SPATIAL APPROACHES TO REDUCING ERROR IN GEOCODED DATA

by

Daniel Wright Goldberg

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

May 2010

DEDICATION

This dissertation is dedicated to the life and memory of Michael Owen Wright-Goldberg – beloved son, husband, and brother. My overwhelming hope is that the research presented herein contributes in some small part to the ongoing efforts of more effectively diagnosing, treating, and ultimately curing cancer and other such horrific diseases.

ACKNOWLEDGMENTS

I would first and foremost thank my advisor and chair of my dissertation committee, Dr. John P. Wilson, for his continued guidance, encouragement, and sense of humor through the many years I have been involved in my academic pursuits. Without his constant support and timely advice this dissertation would not have been possible. Being able to observe a world renowned scholar from such a short distance has provided me with a rich set of experiences which I can draw from as I move forward in my career in academia. Many students in my position do not have a chance to stumble upon the professional and personal experiences my relationship with John has afforded me, and I consider myself truly fortunate to be among his students. I look forward to even just a small fraction of his success in my own career.

Another committee member, Dr. Myles G. Cockburn, has really gone above and beyond in helping me throughout my graduate studies. Without meeting and working with Myles, I do not think that my research or dissertation would be nearly as complete and focused a picture as I hope it is. I would also like to thank my other qualifying examination and dissertation defense committee members Drs. Craig A. Knoblock, Ulrich Neumann, and Cyrus Shahabi, each of whom has helped shape the focus of my research through their courses and guidance and has provided valuable insight, encouragement, and different points of view when they were needed the most.

It goes without saying that I owe the entirety of my life and all of my successes to my parents. I consider myself to be extremely lucky to have grown up in a wonderful household with two wonderful parents who wanted nothing in the world more than for me to succeed at every part of my life, from tying my shoes to completing a dissertation. Without their unwavering love and continuous support I would not be where I am today. My brother, Jon, and my sister Liz each played a role in helping me succeed in my graduate studies, whether they know it or not. It is with great sadness that we recently lost our older brother Mike, and without their support I don't think that I would have made it through another day let alone a dissertation. I could not ask for better siblings, nieces, nephews, and sisters-in-law and I look forward to spending more time together in the most beautiful place on earth, Long Beach Island.

I could not have survived graduate school without the support of the friends I have made here at USC: Shawn Allison, Tom Mernar, and Martin and Sarah Michalowski. I am grateful for the time we shared together before you all beat me to graduation and moved away to bigger and better things.

Each of my lab-mates here at the USC GIS Research has played an important role in my success as a graduate student and I look forward to watching each of you succeed in your studies and careers moving forward.

I would like to thank the North American Association of Central Cancer Registries, Charlie Blackburn, the member of the GIS committee, and particularly Frank Boscoe, Kevin Henry, and David Stinchcomb for providing me the opportunity to get my geocoding thoughts together in order in the form of a book, out of which sprung much of the inspiration for this dissertation.

Last but in no way least, I would like to thank Mona Seymour. You have been the best part of graduate school and were always there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The process of geocoding, converting geospatial textual information into one or more representative geographic locations or areas, is a fundamental geospatial operation essential to many diverse scientific fields such as environmental epidemiology, homeland security, sociology, political science, and transportation logistics. These geocoded data typically form the underlying data from which geographic mapping and visualization can occur, spatially-based research questions can be posed and investigated, and network routing and planning can be conducted. Although varied and diverse in terms of their applications and usages, this wide set of geocode users all require spatially accurate geocoded results as well as metrics capable of describing the accuracy.

The current state-of-the-art in geocoding technology is often capable of producing industry-accepted spatially accurate geocoded results under the ideal situations when input data are of high quality and expensive reference data layers are available. For most consumers of geocoded data and geocoding tools, difficulties in producing high quality data still remain. Even more so, the existing methods used to describe the quality of these data are severely deficient for use in scientific studies.

In this dissertation, I describe a geocoding system that addresses many of the fundamental underlying problems that cause inaccurate spatial results using uncertainty minimizing techniques. Specifically, we present a set of novel advances to geocoding algorithms which both increase the spatial accuracy of the output data and reduce the spatial uncertainty inherent in this information. To accomplish these tasks, we first outline a strategy for picking nearby candidate geocodes when a specific known reference feature is not available for a particular input address. This approach uses a spatially-varying block distance metric to define a local region of interest within which candidates are scored based on their spatial distance and attribute similarity. I next develop the concept of a spatial uncertainty-driven approach to candidate feature selection, integration, and interpolation that uses the characteristics of all available candidate features, their individual uncertainties, and their topological relationships to deduce the most likely candidate outcome. We finally turn our attention to the case of ambiguous results and develop both rule- and spatial neighborhood-based approaches for choosing the appropriate candidate feature based on the relationships between ambiguous candidates and the characteristics of the local regions around them.

Together, these three branches of the research presented serve to increase match rates (the number of successfully geocoded results), reduce spatial error (the distance from the computed output location to the ground truth position), and reduced spatial uncertainty (the number/scale of equi-probable locations to which a geocode could belong) in geocoded information. These advances increase the quality of geocoded data used in scientific studies and will play a key role in developing the next generation of spatial analysis approaches that utilize spatial uncertainty-based approaches to understanding geospatial phenomenon across scientific disciplines.

CHAPTER 1: INTRODUCTION

1.1 The Role of Geocoding in Research and Practice

Geocoding is the process of converting locationally descriptive textual information into one or more representative geographic locations or areas, typically applied to convert postal addresses into geographic coordinates (latitude and longitude) (Boscoe 2008; Rushton et al. 2006). Most often, this process is performed by converting postal address data into geographic coordinates (latitude and longitude) for use in locating people, places, objects, and events at a geographic position at a particular point in time.

This process has been used as the foundation for spatially-based research and practice throughout many diverse fields in academia and industry for decades, and continues to be the primary source of spatial data for countless research projects and applications (Bell et al. 2006; Chainey et al. 2005; Costello et al. 2009; Krieger 2003; Ratcliffe 2004; Ritz et al. 2009; Suarez et al. 2007; Vine et al. 1998). The spatial locations produced through the process of geocoding are used by researchers and practitioners to perform tasks and develop products and services that require an underlying spatial foundation. This high-level description is depicted graphically in Figure 1.1.

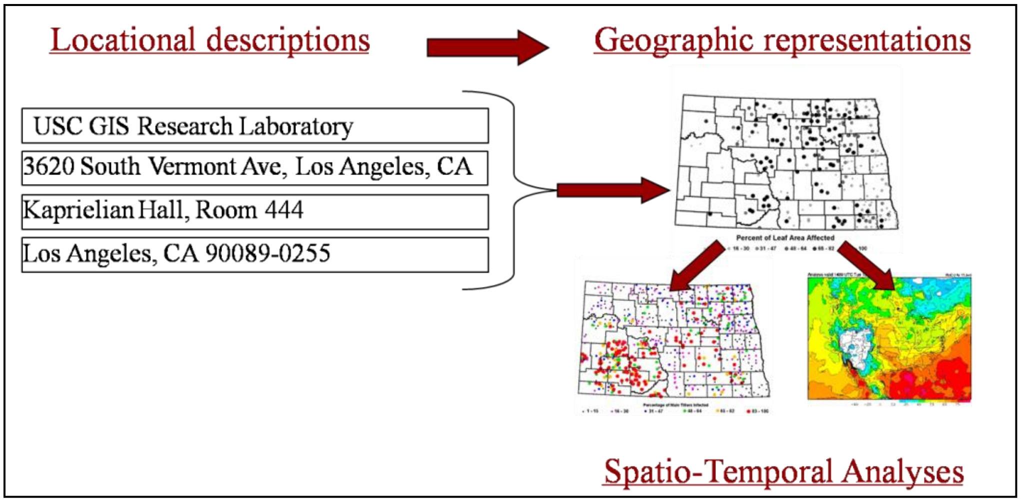

Locational descriptions
Geographic representations
USC GIS Research Laboratory
3620 South Vermont Ave, Los Angeles, CA
Kaprielian Hall, Room 444
Los Angeles, CA 90089-0255
Percent of Leaf Area Affected
Spatio-Temporal Analyses

Figure 1.1: High-level depiction of the role of geocoding

For example, suppose an epidemiologist has an intuition that living close to a freeway may increase the risk of developing mesothelioma, a type of cancer associated with exposure to asbestos. This researcher thinks these two might be related because automobile clutch linings and brake pads are often made from asbestos, and those living close to freeways are continuously exposed to air pollution from the freeway in the form of particulate matter that may contain trace amounts of this material.

To investigate the causal relationship between residential proximity to freeways and likelihood of developing mesothelioma, the first step this researcher will take will be to determine the binary exposed/non-exposed and incidence/non-incidence relationships for a set of cases (a sample of the population with mesothelioma) and controls (a random sample of the population). If these relationships indicate a preponderance of those living near freeways developing mesothelioma, the researcher may then take the further step of seeking to quantify the potential amount of particulate asbestos matter that an individual may have been exposed to in order to derive quantitative estimates of how much exposure is dangerous.

In both of these steps, the spatial locations of the individuals in both the case and control group form the foundation of the epidemiologist's raw data. The spatial locations of the cases are, in the majority of situations, derived by geocoding the address of the individual at the time they were first diagnosed as having the disease. In the first binary classification case (exposed/unexposed) epidemiologists typically perform a spatial intersection of the spatial locations of the individuals and a spatial surface describing the known, estimated, or probable spatial distribution of the phenomenon in question – in this case the spatial distribution of the existence/extent of the particulate matter. The locations of individuals that intersect with the distribution of particulate matter are classified as exposed while those not intersecting are classified as unexposed. An example of exposure misclassification is shown in Figure 1.2. The red point represents a point source emission with the red area being its probable distribution. The blue points represent the true geocodes that should be associated with the individuals based on address-range geocoding. The yellow points are incorrect U.S. Postal Service (USPS) ZIP code centroid based geocodes. This example shows that an individual in the green area would be misclassified as exposed if the USPS ZIP code centroid geocode was used, while an individual in the orange area would have likewise been misclassified as unexposed.

