given area. This reduces false positives and therefore improves the accuracy. CenterMBR performs better than childrenMBR as it constructs a square of a given area from the given center and since the area around the center is often included in the true MBR, it results in a better MBR than the childrenMBR. Furthermore, due to large number of points (on average $\approx$ 2789 children per place, see Table 5.4), childrenMBR_woOutlier outperforms the baseline as it removes the outliers.

From Table 5.3, it is clearly observed that hybridMBR outperforms other methods on the OSM-160 testset, with around 19% improvement over our baseline. The reason for this is that the majority of the locations in OSM-160 have children which are away from the given center. Additionally, most of these child locations are wrongly placed, which is evident from high value of FP_{area} and FP_{point} for childrenMBR. Thus, in such cases where child locations are less reliable and concentrated in a certain region, using conflation techniques such as hybridMBR performs better as compared to other methods.

In both the Tables 5.2 and 5.3, the point overlap accuracy is seen to be higher than the area overlap accuracy over all methods. It is because, a ma-

Parameters	Geo-140	Geo-540	OSM-160	OSM-1500	Geo-50
child count (mean)	2789	1015	111	71	40,210
area (mean)	78672.95	81026.98	125.11	367.05	489,324.96
places with area under 100 km^2	21	76	144	1275	0
child count (mean)	53	19	96	54	NA
area (mean)	43.54	46.90	17.39	14.04	NA
places with child count less than 25	11	379	35	889	0
child count (mean)	10	6	11	8	NA
area (mean)	683.35	7710.38	3.93	223.71	NA

Table 5.4: Statistics across different testsets.

jority of the points are part of a true MBR and this increases POA but the number of points are not enough or the points are not spatially extended in right directions to define the geographical boundary of the locations.

Since OSMNames provides bounding boxes of places, we calculate the area overlap accuracy for the places in OSM-160 using their bounding boxes from OSMNames. The area overlap accuracy of the MBRs obtained from OSMNames is 31.48% whereas our hybrid approach achieves 55.33%; this means our methods can construct better MBRs than those in OSMNames.

Strategies	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	68.03	29.06	**3.25**	97.95	2.04	**2.0**
Center	64.58	23.88	20.83	77.64	2.02	12.33
Children_woOutlier (bag)	**90.87**	**2.49**	7.05	**97.34**	**2.02**	2.63
H-enlarge	76.24	14.32	13.85	91.97	2.02	8.00
H-scale	76.00	15.17	14.67	92.65	2.02	7.32

Table 5.5: Evaluation result for **Geo-50** dataset in (%)

(2) Overall accuracy on Geo-50, Geo-540 and OSM-1500: In this experiment, we use Geo-50, Geo-540 and OSM-1500 testsets to evaluate each of the proposed methods except POM which is only applicable to 43 places in Geo-50 whose result is not mentioned in Table 5.5 but discussed when we present the results of Geo-50. The probabilistic approach cannot be applied to places in Geo-540, OSM-1500 and to few places in Geo-50 as the given information of places do not satisfy the stated criteria of POM.

Our result on the US states dataset (Geo-50) in Table 5.5 shows that childrenMBR_woOutlier outperforms all other strategies with AOA and POA respectively at 90.87% and 97.34%, producing the lowest number of false positives and negatives. This is due to a large number of child locations per place in Geo-50 (see Table 5.4), which also results in high accuracy of childrenMBR. On further applying bagplot on childrenMBR, outliers are removed which results in more accurate bounding box; hence a high accuracy for childrenMBR_woOutlier. There are also 43 places in Geo-50 on which POM is applied which gives an AOA and POA of 97.50% and 99.97% respectively. Such high accuracy shows that POM performs best when the number of children are high and can define the place boundary.

In the subsequent parts of this chapter, we will focus on discussing the results from more diverse testsets, namely Geo-540, OSM-1500, Geo-140 and OSM-160.

Strategies	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	34.20	6.27	61.68	94.49	5.50	**1.85**
Center	65.39	22.71	22.23	91.73	**4.29**	7.49
Children_woOutlier (box)	34.80	**5.44**	61.72	**94.52**	5.45	1.86
H-enlarge	**71.86**	17.95	**17.85**	94.28	4.75	4.28
H-scale	63.74	24.37	25.04	94.49	4.73	4.06

Table 5.6: Evaluation result for **Geo-540** dataset in (%)

Tables 5.6 and 5.7 show that the center and area information play a vital role in Geo-540 and OSM-1500. For both testsets, childrenMBR has less accuracy as most of the places do not contain enough spatial points to repre-

Strategies	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	26.27	10.85	66.04	90.20	9.79	**2.6**
Center	63.25	24.71	24.43	91.87	**4.56**	6.76
Children_woOutlier (box)	25.97	**10.83**	66.37	90.20	9.78	2.61
H-enlarge	**64.29**	23.79	**23.63**	94.15	4.90	4.08
H-scale	59.43	28.16	28.02	**94.32**	4.91	3.87

Table 5.7: Evaluation result for **OSM-1500** dataset in (%)

sent the place boundary (≈ 70% of places have less than 25 children per place in Geo-540 while it is ≈ 59% in OSM-1500, see Table 5.4). Due to the same reason childrenMBR_woOutlier performs worse than the baseline. On the contrary, centerMBR covers major parts of places using the area information and combining it with childrenMBR_woOutlier improves the area overlap accuracy remarkably to 71.86% on Geo-540 and 64.29% on OSM-1500.

We also calculated the area overlap accuracy for the places in OSM-1500 using their bounding boxes from OSMNames. The area overlap accuracy obtained is 66.53% which shows that MBRs obtained from our methods are pretty close to those in OSMNames. Note that our hybridMBR shows an overlap accuracy of 64.29% on the same dataset.

The study of overall accuracy across different testsets suggest that the factors which may affect the accuracy of an MBR of a place are number of child locations, their distribution and area of the place. In terms of applicability of our strategies, our heuristic approaches seems to perform better for places which have area and number of children less than the average. For places with comparatively larger area and number of children, our probabilistic approach yields better result.

(3) Performance of the expansion strategies: In this experiment, we study the performance of our expansion strategies. More specifically, we study the effect of applying uniform scaling and uniform enlargement operations on the accuracy of hybridMBR.

Figure 5.2 shows the number of instances in which a particular operation

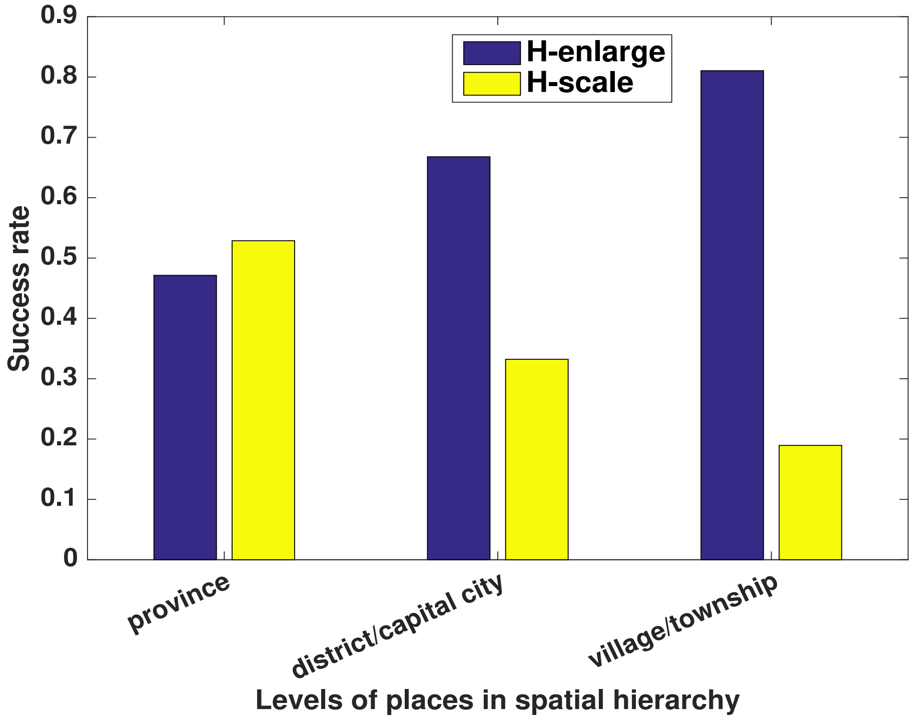
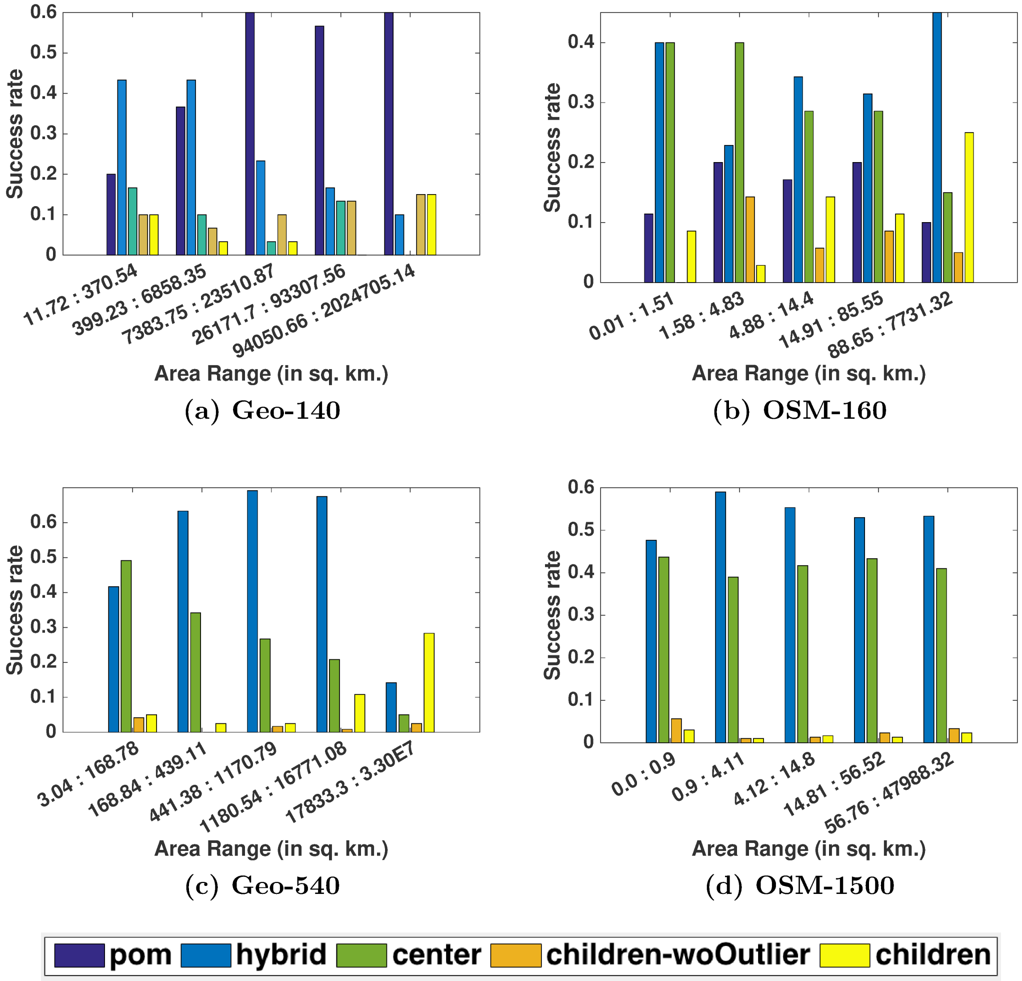