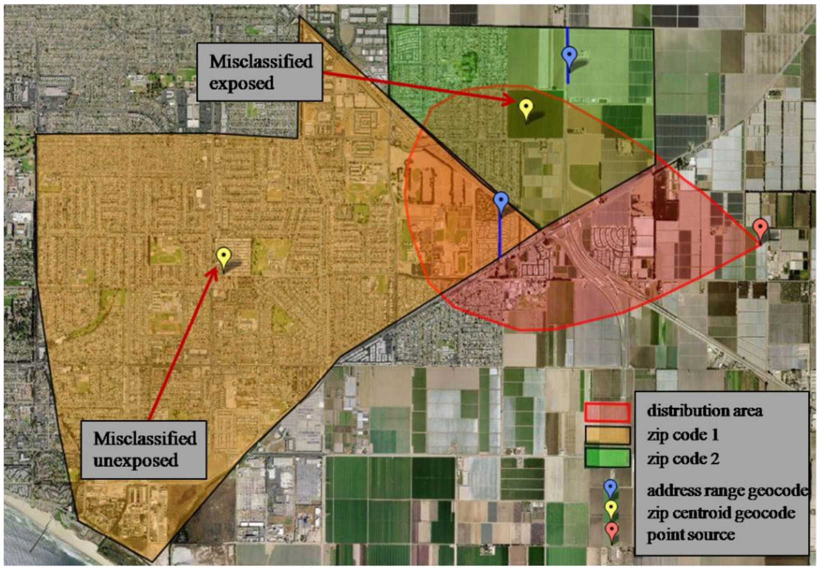


Figure 1.2: Hypothetical binary exposure misclassification example resulting from USPS ZIP code-level geocodes

The second case is very similar except spatial buffers are placed around the locations of the individuals to represent individual catchment areas, and the distribution of the particulate matter is represented by a surface that indicates quantities of particulate matter at a location instead of just existence/non-existence. A spatial intersection is performed and the intersecting values between catchment areas and particulate matter are aggregated to determine specific amounts of material that the individual may have been exposed to. A typical example of a pollutant distribution surface is shown in Figure 1.3. Here, the red and blue rings indicate 500 and 1,000 m catchment areas, respectively, while the points labeled a, b, and c represent the geocodes resulting from USPS ZIP code centroid, city centroid, and address range geocoding for the same individual. Again, this figure shows that miscalculation of exposure estimates that can occur when using geocodes of different qualities.

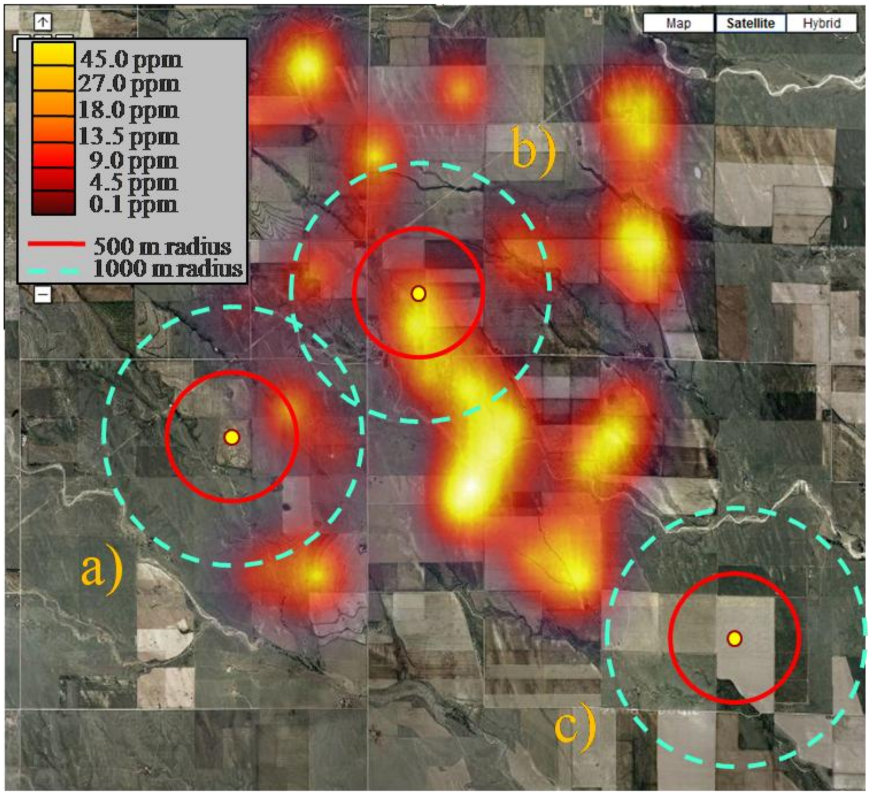


Figure 1.3: Hypothetical quantitative carbon monoxide exposure estimates varying by geocode type

A generalized version of the use of geocoded data in the scientific process is shown again in Figure 1.4. From this, we can see that errors or inaccuracies introduced during the conversion of address data to geocoded locations can draw any conclusions derived from a study into doubt. Further, we see that because these geocoded data underlie all subsequent steps in the research workflow, the relative magnitude of error that they introduce if incorrect or inaccurate is quite large proportional to the other steps of the process leading to the eventual research outcome. Therefore, researchers need to be certain that their geocoded data are accurate and can be used appropriately in their research projects.

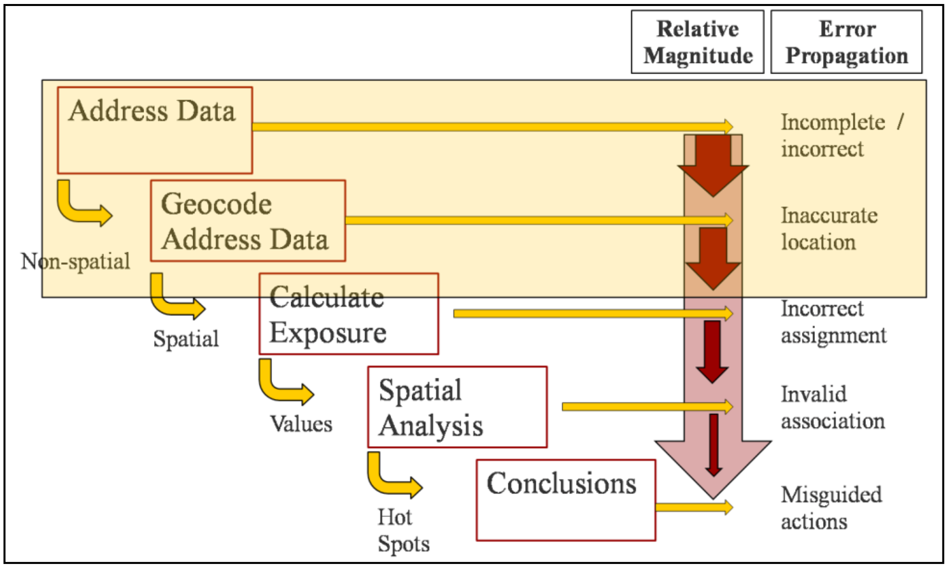


Figure 1.4: The role of geocoding and the potential for error introduction

1.2 Motivation and Problem Statement

Given its widespread use within many disciplines, there has been considerable research into both defining the technical aspects of the process and methods for improving its accuracy (Bakshi et al. 2004; Beyer et al. 2008; Block 1995; Cayo et al. 2003; Davis Jr. et al. 2007; Gilboa et al. 2006; Goldberg et al. 2007; Hurley et al. 2003; Karimi et al. 2004; Krieger, Waterman et al. 2002; Zandbergen 2008a; Zandbergen et al. 2007). Accordingly, a substantial amount of effort has been expended into determining how and why these errors and inaccuracies occur in the geocoding process and what affects these errors and inaccuracies may have upon subsequent studies. A particularly large effort has been undertaken to show that errors and inaccuracies in geocoded results may be systematic and/or non-random and can introduce bias along demographic and geographic boundaries if not accounted for (Gabrosek et al. 2002; Gilboa et al. 2006;

Hurley et al. 2003; Krieger et al. 2001; Krieger, Waterman et al. 2002; McElroy et al. 2003; Rushton et al. 2006; Sheehan et al. 2000; Zandbergen 2007; Zandbergen et al. 2007). Because of this, in all situations where geocoded data are used the consumer of these data must be able to determine fitness-for-use with regard to the requirements for his/her particular study or application. Without this knowledge, it is impossible for a user to determine if the geocoded data can and/or should be used to investigate spatial phenomena through the use of spatial analyses. Further, in the absence of quality measures about the underlying geocoded data, it is not clear whether the conclusions drawn from spatially-based research can and/or should be considered valid and/or meaningful.

In order to facilitate reliability and trust in studies and applications using geocoded data (as well as any products or conclusions derived or drawn from them), consumers must be furnished metadata describing quality characteristics about each individual geocode. To address these needs, prior academic research and commercial endeavors have created geocoding platforms that return varying degrees of descriptive metadata along with a geocoded result within two classes: (1) the level in the feature hierarchy matched to (Hofferkamp et al. 2008), and (2) the quality of the match (Boscoe 2008).

Despite this prior work, geocode consumers are still faced with geocoded data that: (1) have quality metrics that are not adequately descriptive enough of how the geocode was produced to be directly applicable in determining fitness-for-use, and (2) continue to be of low quality. For example, a typical quality description researchers encounter is that the address of the University of Southern California GIS Research Laboratory, “3620 S. Vermont Ave, Los Angeles, CA 90089-0255, Kaprielian Hall Room 444” was geocoded “to a street-level geographic point”, and “geocoded with 83.04% certainty”. Although these two pieces of metadata do provide some insight into the quality of this particular geocode with regard to the type of reference feature selected and the quality of that selection, they constitute poor quality metrics and provide limited information for determining the fitness-for-use of this result for one’s application. Specifically, even when these somewhat detailed pieces of process information are returned along with a geocoded result, they are meaningless when one attempts to determine the true uncertainty that is associated with a geocode. These quality descriptions do not indicate any quantitative values that can be used, for example, to: (1) determine a proportional exposure estimation value resulting from a point source pollutant at a known location based on the likelihood that the geocode is correct at a specific location, or (2) provide upper and/or lower error bounds on the length of a shortest path between geocoded locations. These example use cases are currently outside the realm of possibility because the example quality description above lacks a sense of a discrete region within which one can be sure the true geocode falls, along with a likelihood of the geocode being at any particular location within that region based on the solution space identified during the production of the geocode.

1.3 Geocoding Quality Metrics

The classic definition of geocode accuracy is spatial error, a quantitative measure of the distance between a computed geocode and the location known ground truth position (Karimi et al. 2004; Whitsel et al. 2006; Zandbergen 2008b; Zimmerman et al. 2010). This distance is often used to measure advancements to every aspect of the geocoding process from the underlying reference data files (Wu et al. 2005), to the feature matching algorithms (Christen et al. 2005; Christen et al. 2004; Churches et al. 2002; Jaro 1989), to the feature interpolation methods (Bakshi et al. 2004). This notion is used to show that applying some new approach reduces the distance between the computed value and the true value, either for a single record, or when viewing the spatial error in a large dataset as a whole. This metric describes an average straight-line distance and direction between every computed geocode and its corresponding true location in a dataset to provide some degree of spatial accuracy information (Zimmerman et al. 2010; Zimmerman et al. 2007).

In contrast, the spatial uncertainty associated with a geocoded location describes a quantitative measure of the number of other locations that are all equally likely to be the true location (as described in **Chapter 4**). This metric can be used to provide a spatially-based approximation of a region around the computed geocode within which the true geocode is most likely to be located. This approach represents the notion that choosing any location at random within this area will have the same probability of being the correct location.

A separate metric often used to describe the accuracy of attempting to geocode a dataset of input address data is known as the match rate, or the proportion of data that were able to be geocoded. Research has shown that this metric is often related to the character of either the input address or the type of region the addresses stem from (McElroy et al. 2003; Zhan et al. 2006; Krieger et al. 2001; Ratcliffe 2004) and can have a biasing effect on studies that utilize it alone as the sole indicator of geocode quality (Zandbergen et al. 2007; Shi 2007; Beyer et al. 2008; Krieger, Waterman et al. 2002; Grubesic 2008).

The most important distinction between these three metrics is that the spatial error of a geocode provides an indication of how far away the computed value is from the true value, while the uncertainty metric describes the size and shape of the region within which the true geocode resides, and a the match rate describes the recall of a geocoder without any spatial metric at all. Methods that strive to improve the first of these aims seek to lower the average distance that any computed geocode will be from its true location. Methods that strive to improve the second strive to shrink the region that the true geocode should be located in. In doing so, these methods result in less overall area which could be chosen at random as the output location, thus lowering the uncertainty with all locations and raising the probability that any location chosen at random is correct. Methods that strive to improve the third simply aim to increase the number of matches found in input dataset, typically without regard to resulting spatial accuracy.

The research presented herein does attempts to tackle specific aspects of each of these geocode quality metrics. In what follows, we offer approaches geared toward improving the match rates and the spatial error and spatial uncertainty associated with a geocoded address. In doing so, the approaches we present are explicitly aimed at improving the location of geocoded data in contrast to a ground truth equivalent (spatial error) as well as shrinking the region a geocode is believed to be in order to reduce the number of other location that could be equally likely (spatial uncertainty). To accomplish these goals in, we utilize several spatially-based approaches to explain, understand, and reduce both the spatial error and spatial uncertainty in geocoded data.

1.3 Thesis Statement

In this thesis I test the hypothesis that by using spatially-based approaches to understand, describe, and model and the potential sources of spatial error and uncertainty in the geocoding process, one can develop improvements to the underlying components of geocoding systems which significantly enhance the quality of geocoded datasets and provide researchers the quantitative metrics needed to identify fitness-for-use.

1.4 Contributions of the Research

The key contribution of this thesis is a geocoding framework performs better than current state-of-the-art approaches and supports the computation of quantitative metrics describing the accuracy of a resultant geocode based upon the process by which it was created. The framework includes the following contributions:

- An exhaustive derivation of the sources and scales of potential spatial error and uncertainty associated with the geocoding process.

- A novel feature matching component that uses a spatially-varying neighborhood metrics to dynamically score nearby candidate reference features.
- A novel method to combine multiple layers of reference features using uncertainty-, gravitationally-, and topologically based-approaches to derive the most likely candidate region.
- A novel feature matching and interpolation technique called composite feature matching that reduces the spatial uncertainty in geocoded data.
- A novel tie-breaking strategy that picks the appropriate candidate reference feature based on the characteristics of the local region surrounding and the relationships between ambiguous candidate reference features.
- An open source and extensible geocoding framework for developing, testing, and evaluating geocoding techniques.

1.5 Outline of the Dissertation

The remainder of this dissertation is organized as follows. **Chapter 2** provides a detailed background to the current state-of-the-art in geocoding techniques through an in-depth review of the rich existing literature related to the topic of geocoding, both historic and current. **Chapter 3** provides a thorough account of the major components of the USC geocoding system developed as part of this dissertation and describes our novel method for computing quantitatively-based feature match scores for nearby candidate reference features using a spatially-varying block distance metric. **Chapter 4** develops the concept of a best-match criterion and describes our novel uncertainty-based approach to resolving the most likely output location for a geocode based on the complete set of candidate features available, their inherent spatial uncertainties, and their topological relationships to one another. **Chapter 5** outlines the existing approaches taken to handle ambiguous geocode results and presents our novel method for tie-breaking using both a set of rules about the known relationships between candidate feature attributes and characteristics about the local geographic regions surrounding each ambiguous candidate reference feature. **Chapter 6** concludes the dissertation by highlighting key findings and describing the potential for future work.

CHAPTER 2: FROM TEXT TO GEOGRAPHIC COORDINATES – A REVIEW

This chapter was published as:

Goldberg, D.W., Wilson, J.P. and Knoblock, C.A. 2007.

From text to geographic coordinates: The current state of geocoding.

URISA Journal 19(1), 33-47.

2.1 Chapter 2 Introduction

The process of geocoding forms a basic fundamental component of spatial analysis in a wide variety of research disciplines and application domains, e.g., health (Vine et al. 1998; Boulos 2004; Rushton et al. 2006); crime analysis (Olligschlaeger 1998; Ratcliffe 2001; Haspel et al. 2005); political science (Haspel et al. 2005); computer science (Hutchinson et al. 2005b; Bakshi et al. 2004). This act of turning descriptive locational data such as a postal address or a named place into an absolute geographic reference has become a critical piece of the scientific workflow. However, the geocoding of today is a far cry from the geocoding of the past. Geocoding data that used to cost $4.50 per 1,000 records as recently as the mid-1980s (Krieger 1992) quickly moved to $1.00 per record in 2003 (McElroy et al. 2003), and can now be done for free with online services (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b; Locative Technologies 2009; Goldberg 2009), with far greater spatial accuracy and match rates.

As the availability and accuracy of reference datasets have increased over the past several decades (Dueker 1974; Werner 1974; Griffin et al. 1990; Higgs et al. 1995; Martin et al. 1996; Johnson 1998a; Martin et al. 1999; Boscoe et al. 2004), geocoding has undergone marked transitions to accommodate and exploit changes in both data format and user expectations. These transitions can clearly be seen in the input, output, and internal processing of the geocoding process. The input data suitable for geocoding have expanded from simple postal addresses (O'Reagan et al. 1987) to include textual descriptions of relative locations (Levine et al. 1998; Davis Jr. et al. 2003; Hutchinson et al. 2005b). The output capabilities of the geocoding process have moved from simple nominal geographic codes (Tobler 1972; Dueker 1974; Werner 1974; O'Reagan et al. 1987) to full-fledged three-dimensional (3-D) geospatial entities (Beal 2003; Lee 2004). Likewise, the internal processing mechanisms that produce the geographic output have moved from simple feature assignment (O'Reagan et al. 1987) to complex interpolation algorithms using a variety of heterogeneous data sources (Bakshi et al. 2004; Hutchinson et al. 2005a, 2005b).

While significantly improving the usability, reliability, and accuracy of the geocoding process, these developments have brought with them a host of issues that a potential user must recognize and be prepared to contend with. Specific issues include the assumptions made during the interpolation process (Dearwent et al. 2001; Karimi et al. 2004), the underlying accuracy of the reference dataset (Gatrell 1989; Block 1995; Drummond 1995; Martin et al. 1996; Chung et al. 2004), the uncertainty in the matching algorithm (O'Reagan et al. 1987; Jaro 1984), and the choice of areal unit geocoded to (Krieger 1992, 2003; Krieger et al. 2003; Geronimus et al. 1998; Geronimus et al. 1995). These topics have received considerable research in recent times, and a great deal of literature is available.

This chapter will survey the field of geocoding through a cross-disciplinary study of the geocoding literature focusing foremost on the technical aspects of the process. The changing concept of geocoding will be described, and the fundamental components of the geocoder will be outlined. Potential sources of error in the geocoding process will be explored, and particularly difficult geocoding scenarios requiring further research will be highlighted. The primary contributions of this research will be to inform the reader of the state-of-the-art in geocoding through a discussion of its evolution over time and to warn of potentially sticky situations that can arise in the geocoding process if one is not aware of how one's decisions and assumptions can affect the geocoded results.

This work should be seen as distinct from the recent work published by (Rushton et al. 2006), which also offers a review of the geocoding process, but is focused on its application to health research, in particular cancer studies. Their work takes a narrow and limited view of geocoding and does not delve so deeply into the evolution or technical aspects of the geocoding process as does that presented here. As such, this chapter can be seen as a more comprehensive, technically targeted, broadly visioned journey through the geocoding process and should be used as companion material to field-specific reviews such as that of (Rushton et al. 2006).

2.2 The Concept of Geocoding

Over the years, the changing availability of geographic data has forced the concept of geocoding to remain flexible and adaptive in terms of its requirements and capabilities. The increasing availability, accuracy, and reliability of digital geographic reference datasets has meant that the geocoding process has continually evolved to keep pace with the underlying datasets that facilitate its use. As such, practitioners have been pushing the boundaries of what types of information can be geocoded using different information sources from the very beginning. Early geocoding systems used by the U.S. Census in the 1960s simply turned postal addresses and named buildings into geographical zones delineated by numerical codes (O’Reagan et al. 1987), not the valid geographic objects such as points, lines, areas, or surfaces with which consumers of geocoded data are accustomed to today. More modern attempts at geocoding have tackled the problems of assigning valid geographic codes to far more types of locational descriptions such as street intersections (Levine et al. 1998), enumeration districts (census delineations) (Sheehan et al. 2000), postal codes (zip codes) (Gatrell 1989; Collins et al. 1998; Sheehan et al. 2000; Krieger, Waterman et al. 2002; Hurley et al. 2003), named geographic features (Davis Jr. et al. 2003; United Nations Economic Commission 2005), and even freeform textual descriptions of locations (Hutchinson et al. 2005a, 2005b; Wieczorek et al. 2004).