Figure 5.2: Success rate over Geo-540 varying the level of places in the spatial hierarchy.

outperforms the other operation at different administrative levels in Geo-540. One can observe that the performance of H-enlarge is comparable with H-scale for places at a higher level (e.g. province) while it performs better than the H-scale for lower level places (districts, cities and villages) in the spatial hierarchy. It is because the scaling operation extends the intersection region without changing the shape of the intersection region greatly. This seems to work better for places whose MBRs are geometrically close to rectangle. In case of uniform enlargement, the intersection region is equally incremented in both horizontal and vertical direction which result in MBR shape closer to a square. This seems to work better for places at coarser granularity since it covers majority of area around the intersection region.

(4) Varying the MBR area of places: In this experiment, we aim to derive a relationship (if any) between the MBR area of a place and the accuracy of our proposed methods. More precisely, we want to (1) study the performance of different methods across different testsets by varying MBR area, and (2) decide which method is best for a place with a given MBR area. To study this, we calculate the success rate of a method, defined as the ratio of the number of instances for which the method gives the best accuracy over total number of instances in a given area range. We then plot it against the corresponding MBR area range as shown in Fig. 5.3.



(a) Geo-140 **(b) OSM-160**

(c) Geo-540 **(d) OSM-1500**

(e) color legend for the graphs

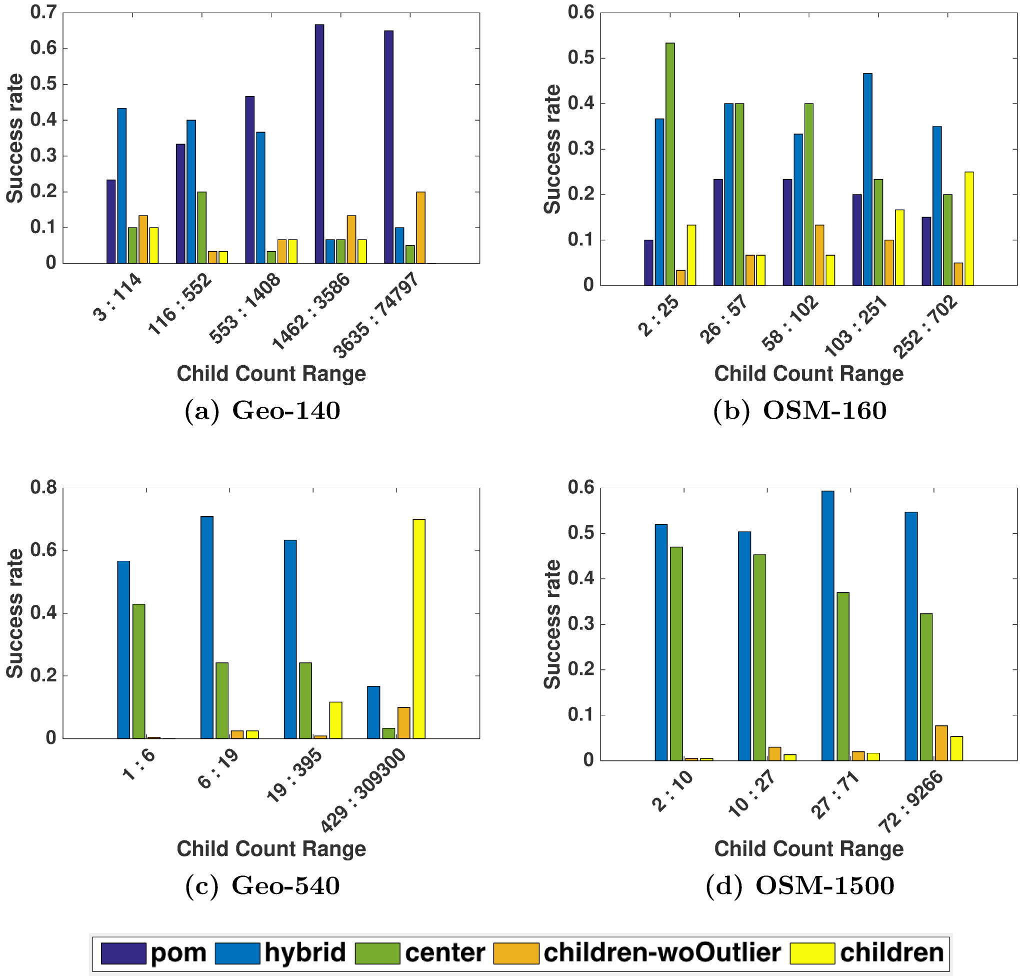
Figure 5.3: Success rate of the methods varying the area of MBR of the places.

From Figure 5.3 (a), it can be observed that the probabilistic approach performs best for places with area greater than 7383.75 km^2, which covers places that are mostly large cities, districts, provinces and small countries. For places with small area, conflation techniques such as hybridMBR seems to work well. Fig. 5.3 (b), (c) and (d) shows that the MBRs generated using our heuristic approaches, i.e. centerMBR and hybridMBR, outperform other strategies in almost all area ranges. This is because generally smaller regions are not covered well by gazetteers and the heuristic methods which uses the area and the center information seems to generate a better MBR. One may also notice that for spatially extended places such as territories or countries (see Fig. 5.3 (c)), childrenMBR outperforms all other methods. This is mainly due to number of child locations which is explained later.