These fundamental shifts in geocoding attitudes and opportunities can be traced directly to the technological advances made to the underlying reference datasets on which they are based. The early attempts at geocoding were hindered by the lack of digital geographies to use in the assignment of codes, and were limited by their use of flat text-based files. This resulted in low-resolution non-geographic output, turning addresses and building names into the census block to which they belonged. The development of true digital geographies in the form of products such as the U.S. Census Bureau’s Dual

Independent Map Encoding (DIME) files enabled the assignment of true geographic codes, but their structure limited the processing that could be applied to derive the output. The introduction of the vector-based geographic datasets such as the U.S. Census Bureau's Topographically Integrated Geographic Encoding and Referencing (TIGER) (U.S. Census Bureau 2009b) database have enabled new generations of geocoding algorithms to approximate representations for the geographic output using interpolation-based approaches, greatly increasing the resolution of the geographic output (Dueker 1974; O'Reagan et al. 1987; Martin 1998; Ratcliffe 2001; Nicoara 2005).

Taking this a step further, the creation of precompiled geocoded national address registers such as the ADDRESS-POINT (Ordnance Survey 2010) and G-NAF (Paull 2003) databases in the United Kingdom and Australia, respectively, have facilitated highly precise geocoding capabilities at national scales (Higgs et al. 1995; Martin 1998; Ratcliffe 2001; Churches et al. 2002; Higgs 2002; Christen et al. 2005; Christen et al. 2004; Murphy 2005). Furthermore, the emergence of high-resolution digital parcel and property boundary files may enable even more accurate digital geographic results to be returned (Dueker 1974; Olligschlaeger 1998; Dearwent et al. 2001; Ratcliffe 2001; Rushton et al. 2006), but these developments are pushing the limits of what form the output of geocoding should take. Likewise, the development of multiresolution gazetteers defining geographic footprints for named geographic places such as the Alexandria Digital Library Gazetteer (Frew et al. 1998; Hill 1999; Hill et al. 2000) are pushing the limits of what type of geographic features can have geographic codes assigned to them (Davis Jr. et al. 2003; United Nations Economic Commission 2005), as well as the role of the geocoder in the larger geospatial information-processing context. The proliferation of a variety of diverse types of locational addressing systems throughout the world precludes a “one size fits all” geocoding strategy that will work in all cases (United Nations Economic Commission 2005; Fonda-Bonardi 1994; Lind 2001; Davis Jr. et al. 2003; Walls 2003).

The result of this evolution is a somewhat “fuzzy” concept of geocoding, tailored to the specific requirements and data availability of the person performing the geocoding. For example, almost everyone involved in or using geocoding today would agree that turning a postal address into a geographic point is most certainly included in the set of geocoding operations. Likewise, they would probably agree that turning a portion of the postal address such as the post code (ZIP code) into a geographic point or polygon is also part of the geocoding process. However, continuing this line of reasoning presents a slippery slope because a series of fundamental questions arise. What should the point returned as representative of the postal code be? Should it be the center of mass (centroid)? Should it be weighted by the population distribution? Furthermore, if the digital boundary of the postal code is available, why not return it instead of just a single point? Questions such as these are just the beginning. If the postal code can be geocoded, can the city be as well? If so, what is the difference between the geocoder returning a geographic representation of the city and the gazetteer doing the same? And if they are, in fact, performing the same operation, why is it commonly understood that a gazetteer can provide geographic representations for a wide variety of geographic features such as rivers, mountains, and shorelines, while these are seldom thought of as candidates for the geocoding process? We can see through this discussion that the term geocoding can mean different things to different people, and their perception will be based on their primary experience or usage with a particular geocoding tool. To some, “geocoding” is synonymous with “address matching” (Bonner et al. 2003; Drummond 1995; Vine et al. 1998), highlighting its prevalent use of transforming postal addresses into geographic representations (Drummond 1995, p. 250). For others, “geocoding” is understood to produce a valid geographic output, but its input is not necessarily limited to simple postal addresses, e.g., (Levine et al. 1998), and still further distinctions can be drawn between the two terms (Johnson 1998a, p. 25). Taken literally, geocoding means “to assign a geographic code.” This definition stems from the two root words: geo, from the Latin for earth, and coding, defined as “applying a rule for converting a piece of information into another” and is similar to that defined early on in the geocoding literature (Dueker 1974, p. 320). Notice that this literal definition does not imply nor constrain in any way the input to the geocoding system, the processes or data sources used to assign the geographic code, or even what the geographic code returned as output must be. It is precisely this relaxation of formal constraints on the geocoding process that has allowed it to mature and prosper to the many forms that we use today, and that will in turn drive the technological advances of tomorrow.

2.3 Geocoding Fundamentals

Even with this varied notion of geocoding, it is still possible to characterize it in terms of its fundamental components: the input, output, processing algorithm, and reference dataset (Levine et al. 1998; Karimi et al. 2004; Yang et al. 2004; Nicoara 2005). The input is the locational reference the user wishes to have geographically referenced that contains attributes capable of being matched to some datum that has been previously geographically coded. The most common data to be geocoded are postal addresses. In fact, there are very few geocoding services that geocode anything other than postal address data. The simple reason for this is that postal address data are among the most prevalent forms of information (Eichelberger 1993), and address geocoding is cited often throughout the literature as a national health goal that will “be the basis for data linkage and analysis in the 21st century” (U.S. Department of Health and Human Services 2000, goal 23-3). Address data are how people locate, situate, and navigate themselves, and are presently the easiest method by which to describe one’s location (Walls 2003). In the future when all cellular phones come equipped with reliable global positioning system (GPS) units and all homes and businesses are geographically referenced with coordinates available via wireless location-based services, the postal address may, in fact, become obsolete. But for the foreseeable future, the postal address will remain the critical and ubiquitous data throughout most forms of information processing.

As previously noted, however, address data are not the only type of locational data that can or should be geocoded. Even the earliest geocoding systems of the U.S. Census accounted for the geocoding of named buildings (O’Reagan et al. 1987), but the task of associating geocodes with geographic features other than addresses is most commonly associated with the services provided by a gazetteer (Hill 2000). The problem with this, though, is that a gazetteer typically does not contain the functionality to generate the geocodes that it returns, instead acting as a storage mechanism after the geocodes have already been determined using other methods. As such, the geocoder is commonly employed to produce the geocodes for features in the gazetteer that are address-based, emphasizing the crucial connection between the two components as part of a larger spatial query and analysis framework. This situation is displayed in Figure 2.1, where the geocoder is shown to be one of many possible sources of footprint data for a gazetteer, with itself being composed of several data sources.

The output is the geographically referenced code determined by the processing algorithm to represent the input. In most situations, the output is a simple geographic point, but nothing forbids it from being any valid type of geographic object. The development of detailed spatial datasets enables the output of increasingly detailed multidimensional geographic features, including the emergence of 3-D indoor geocoding solutions (Beal 2003; Lee 2004).

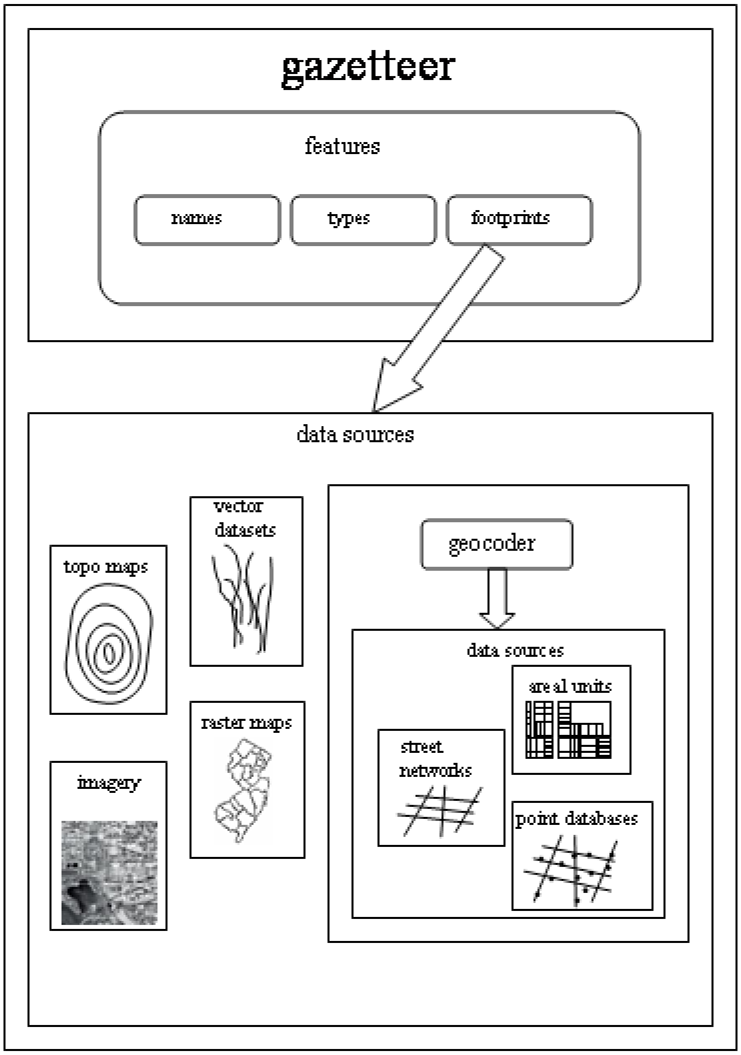


Figure 2.1: Relationship between gazetteer and geocoder

The processing algorithm determines the appropriate geographic code to return for a particular input based on the values of its attributes and the values of attributes in the reference dataset. This is by far the most complicated portion of the geocoding process in which the most research has been invested. The key topics involved in the process include the standardization and normalization of the input into a format and syntax compatible with that of the reference dataset (Johnson 1998b; Churches et al. 2002; Nicoara 2005; Laender et al. 2005), the matching algorithm that picks the best feature in the reference dataset (Bakshi et al. 2004; Davis Jr. et al. 2003; Drummond 1995; Vine et al. 1998), and the final geocode generation mechanism that determines what to return based on the reference feature selected as the best match (Cayo et al. 2003; Davis Jr. et al. 2003; Drummond 1995; Levine et al. 1998; Ratcliffe 2001).

Figure 2.2 shows a schematic diagram of how a simple deterministic processing algorithm could proceed using standardization, normalization, and attribute relaxation. The standardization and normalization process can vary in complexity from simple token parsing with lookup tables for standardizing abbreviations to advanced probabilistic methods using machine learning techniques such as hidden Markov models that can handle attribute misspellings and misplacements (O’Reagan et al. 1987; Fulcomer 1998; Christen et al. 2005; Christen et al. 2004; Churches et al. 2002; Nicoara 2005; Yang et al. 2004). In general, the key role performed in this step is to determine what each piece of the input is and to turn each into versions consistent with those in the reference dataset. Once the input has been sufficiently massaged to be compatible with the reference dataset, the matching process picks the best candidate to be used to derive the final output.

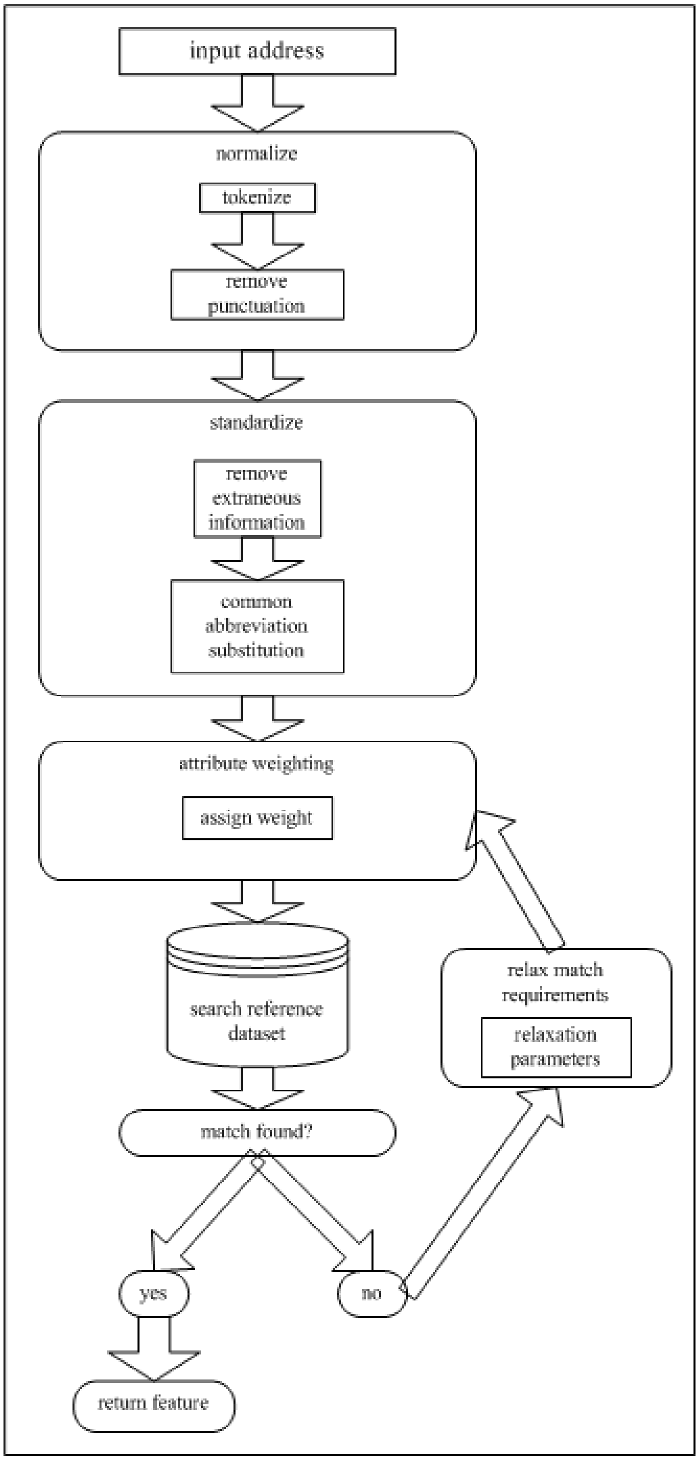


Figure 2.2: Schematic of deterministic address matching with attribute relaxation

Tricks such as word stemming, using Soundex, and relaxing the requirement of matching all attributes can be used to improve the probability of finding a match in the reference dataset (Drummond 1995; Fulcomer 1998; Johnson 1998a; Levine et al. 1998; O'Reagan et al. 1987; Gregorio et al. 1999; Boscoe et al. 2002; Beal 2003; Christen et al. 2005; Christen et al. 2004; Churches et al. 2002; Nicoara 2005; Yang et al. 2004). Here the issue may arise that zero, one, or more than one reference features can be the best possible match.

In the case of one match, the algorithm will use it to determine a geocode. In the case of zero, the matching algorithm may prompt the user for more information, attempt to geocode at a lower resolution with additional datasets, or try to find additional information in other datasets to enable a match (Laender et al. 2005). Likewise, in the case of multiple matches, the algorithm may prompt the user to determine the appropriate one or consult additional datasets for more information to use in breaking the tie (Hutchinson et al. 2005a, 2005b). **Chapter 5** of this dissertation deals with the challenges and opportunities available in such instances.

In any case, once the appropriate reference feature has been selected, the algorithm must determine the appropriate geocode for output based on the input and the reference feature. In the case of a precompiled geocoded dataset such as the ADDRESS-POINT (Ordnance Survey 2010) and G-NAF (Paull 2003), the algorithm can simply return the existing geographic representation. However, in the case of U.S. Census Bureau TIGER/Line files (U.S. Census Bureau 2009b), the output geography must be derived based on the line segment determined to be a match. Here interpolation algorithms deduce the appropriate output geography based on attributes of the street segment such as address ranges and polarity (Cayo et al. 2003; Davis Jr. et al. 2003; Drummond 1995; Levine et al. 1998; Ratcliffe 2001).

In general, these interpolation algorithms work by first identifying the correct street segment in the reference data source based on the attributes of the address to be geocoded and the attributes of the street segment (address ranges associated with both sides of the segment, street name, street suffix, etc.). Once found, the appropriate side of the street segment is ascertained using the polarity (even/odd) of the address and each of the street segment sides. The correct location along the street segment is then determined by computing where the addresses in question would fall as a proportion of the total address range associated with the appropriate side of the street segment. This proportion is then applied to the total length of the street segment to obtain a location along the centerline of the street, and additional parameters such as distance and direction from the street center and offset from the endpoints of the street can be introduced to further improve the accuracy (Cayo et al. 2003; Ratcliffe 2001). Additional data sources can be consulted to obtain knowledge about the number of parcels on the street and their geographic distribution (Bakshi et al. 2004) to overcome the parcel homogeneity assumption (Dearwent et al. 2001) that all parcels within an address range truly exist and have the same dimensions. In Figures 2.3 through 2.6 these points are illustrated.

Figure 2.3 shows the parameters for the interpolation algorithm, d and θ, the street centerline offset distance and angle, q, the corner offset distance, and v, the interpolated distance to the center of the parcel. Also shown are the address ranges for each side of the segment, 601 through 649 on the odd parity side, and 600 through 648 on the even parity side. Figure 2.4 shows a sample block segment with the geocoded position of 631 Main Street displayed. Figure 2.5 displays how the parcel homogeneity assumption divides the segment into equal portions for all addresses within the range of the street segment, placing the geocoded point for address 631 at the wrong location (shown as ring) compared to the true location (shown as shaded ring). Figure 2.6 also displays the parcel homogeneity assumption, but in this case the true number of parcels on the street is known and the resulting geocoded point for address 631 is at a closer location (shown as ring) to that of the true location (shown as shaded ring). When using area-based reference features such as postal code and parcel polygons to compute point geographies to return as output, the algorithm must calculate an appropriate centroid (Stevenson et al. 2000; Dearwent et al. 2001; Ratcliffe 2001). It may simply return the center of mass of the object, or it may perform more complex calculations in conjunction with other information such as population distributions across an area to determine a more representative weighted centroid (Gatrell 1989; Durr 2002).

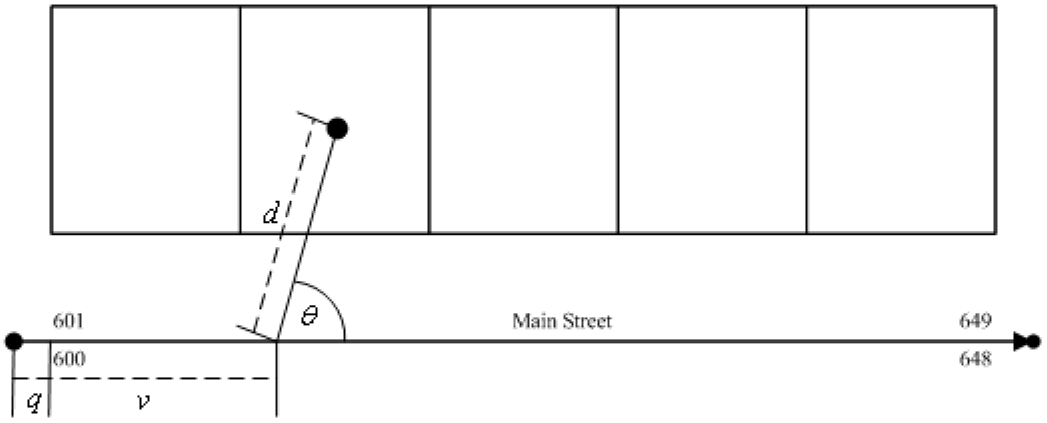


Figure 2.3: Sample block showing parameters of the geocoding algorithm

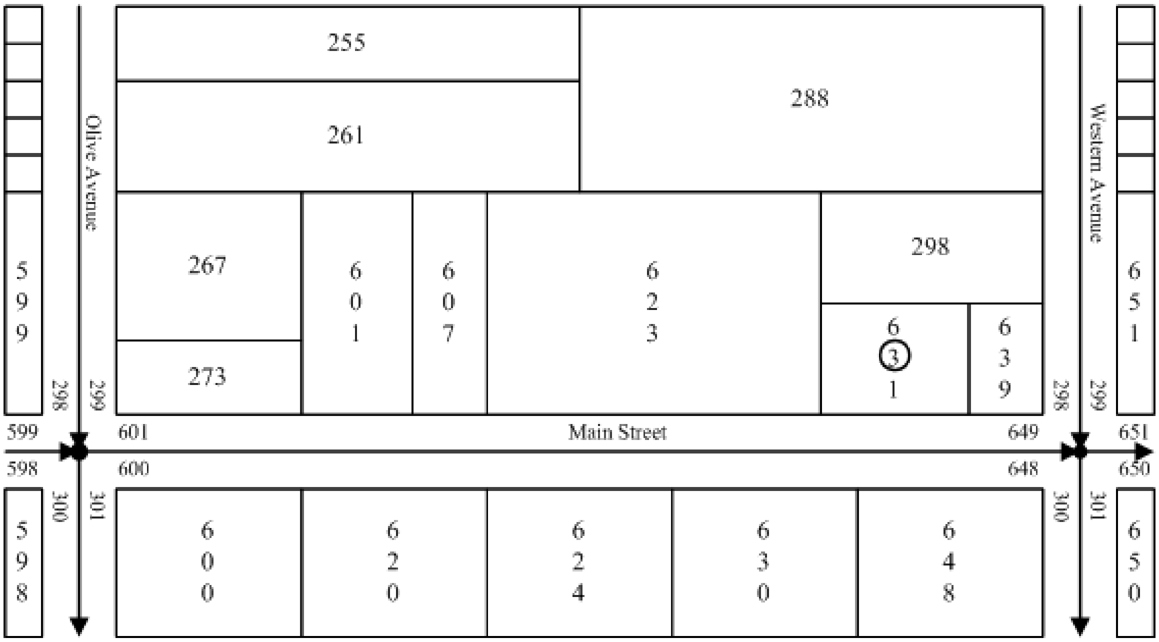


Figure 2.4: Sample address block with true parcel arrangement showing true geocoded point as ring