Based on these empirical results, we can conclude that for smaller places such as localities and neighbourhoods, the heuristic approaches give a better result, and for places with large areas such as districts and states, the probabilistic approach seems to perform the best. Although in this experiment we observe the variation of area overlap accuracy with the area of MBR of the place, we find that it is the child locations which actually affects the accuracy of MBRs. Thus, we study the results by varying the number of child location in the next experiment.

(5) Varying the number of children for places: In this experiment, we perform a study similar to the previous one but instead of varying the area, we analyze changes in the area overlap accuracy of the strategies over different child count ranges.

From Fig. 5.4, it is seen that our probabilistic approach seems to perform better in cases when the child count is large while hybridMBR performs better when there are less number of children. Furthermore, for places with less number of children in OSM-160 (as shown in Fig. 5.4(b)), centerMBR performs better since it uses the area information to construct an MBR. However as the child count increases, hybridMBR is seen to be best for majority of locations as it is formed by using both child locations and area of MBR of the place. It can also be noted that in range [252, 702], childrenMBR works better than childrenMBR_woOutlier and probabilistic approach even though the child count is not large. This is because several places in child count range [252, 702], contains children which are not evenly distributed around the center. Thus, the child locations which are away from center and occur



(a) Geo-140 **(b) OSM-160**

(c) Geo-540 **(d) OSM-1500**

(e) color legend for the graphs

Figure 5.4: Success rate of the methods varying child count.

sparingly are considered as outliers and not included in POM-based MBR and childrenMBR_woOutlier; this leads to smaller and inaccurate MBR.

Fig 5.4 (c) and (d) show same characteristics as those seen in Fig. 5.3 (c) and (d), where our geometric and heuristic approaches performs best for majority of instances. One may also observe that childrenMBR is seen to be the best strategy when the child count is very large. This is because for a location with large number of children, its MBR based on children is more likely to be close approximation of true bounding box.