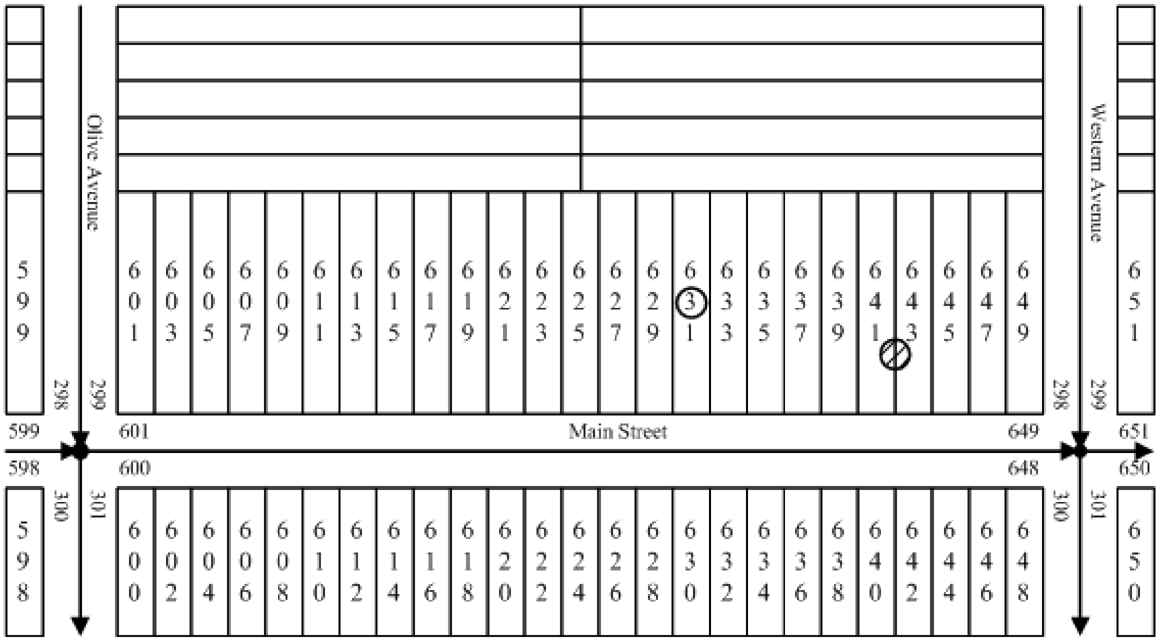


Figure 2.5: Sample address block with parcel homogeneity assumption using address range showing erroneous geocoded point as ring and true geocoded point as shaded ring

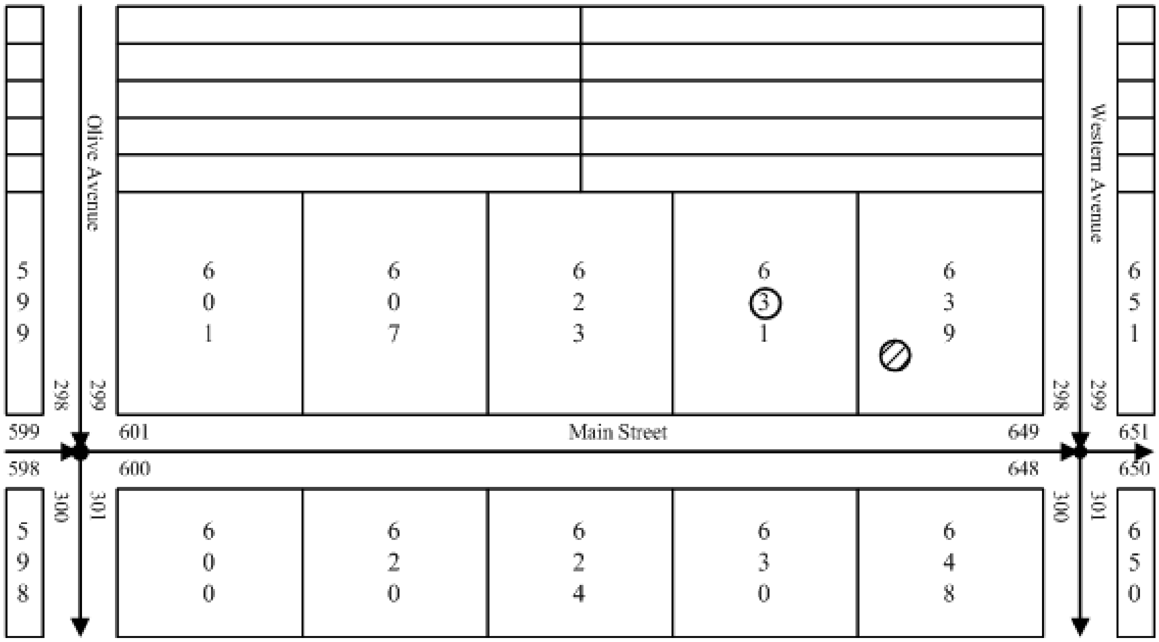


Figure 2.6: Sample address block with parcel homogeneity assumption using actual number of parcels showing erroneous geocoded point as ring and true geocoded point as shaded ring

The reference dataset consists of the geographically coded information that can be used to derive the appropriate geographic code for an input. As noted earlier, the datasets used as geocoding reference files have changed rapidly over time and are responsible for driving new technological breakthroughs in geocoding methodologies. The early datasets of text-based lists have given way to true digital geographic datasets, and are rapidly moving toward advanced 3-D representations. The underlying advances in terms of efficient storage, retrieval, and indexing have allowed these datasets to grow expansively in size, detail of resolution, and speed of access. The only constraint on these datasets is that they need to maintain attributes in a consistent fashion throughout, so that the standardization and normalization algorithms can work toward transforming the input data to be appropriate for finding a match.

2.4 Geocoding Error

This broad definition of geocoding also brings with it a significant burden in the form of anticipating and/or quantifying geocoding error. Even simply defining what the error of the geocoding process is presents an arduous task. When speaking of geocoding error, is reference made to the positional accuracy of the returned geographic object, the probability that the feature returned is the one that was desired, or the validity of one or more assumptions used by the geocoding algorithm? Further definitions could include the error caused by the match rate, the weighting and relaxation techniques used in the standardization process, or the confidence cutoffs used during probabilistic matching.

Common causes and effects of errors in each stage of the geocoding process are listed in Table 2.1.

Table 2.1: Common causes and effects of errors in stages of the geocoding process

Stage	Cause of error	Effect of error
Matching		
	Attribute relaxation	Incorrect feature
	Probabilistic confidence level	Incorrect feature
Derivation		
	Parcel homogeneity assumption	Wrong distribution
	Address range existence assumption	Wrong number
Reference Data		
	Spatial accuracy	Results inaccurate
	Temporal accuracy	Results inaccurate

It becomes obvious from this (not even close to exhaustive) list of commonly described error metrics that evaluating the error associated with a geocoded result is difficult at best, and at worst not even taken into consideration. It is an unfortunate reality that even though a broad range of literature exists specifically geared to exposing how minor error in geocoding accuracy can affect results based on detailed spatial models (Gatrell 1989; Ratcliffe 2001; Higgs 2002; Bonner et al. 2003; Cayo et al. 2003; Krieger 2003; Krieger et al. 2005), recent research initiatives continue to employ geocoded data without regard for how the accuracy can introduce possible inconsistencies or bias into the results (Diez-Roux 2001; Brody et al. 2002; Haspel et al. 2005).

Several studies have attempted to quantify the error associated with the geocoding process, highlighting error introduction from specific aspects of the geocoding process (Davis Jr. et al. 2003; Karimi et al. 2004). On evaluating a potential geocoding strategy, one should consider several key factors to determine if the outcome will meet the needs. First, what areal unit will the data be geocoded to? Will the output be to the granularity of individual postal addresses, or will it be to a larger delineation such as a census block or zip code, and will the implicit aggregation of using a larger unit have an effect on the results? This decision is a divisive topic in the geocoding literature and several studies have demonstrated that areal unit choices both have an effect and do not have an effect on the outcomes of the results (Geronimus et al. 1998; Geronimus et al. 1999a, 1999b; Geronimus et al. 1995; Krieger et al. 1999; Smith et al. 1999; Soobader et al. 2001; Krieger, Chen et al. 2002; Krieger et al. 2003; Gregorio et al. 2005). Evaluating one's confidence in the available scholarship will require personal judgment to determine if this could be an issue given a particular dataset and research objective.

Second, how accurate is the underlying data used as the reference dataset? Included in this discussion should be the concepts of spatial accuracy – how close are the features in the dataset to what is found on the ground (Karimi et al. 2004; Wu et al. 2005)?, temporal accuracy – how close are the features in this dataset to how they were at the time period of interest to me (McElroy et al. 2003; Han et al. 2005)?, original collection purpose – what were these data originally collected for (Boulos 2004)?, and lineage – what processes have been applied to this data (Veregin 1999)? These aspects may be difficult to quantify because the accuracy measurements associated with datasets are estimates over the entire dataset, not on a per-feature basis. For example, while achieving an acceptable accuracy for short street segments in urban areas, the U.S. Census Bureau TIGER/Line files (U.S. Census Bureau 2009b) most commonly used for linear interpolation geocoding in the United States are known to be far less accurate for geocoding in rural areas with longer street segments (Bonner et al. 2003; Cayo et al.

2003; Drummond 1995; Vine et al. 1998; Wu et al. 2005). Assuming a consistent accuracy value for a dataset throughout the entire area of coverage is rarely discussed or noted as a point of contention in the determination of geocoding accuracy.

A third related issue arises when one considers multi-tiered geocoding approaches using multiple data sources. For example, in numerous instances, geocoding match rates in rural areas are far less than in urban areas (Gregorio et al. 1999; Kwok et al. 2001; Bonner et al. 2003; Boscoe et al. 2002; Cayo et al. 2003). The typical approach to solving this problem involves a decision of whether to geocode to a less precise level or to include additional detail from other sources to determine the correct geocode. Choosing either case creates a resulting dataset with varying degrees of accuracy as a function of location, a condition recently defined as “cartographic confounding” (Oliver et al. 2005) that has been alluded to many times, yet remained undefined throughout the history of geocoding research (Block 1995; Cayo et al. 2003; Gregorio et al. 2005; Ratcliffe 2001, 2004; Nuckols 2004). A per-geocode accuracy is rarely maintained as a result of the geocoding process other than the level of geography matched to (i.e., census tract versus block group), and rarely do spatial models include variables to model this phenomena, although some researchers have begun developing models to account for it (Openshaw 1989; Arbia et al. 1998; Cressie et al. 2003; Gabrosek et al. 2002). Despite this, information describing the varying degrees of accuracy of each individual geocode is not typically represented during subsequent spatial analysis.

Fourth, one needs to determine if the assumptions made by the geocoding algorithm are applicable to one’s needs. As previously mentioned, the most common form of geocoding, linear interpolation–based, makes several key assumptions that can affect the level of accuracy of the results. First, it assumes that all addresses within an address range exist. Thus, when it determines the correct location for a particular address along a street segment by identifying the proportion along the segment where an address should fall, it will overestimate the number of addresses placing it at the wrong location. Second, it assumes a homogeneous distribution of addresses in terms of lot placement and size, known as the parcel homogeneity assumption (Dearwent et al. 2001). This means that each lot on the street is assumed to have the same dimensions, and oriented in the same direction, which is typically not a realistic assumption. Furthermore, it does not take into account that the corner lot on a segment may belong to the segment in question, or to the segment that forms the corner (Bakshi et al. 2004). While the magnitude of error introduced by these assumptions is small, on the order of half the length of the street segment (Wu et al. 2005, p. 596), it can have dramatic effects when the variable and/or relationships of interest, e.g., environmental exposure doses to pesticide (Brody et al. 2002; Kennedy et al. 2003), air pollution (Wu et al. 2005), or proximity to voting precincts (Haspel et al. 2005) vary over tens or hundreds of meters, and becomes amplified as the landscape becomes more rural. Additionally, it has been shown that when geocodes are used for point-in-polygon operations to derive attributes from other datasets, small spatial errors in geocodes that lie along borders between the larger level features can cause serious misclassifications in combined data (Ratcliffe 2001; Schootman et al. 2004).

Fifth, one needs to consider the uncertainty created by the aggregation or randomization performed on the resulting point to protect the identity of the geocoded object. This is most often the case in the geocoding of health data, where confidentiality requirements necessitate the geocode for an individual's location to be non-identifying. Research has shown that there are ways to trade off between the usefulness of data returned for spatial analysis versus specific confidentiality requirements, but further work is required to quantify the effect of this in a geocoding context (Armstrong et al. 1999). For a more thorough description of the issues involved specifically geared toward health research, refer to (Boscoe et al. 2004) and (Rushton et al. 2006).

Finally, one needs to determine if the intended spatial analysis can deal with uncertain geographic values or not. Here a fundamental decision must be made whether probabilistic matching methods can be used or strictly deterministic ones (O'Reagan et al. 1987). When interpreting an input query, the geocoding system must go through several steps to determine the "best" match in the reference dataset (Levine et al. 1998). If the input can be matched directly to an existing geography, it can be returned immediately. However, it is more often the case that one needs to massage the input data and transform it into a format consistent for finding the best match. Locational data, and in particular postal address data, are notoriously "noisy, very often, extraneous information, missing information, or confusing non-standardization is contained in the input (Fulcomer 1998; Ratcliffe 2001, 2004; Murphy 2005; Nicoara 2005). In these cases, the geocoding algorithm is forced to either attempt to correct the input so that a match can be found or return a non-match. It has been shown that with deterministic approaches such as relaxing the constraint that all attributes must match exactly and allowing partial matches with a variety of attribute weighting schemes, a higher match rate can be achieved, but at the price of accuracy. In particular, studies have found that relaxing the street name portion of an address will greatly reduce the accuracy of the geocoded results (Lixin 1996; Bonner et al. 2003; Cayo et al. 2003; Krieger 2003; Rushton et al. 2006). In contrast, probabilistic approaches to standardization (Jaro 1984) have been used since very early on in the geocoding literature with much success (O'Reagan et al. 1987) and continue to improve (Christen et al. 2005; Christen et al. 2004; Churches et al. 2002), but one must recognize the risk that these results may not be accurate as they are relying on the confidence level of their uncertainty measures, and they will in some cases produce erroneous results.

2.5 Persistent Geocoding Difficulties

For all the technological advances and improvements that have been made to the geocoding process and the underlying reference datasets, the geocoding difficulties identified early on still exist. In developing countries with little GIS data infrastructure, the main roadblock to accurate geocoding is the simple nonexistence of reference datasets or GIS data infrastructure (Croner et al. 1996; United Nations Economic Commission 2005). The development of basic GIS reference datasets is hindered by the existence of slum-like areas that change frequently, contain geographic features that are not street addressable, and where many areas lack a consistent addressing scheme (Davis Jr. 1993; Davis Jr. et al. 2003; United Nations Economic Commission 2005; Oppong 1999).

Efforts are under way to remedy these situations by developing standardized addressing systems that include facets for encouraging public participation aimed at promoting acceptance and eventual adoption, but these are costly endeavors being undertaken in areas with few economic resources to dedicate to the task (United Nations Economic Commission 2005).

Even in developed countries such as the United States, the existence of rural addresses and P.O. boxes impose a continual headache for geocoding practitioners (Gregorio et al. 1999; Boscoe et al. 2002; Hurley et al. 2003; McElroy et al. 2003; Schootman et al. 2004; Gaffney et al. 2005; Oliver et al. 2005). In the P.O. box case, it is not possible to determine an accurate geocode because the information available about the address is just not specific enough. The best that one can do is to geocode to a lower resolution such as a postal code centroid, but several studies have explored how this can introduce bias into the results produced with the geocoded data (Sheehan et al. 2000; Krieger, Waterman et al. 2002; Hurley et al. 2003). Research initiatives have recently undertaken creative ways to obtain enough specific information to produce a more accurate geocode by using secondary sources including obtaining the P.O. box renter’s address from the postal service, utility company records, and administrative records from government agencies. These tasks require human intervention and are quite expensive (Levine et al. 1998; Hurley et al. 2003; McElroy et al. 2003; Han et al. 2005). While capable of producing highly accurate results to within a few meters, the practice of using a global positioning system (GPS) technology to record point locations for addresses is an option for producing geocoded results, but this has its limitations (e.g., timeconsuming, expensive, and labor-intensive) (Ward et al. 2005; Bonner et al. 2003). The increasing prevalence of parcel data and its use when GPS data are unavailable is an alternative option that has been proposed throughout the history of the literature (Dueker 1974; Rushton et al. 2006). A recent U.S. government report found that there is an increasing surge in the amount of survey quality digital parcel boundary data becoming available (Stage et al. 2005), with some states actually passing legislation requiring its release (Lockyer 2005), from which accurate centroids could be derived and used as substitutes where GPS data are not available (Ratcliffe 2001).

Likewise, the mandatory introduction of the Enhanced 911 (E911) system in the United States for all structures with telephones is improving geocoding by increasing the number of rural addresses reported as address data and creating more accurate reference datasets (Johnson 1998a; Cayo et al. 2003; Levesque 2003; Oliver et al. 2005; Rose et al. 2004), but historical data frequently used in research are not being updated, so the problem still remains. Again in this case, the geocoding practitioner is forced to obtain secondary information to identify what an appropriate city-style address would be for the location so it can successfully be geocoded. E911 geocoding typically results in an “absolute” geocode, as opposed to a “relative” geocode, as in traditional interpolation-based geocoding. “Absolute” geocoding, as used here, refers to the fact that the resulting geocode is based on a linear addressing system, describing a known point (e.g., a milepost) and the distance one would have to travel to find the actual location from that point.

“Relative” geocoding, in contrast, results in a geocoded result that is an interpolation along or within a geographic feature (e.g., a percentage of the distance along a street segment or the center of mass of a parcel). As people move away from traditional land-line phones with the adoption of cell phone technology, some may argue that the promise of E911 solving addressing issues will begin to disappear. However, while it is true that in the future more calls will undoubtedly be made from cell phones, this is irrelevant for most municipalities still assume that structures will have phones and legislation is often in place that requires the E911 system to be kept up-to-date and accurate. As such, when official addresses are requested for new construction, the department responsible for maintaining the E911 system will most likely be required to visit the property and assign the E911-based geocode for the address.

A further problem, which the evolution of reference datasets may help solve, is that of sub-parcel geocoding. This case occurs when multiple structures are residing on the same land parcel such as in apartment/condominium-type properties and large campuses such as universities and business parks or in the case of large farms where a single small structure may be located somewhere within a much larger parcel. Here geocoding to the centroid of the property may not present sufficient accuracy for the detailed applications previously described (Gaffney et al. 2005). However, including secondary data sources and operations such as high-resolution imagery in conjunction with computer vision techniques to identify and separate buildings may help lead the way in this arena (Hutchinson et al. 2005b). Like all reference data sources though, when employing imagery data in a geocoding solution, one must be aware that the accuracy ultimately achieved can be greatly affected by the preprocessing applied (or lack thereof), typically the rectification and registration processes. For in-depth historical and state-of-the-art reviews, consult (Gottesfeld-Brown 1992), (Pohl et al. 1998), and/or (Toutin 2004). Additionally, integrating and conflating existing detailed maps of campuses (Chen et al. 2003; Chen et al. 2004) may enable the extraction of highly accurate polygons for building footprints, but automating this task is still an open research problem. Of course, the reliance on two-dimensional (2-D) GIS data sources of the traditional and commonly used GIS platforms precludes the ability for highly precise geocoding of 3-D structures with multiple addresses such as multistory buildings.