Our experiments in this section show that the number of children location is a major factor that affects the accuracy of different methods. The results can easily be understood by knowing the number of children and the way points are distributed. Based on the empirical results, it is seen that our probabilistic approach is the best choice for places which are moderately or well covered by a gazetteer. However, if the number of child locations are very high as in case of developed states or places at the country level, childrenMBR or childrenMBR_woOutliers may perform best. Lastly, in case of locations with less number of children (often seen for sparsely populated regions), having center and area information is useful and the heuristic methods can give a better abstraction of a place.

5.1.4 Summary of the Results

The conclusions drawn from the experiments conducted in this section can be summarized as follows:

- The overall accuracy of the proposed strategies exceeds that of a baseline across testsets. Although the testsets are diverse, as seen in Table 5.4, the empirical results show that the proposed methods are robust to changes in parameters such as the number of children and the area of an MBR of a place.

- Comparing the accuracy of the bounding boxes obtained using our methods with those in OSMNames shows that MBRs generated through our methods are comparable for places in the OSM-1500 testset while they are more accurate in case of the OSM-160 testset.

- A comparison of scaling and enlargement operations shows that enlargement works well for lower level places while scaling works better for places up in the hierarchy.

- The area of a place can indicate how a particular strategy performs at a given level in the hierarchy. But it does not help to understand why a particular method works better in a given area range.

- The accuracy of an MBR generated by the probabilistic and hierarchical approaches is directly dependent on the number of children. It also depends on how the points are distributed.

- For places with a large number of children (e.g. child count greater than 553 as seen in Geo-140) such as a province, a capital city, or a district, our probabilistic approach estimates a better MBR (more than 33% improvement as compared to the baseline). On the other hand, if a place contains a very large number of children (e.g. child count of order 10^5), the hierarchical approach generates an accurate MBR. However, such places are less frequent in a gazetteer as they usually represent a country or a territory. Finally, for a place with less number of children, an MBR constructed using heuristic methods represents a better abstraction of the geographical scope of a place.

5.2 Effectiveness of an MBR

Effectiveness of an MBR is measured in terms of its usability in some of the applications discussed in chapter 4. In this section, we present our evaluation setup and results on two applications: gazetteer refinement and enrichment.

5.2.1 Gazetteer Refinement

Our approach for refining a hierarchy requires as input (1) the spatial hierarchy expressed in a gazetteer, and (2) the MBR of the places. Each node in a hierarchy is checked for a valid parent-child relationship based on the containment relationship between their MBRs, and if a node is positioned incorrectly, then it is moved as per the two operations introduced in Section 4.1.

DataSet: The dataset used was GeoNames with its spatial hierarchy constructed as discussed in Section 5.1.2. We queried for places in GeoNames having at least one child location and retrieved a total of 16,120 places. Out of these 16,120 places, we filtered places which have no geographic extent due to duplicate coordinates and this reduced the number of places to 14,635.

Furthermore, we extracted a list of places along with their area from a public domain site (in our case Wikipedia Infobox) and matched these places with places in GeoNames based on both latitude/longitude and place name and this returned 78,639 places. The intersection of the two aforementioned sets gives 420 places that satisfy the applicable criteria of POM. Hence, we constructed their MBRs using our probabilistic approach. For the remaining places obtained earlier from GeoNames having at least one child location (14,215 places), their MBR was constructed using our hierarchical approach. For other set of places with no children but with area information i.e. 78,219 places, we obtained theirs MBRs using our geometric approach. In the end, we had MBRs for 93,274 locations, and those MBRs were used to restructure the hierarchy.

One point to observe here is that we used the area of each place, instead of the area of MBR of the place; this is a more realistic setting since the MBR area of a place is not available when the MBR is not known, whereas the area of a place can be found easily in public domains.

Evaluation Measure: To evaluate the refined hierarchy, we obtained a random sample of 100 places that were identified as inconsistent under Operation 1 and another 100 random sample of places that were moved deeper in the hierarchy under Operation 2 (see Section 4.1). We manually verified the detected inconsistencies and the moves using the information from Wikipedia and Google Maps. More precisely, for Operation 1 we verified whether a place which was identified as inconsistent was actually inconsistent and for Operation 2, we verified whether a location was part of another location. This was done manually by looking into Wikipedia text or Google Maps.

Results and Discussions: The total number of places identified as wrongly placed under Operation 1 was 67,820 while the total number of places moved deeper in the hierarchy under Operation 2 was 2,081,709. It can be seen that the number of locations to be moved deeper in the hierarchy is much higher than the number of places found inconsistent under Operation 1. In the absence of a geographic scope, there are many places which are kept directly below the root level. Furthermore, our evaluation result shows that 91% of places in our sample (91/100) are moved correctly down the hierarchy. This provides a strong evidence in support of an accurate restructuring. For Operation 1, the fraction of places which were actually inconsistent was 3/100; this empirically shows that the MBRs are robust enough to support the movement of places deeper in the hierarchy (vertical movements) but inconsistent for moving the nodes across the hierarchy (horizontal movements). The children places which were identified as wrongly placed are often streams, forests or places which lie near the geographical boundary of a parent. This is mainly due to vague geographic scope of natural landscapes and mis-alignments of a child MBR, which does not allow the MBR to fall completely under the parent node.

5.2.2 Gazetteer Enrichment via Geotagging

In this experiment, we want to augment GeoNames with photos/videos from Flickr by predicting the most likely MBR using the textual features such as user tags. Once an MBR is predicted for an instance, the geo-coordinates of the instance is obtained based on the geo-coordinates of the training instances mapped to the predicted MBR, as explained in [43].

DataSet: The dataset used for training and testing was extracted from the data provided in MediaEval Workshop 2016 for the 'Placing task', in which the organizers extracted the train and testsets from a huge corpus (YFCC100M) containing Flickr photos and videos. The extracted dataset consisted of 5,016,634 locations for training and 500,000 locations for testing. For each instance, title, description and user tags were preprocessed to remove special characters and stopwords. Furthermore, the data was stemmed using Snowball Stemmer[33]. The tag set for an instance consisted of the preprocessed tags, if any. Otherwise, associated description and title were used as tags. Instances which did not have any tags were removed. Thus, after preprocessing, total training data consisted of 4,631,717 photos/videos while the number of test instances remained the same.

To prepare the input for our experiment, we mapped each training instance to an MBR; there were 52,294 MBRs each with at least one photo or video. As a baseline for comparison, we also divided the surface of earth into cells of size 0.3 degree latitude and longitude and mapped the training data to the cells; there were 44,926 cells each with at least one training photo or video. The reason for generating cells of size 0.3 degree was to keep the number of cells comparable with the number of MBRs; otherwise an approach with large size cells or MBRs is expected to perform better but is computationally expensive. With this setup, we apply our approach, discussed in Section 4.2, to geotag photos and videos using both MBRs and grid cells separately and compare their error distance. The ground truth (i.e. geo-coordinates for test photos and videos) were provided in the dataset.

Evaluation Measure: A predicted geo-coordinate for a photo or a video is evaluated by calculating the Average Distance Error (ADE), defined as the Haversine distance between the predicted coordinate and the true coordinate of the photo or video. We also measured the prediction accuracy for MBRs (and similarly for cells) defined as the ratio of the number of instances for which a predicted MBR (cell) is the same as the true MBR (cell) to the total number of instances in the testset.

Results and Discussions: The aim of our experiment is (1) to determine how accurately a multimedia object is mapped to a gazetteer location using MBRs, and (2) to study the impact of using MBRs instead of cells.

Our experiment gives an ADE of 2561.114 km for MBRs compared to 3039.674 km for cells, with a prediction accuracy of 41.23% for MBRs and 32.37% for cells. This clearly shows that geotagging using MBRs is more accurate than the grid-based approach. The analysis of the results show two major reasons for wrong MBR or cell prediction which leads to large distance error; (1) There are several instances of photos or videos in the testset which only contain tags which are general terms (e.g. ‘affect’, ‘ipad’) and they do not carry any location-specific information. These tags can occur anywhere on world map and therefore difficult to predict. (2) There are cells and MBRs which are sparsely populated, i.e. they have very few users assigned (1 or 2). Thus, even though there are several dense MBRs or cells containing multiple tags of a test instance, a sparsely populated cell or MBR is predicted as best cell or MBR for the test instance. This is because such cells and MBRs have a very few users assigned as compared to populated cells or MBRs which have a number of users in the range of 10^5. Also, such cases are seen more often with cells since cells are created randomly without any knowledge of geographical scope of locations.

5.2.3 Summary of the Results

The conclusions drawn from the experiment in this section can be summarized as follows:

- Publicly managed gazetteers have places usually positioned directly under a country or a province. In our experiment on hierarchy refinement, we move more than 2M places down the hierarchy i.e. under a province, district or a city, with an accuracy of 91%. This determines the effectiveness of using MBRs in refining and maintaining a gazetteer.

- While our experiment shows promising results for vertical movement of places, it does not work well in identifying inconsistent child locations. Based on the empirical results, we find that the number of places to be moved deeper in the hierarchy are much higher than the number of inconsistent places in a gazetteer and our approach works far better in the former case.

- MBRs also show a great potential in joining external data with a gazetteer. This is shown by our experiment under gazetteer enrichment where GeoNames is augmented with photos and videos from Flickr using MBRs.

- Our methodology to geotag a photo or a video using MBRs yields less distance error as compared to a grid-based approach.

Chapter 6

Conclusions and Future Work

Gazetteers play a major role in geographic information retrieval applications. Hence, it is important to build gazetteers with high data quality standards and develop strategies to identify and possibly fix the inconsistencies. Integrating data from various sources often introduce some degree of anomalies and this seems to hold true in case of gazetteers.

In this thesis, we present different strategies to create the bounding boxes of places using the spatial hierarchy of a gazetteer and information such as the area of places, which are available in public domains. Our contributions include (1) strategies to construct an MBR of a place including a probabilistic optimization model and a few heuristic methods, and (2) an extensive evaluation of our strategies.

Our experimental evaluation on two different gazetteers and on multiple different testsets reveal that our probabilistic model captures the inconsistencies accurately, which results in a significant improvement (more than 33%) over the baseline. Furthermore our evaluation shows that our POM-based approach works best for places at district, provinces or higher level whereas our geometric and heuristic approaches can be employed for places without enough coverage in a gazetteer. Usually these places are sparsely populated regions such as localities, villages and points of interests positioned at the lower levels of the spatial hierarchy. Our experimental analysis shows that the number of child locations and its distribution are the most important factor in estimating an accurate MBR.

We also demonstrate the effectiveness of MBRs in both gazetteer refinement and gazetteer enrichment. A gazetteer refinement is done by restructuring the spatial hierarchy based on the containment relationship between MBRs of the places. Gazetteer enrichment is achieved through geotagging, where we build a probabilistic language model to geotag images and videos from Flickr using MBRs. Experimental analysis in this thesis shows that our MBRs are accurate for moving places deeper in the hierarchy but not accurate in moving places across siblings in the spatial hierarchy. The results of geotagging shows that our methods outperform grid based approach in terms of average distance error. We also provide a set of topological constraints based on MBRs that can be used to prevent dirty updates to gazetteers.

As possible directions for future research, one can improve on the accuracy of bounding boxes using local features such as landscape of places in close proximity, population density, etc. A limitation of our probabilistic model is that it is not applicable when the area of a childrenMBR is smaller than a given area. Also, as seen in our experimental results, our heuristic approach which remove outliers, does not work well when the center of a place is away from its child locations. These are some areas to improve the models or to build better models. Additionally, we believe that using other information such as location specific tags, geotagged images from location aware websites can further help in constructing robust bounding boxes and a better spatial hierarchy.

Bibliography

[1] Dirk Ahlers. Assessment of the accuracy of geonames gazetteer data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pages 74–81. ACM, 2013.

[2] Dirk Ahlers and Susanne Boll. On the accuracy of online geocoders. *Geoinformatik 2009*, 2009.

[3] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. The r*-tree: an efficient and robust access method for points and rectangles. In *ACM Sigmod Record*, volume 19, pages 322–331. Acm, 1990.

[4] Paul V Bolstad and James L Smith. Errors in gis. *Journal of Forestry*, 90(11):21–29, 1992.

[5] Thomas Brinkhoff and Hans-Peter Kriegel. Approximations for a multistep processing of spatial joins. *IGIS'94: Geographic Information Systems*, pages 25–34, 1994.

[6] Nieves R Brisaboa, Miguel R Luaces, Gonzalo Navarro, and Diego Seco. Range queries over a compact representation of minimum bounding rectangles. In *International Conference on Conceptual Modeling*, pages 33–42. Springer, 2010.

[7] Volha Bryl and Christian Bizer. Learning conflict resolution strategies for cross-language wikipedia data fusion. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1129–1134. ACM, 2014.

[8] Cláudio Elízio Calazans Campelo and Cláudio de Souza Baptista. Geographic scope modeling for web documents. In *Proceedings of the 2nd international workshop on Geographic information retrieval*, pages 11–18. ACM, 2008.

[9] Douglas R. Caldwell. Unlocking the mysteries of the bounding box. *URL http://purl.oclc.org/coordinates/a2.htm*, 2005.

[10] Cláudio Campelo and Cláudio de Souza Baptista. A model for geographic knowledge extraction on web documents. *Advances in Conceptual Modeling-Challenging Perspectives*, pages 317–326, 2009.

[11] Jiaoli Chen and Shih-Lung Shaw. Representing the spatial extent of places based on flickr photos with a representativeness-weighted kernel density estimation. In *International Conference on Geographic Information Science*, pages 130–144. Springer, 2016.

[12] Juan Chen, Anthony G Cohn, Dayou Liu, Shengsheng Wang, Jihong Ouyang, and Qiangyuan Yu. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(01):106–136, 2015.

[13] Maria A Cobb, Frederick E Petry, and Kevin B Shaw. Fuzzy spatial relationship refinements based on minimum bounding rectangle variations. *Fuzzy sets and systems*, 113(1):111–120, 2000.

[14] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. 2007.

[15] Xiangyuan Dai, Man Lung Yiu, Nikos Mamoulis, Yufei Tao, and Michail Vaitis. Probabilistic spatial queries on existentially uncertain data. In *International Symposium on Spatial and Temporal Databases*, pages 400–417. Springer, 2005.

[16] Patricia Frontiera, Ray Larson, and John Radke. A comparison of geometric approaches to assessing spatial similarity for gir. *International Journal of Geographical Information Science*, 22(3):337–360, 2008.

[17] Dhomas Hatta Fudholi, Wenny Rahayu, and Eric Pardede. Ontology-based information extraction for knowledge enrichment and validation. In *Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on*, pages 1116–1123. IEEE, 2016.

[18] Judith Gelernter, Gautam Ganesh, Hamsini Krishnakumar, and Wei Zhang. Automatic gazetteer enrichment with user-geocoded data. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 87–94. ACM, 2013.

[19] Maurizio Gibin, Alex Singleton, Richard Milton, Pablo Mateos, and Paul Longley. An exploratory cartographic visualisation of london through the google maps api. *Applied Spatial Analysis and Policy*, 1(2):85–97, 2008.

[20] Antonin Guttman. *R-trees: a dynamic index structure for spatial searching*, volume 14. ACM, 1984.

[21] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[22] Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 2010(1):21–48, 2012.

[23] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50, 1912.

[24] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Geotagging social media content with a refined language modelling approach. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 21–40. Springer, 2015.

[25] Ray R Larson and Patricia Frontiera. Ranking and representation for geographic information retrieval. In *Extended abstract in SIGIR 2004 Workshop on Geographic Information Retrieval*, 2004.

[26] Robert Laurini and Okba Kazar. Geographic ontologies: Survey and challenges. *meta-carto-semiotics*, 9(1):1–13, 2017.

[27] Weimo Liu, Md Farhadur Rahman, Saravanan Thirumuruganathan, Nan Zhang, and Gautam Das. Aggregate estimations over location based services. *Proceedings of the VLDB Endowment*, 8(12):1334–1345, 2015.

[28] Jodi R Norris, Stephen T Jackson, and Julio L Betancourt. Classification tree and minimum-volume ellipsoid analyses of the distribution of ponderosa pine in the western usa. *Journal of Biogeography*, 33(2):342–360, 2006.

[29] Maxwell Guimarães de Oliveira, Cláudio EC Campelo, Cláudio de Souza Baptista, and Michela Bertolotto. Gazetteer enrichment for addressing urban areas: a case study. *Journal of Location Based Services*, 10(2):142–159, 2016.

[30] Dimitris Papadias and Yannis Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographical Information Science*, 11(2):111–138, 1997.

[31] Jian Pei, Bin Jiang, Xuemin Lin, and Yidong Yuan. Probabilistic skylines on uncertain data. In *Proceedings of the 33rd international conference on Very large data bases*, pages 15–26. VLDB Endowment, 2007.

[32] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93. ACM, 2008.

[33] Martin F Porter. Snowball: A language for stemming algorithms, 2001.

[34] Rosanne Price, Nectaria Tryfona, and Christian S Jensen. Modeling topological constraints in spatial part-whole relationships. In *International Conference on Conceptual Modeling*, pages 27–40. Springer, 2001.

[35] Nataliya Prokoshyna, Jaroslaw Szlichta, Fei Chiang, Renée J Miller, and Divesh Srivastava. Combining quantitative and logical data cleaning. *Proceedings of the VLDB Endowment*, 9(4):300–311, 2015.

[36] Jiangwei Yu Rafiei and Davood Rafiei. Geotagging named entities in news and online documents. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1321–1330. ACM, 2016.

[37] C Carl Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[38] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

[39] Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

[40] Timos Sellis, Nick Roussopoulos, and Christos Faloutsos. The r+-tree: A dynamic index for multi-dimensional objects. Technical report, 1987.

[41] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.

[42] Ryan Shaw. Gazetteers enriched: A conceptual basis for linking gazetteers with other kinds of information.

[43] Sanket Kumar Singh and Davood Rafiei. Geotagging flickr photos and videos using language models. In *MediaEval*, 2016.

[44] María J Somodevilla and Fred E Petry. Fuzzy minimum bounding rectangles. In *Spatio-Temporal Databases*, pages 237–263. Springer, 2004.

[45] Kurt Stüwe. *Geodynamics of the lithosphere: An introduction.* Springer Science & Business Media, 2007.

[46] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[47] Mark Wick and Bernard Vatant. The geonames geographical database. *Available from World Wide Web: http://geonames. org*, 2012.

[48] David F Williamson, Robert A Parker, and Juliette S Kendrick. The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921, 1989.

[49] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[50] Jingxiong Zhang and Michael F Goodchild. *Uncertainty in geographical information.* CRC press, 2002.

[51] Wei Zhang and Judith Gelernter. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70, 2014.