2.6 Chapter 2 Conclusions

This chapter has explored the state-of-the-art in geocoding through a discussion of the path geocoding and its reference datasets have taken over the years. This work should serve as a starting point from which potential geocoding projects can be undertaken with regard to identifying the potential pitfalls and challenges that are commonly encountered. Each particular geocoding project will have its own requirements in terms of input and output data structure and format, confidentiality, cost, available tools, and technical know-how, but the survey presented here should allow a more thorough understanding of the ramifications of particular choices made during the process.

CHAPTER 3: IMPROVING GEOCODING WITH SPATIALLY-VARYING BLOCK METRICS

This chapter will be submitted for publication as:

Goldberg, D.W., Wilson, J.P. Knoblock, C.A., and Cockburn M.G. 2010.

Improving Geocoding with Spatially-Varying Block Metrics.

GeoInformatica.

3.1 Chapter 3 Introduction

Geocoding is the process of converting postal address data into geographic coordinates, i.e., latitude and longitude pairs (Boscoe 2008). Given the ubiquity of postal address data, often cited as comprising 80% of all government data (Federal Geographic Data Committee 2006), this process can be seen as critical to nearly every academic, industrial, and government field that seeks to perform any type of spatial analysis or mapping. At the highest level, a geocoding system consists of six main components (Boscoe 2008): (1) the input data to be geocoded; (2) the address parsing and normalization algorithms that identify and standardize the pieces of the input data; (3) the geographic reference data representing real-world geographic features from which an output geocode is interpolated or returned directly; (4) the feature matching algorithms that link the input data to one or more reference features; (5) the feature interpolation algorithms that identify where along or within a reference feature the output should be located; and (6) the output spatial data.

It is well known that the resulting quality of geocodes produced by any geocoding system is closely tied to the quality of the underlying reference data sources available to the geocoding platform. Many researchers have shown that when highly complete and accurate reference data sources are available, there is a high likelihood that a geocoder will return a highly accurate result given a valid input query (Wu et al. 2005; Lee 2009; Frizzelle et al. 2009). Although, highly accurate reference data sources exist at the local level where municipalities and counties have created them, such as a county parcel file derived through field surveys by a local Assessor's Office, their availability and cost varies tremendously nationwide (Frizzelle et al. 2009). The cost of similar, highly accurate, nationwide, commercially derived reference sets is often prohibitive for all but the largest organizations and institutions.

Therefore, the most common form of reference data used in geocoding systems remain street reference files such as the U.S. Census Bureau TIGER/Line (U.S. Census Bureau 2009b) and ESRI StreetMap North America files (Environmental Systems Research Institute 2009d). Although these files are quite complete in terms of their geographic coverage, covering the whole of the US, they suffer from numerous accuracy and completeness errors in terms of both their spatial (geometry) and accompanying non-spatial attributes (address ranges, city names, etc.) (Frizzelle et al. 2009). Most commonly, errors in the address ranges associated with street segments in these files prevent successful geocoding attempts because the address number component of the input query address falls outside of the valid street address ranges in the reference dataset (Wu et al. 2005). In these cases, the commercially available state-of-the-art geocoding tools take one of two paths: (1) report a non-match, or (2) report a match and return the street centroid of a randomly chosen nearby street segment. Neither of these approaches is ideal because in the first, potential partial matches that could have been very close to the true output are thrown away. In the second all but the savviest of users typically unknowingly introduce false positive street-level accuracy matches into their spatial data by automatically accepting the output at face value without inspecting the reference feature from which it is derived to discover that it is an incorrect street segment.

The primary goal of the present research is to improve the match rates within geocoded datasets produced using address range geocoding through the use of nearby candidate reference features. However, it is not sufficient to merely increase the number of successfully matched cases (recall) without ensuring that the results achieved are spatially correct (precision). Therefore, the specific questions we seek to answer herein are two-fold: (1) can nearby reference features be used to improve match rates in geocoders that use street segment reference data sources?; and (2) how spatially accurate are the results of using nearby reference features when compared to those created by commercial geocoding systems with access to better reference data sources?

To investigate these questions, we introduce a candidate match scoring algorithm that computes match score values for nearby non-exact, matching approximate candidate reference features, termed *dynamic nearby reference feature scoring*. These scores are computed based on a spatially varying neighborhood representing the number of blocks a candidate reference feature is away from the input address number. Using these scores, we are able to identify, score, rank, and return the most probable and/or closest candidate reference feature to which the input address feature belongs or is spatially near to.

Our approach has three primary benefits. First, it is a natural extension to the composite weight scoring algorithms utilized in commercial geocoding platforms that currently return no matches (Boscoe 2008). That is, our approach inserts an additional scoring component to the current set of scores used to compute a composite match score, which, if desired, could be set to weight zero to replicate the behavior of current systems. Second, our approach is capable of producing the same output as systems that currently return non-exact matches, but has the distinct advantage that it additionally indicates that the result is a non-exact match. This allows users to easily identify non-exact matches and determine for themselves if the candidate should be included in their study, while at the same time providing a formal parameterized foundation for the current ad-hoc methods used in practice. Finally, our approach could be applied wholesale within any other geocoding engine that returns a list of candidate reference features to compute match scores in a similar manner.

To evaluate the effectiveness of our approach, we compare our approach against four widely used commercially available geocoding platforms. The first, the ESRI Address Locator (Environmental Systems Research Institute 2009c) represents the state-of-the-art in commercial desktop geocoding software and is widely recognized as the most commonly used for large-scale geocoding attempts throughout academia and industry (e.g., Rull et al. 2009; Omer et al. 2008; Blondin et al. 2007; Wagner et al. 2009; Macintyre et al. 2007; Ries et al. 2009; Wrobel et al. 2008; Ruiz et al. 2007; Warden

2008; Meliker et al. 2007; Apparicio et al. 2008). This software package is an example of a geocoding engine that will fail to find an acceptable match if the input address number is not within the address ranges associated with any features in the reference data source. The remaining three are the online geocoding systems provided by Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) which are routinely used by the general public and researchers alike to determine locations, driving directions, and for other mapping and visualization tasks. These systems represent geocoding platforms with access to (what are assumed to be) highly accurate reference datasets that should be less prone (if not completely immune) from the address range attribute errors that can confound address range geocoders relying on less accurate reference data sources such as the U.S. Census Bureau TIGER/Line files.

To perform our evaluation, all four reference geocoding platforms in addition to the USC geocoder are used to process the Los Angeles County subset (n=22,948) of the National Provider Identification (NPI) file, a nationwide set of 3.9 million addresses of Medicare payment recipients representing hospitals, clinics, and doctors offices (North American Association of Central Cancer Registries 2009). This dataset is commonly used in health-related research, so our evaluation herein serves to offer a heretofore missing examination of the results of geocoding these data. Using these systems, we evaluate the improvement in match rates that can be achieved using our approach and show that these improvements are spatially accurate.

The remainder of this chapter is organized as follows. In Section 3.2 we discuss related work. In section 3.3 we detail the implementation aspects of our prototype geocoding engine, excluding the match scoring portion which is presented in greater detail in Section 3.4. In Section 3.5 we evaluate the performance of our approach on the Los Angeles County subset of NPI records. We end with conclusions and future directions in Section 3.6.

3.2 Chapter 3 Related Work

The bulk of the prior work on feature matching and feature match scoring has focused on probabilistic feature matching and the match-unmatch feature matching scoring algorithm (Boscoe 2008). The majority of the geocoding systems in use today continue to make use of the approach developed by the U.S. Census Bureau in the 1970s (O'Reagan et al. 1987; Jaro 1984). The work presented herein takes a similar approach to the relative weight scoring algorithms used in the match-unmatch portion of probabilistic systems, with the distinction that our match scoring algorithm is deterministic, with the weights for each attribute being empirically based. Recently, neural networks have been applied to the similar problem of address matching and/cleaning and have proven to perform quite well when the complete domain of possible addresses is known in advance (Christen et al. 2005; Churches et al. 2002). Our approach does not have the luxury of a complete training corpus being available *a priori*, so it was determined that this would not be a suitable approach given the goals of this chapter. USPS TIGER/ZIP5+4 files (U.S. Postal Service 2009c) would not be suitable for this tasks as they provide you ranges of addresses that are valid, not specific addresses.

The distance decay-based approach we use for determining attribute weight values and corresponding score penalties for the various components of an input postal address has a long history of use in geospatial science (Eldridge et al. 1991) and other related fields that utilize the fundamentals of geography including such diverse disciplines as travel and tourism (Mckercher et al. 2003), environmental planning (Hanley et al. 2003), and biogeography (Poulin 2003). However, instead of modeling the decay rates of a specific geographic phenomenon under investigation, our approach uses distance decay as a measure of uncertainty present as one moves outward from the address in question in terms of attribute value similarity between the input address and the reference features being scored.

3.3 The USC Geocoder

The USC geocoder was built by the University of Southern California GIS Research Laboratory as a research platform for implementing and testing novel geocoding techniques and data sources (Goldberg et al. 2010). This platform implements all standard components found within typical geocoding system architectures, including the representation and storage of reference data layers, a feature matching algorithm including a deterministic candidate scoring scheme, and a set of feature interpolation algorithms. The implementation used for this research (version 2.94) is available online at https://webgis.usc.edu. The following subsections describe the specific implementation details of the various components of the platform, excluding the match scoring algorithm which is covered separately in Section 3.4.

3.3.1 Reference data sources

Geocoder reference data sources are typically classified into three categories commonly referred to as Dime, Nickel, and Penny. These categories are based on the types of reference features they contain and represent two-sided streets, one-sided streets, and situs (single) features, respectively. These types are named for historical reasons dating back to the early geocoder implementations by the U.S. Census Bureau in the 1970s (O'Reagan et al. 1987; Tobler 1972) that relied on the Dual Independent Map Encoding (DIME) format for storing two-sided street segments. The label Nickel is used because in monetary terms it is half a dime, and as a geospatial format it represents a single-sided street segment, half of that represented in the Dime format. The name Penny is used because in monetary terms it represents one, and as a geographic data format it represents a single situs, or location. The USC geocoder supports Dime, Nickel, and Penny reference data sources. Internally, Dime reference data layers are converted into Nickel versions by storing each of the two sides of the Dime reference features as independent reference features.

The data sources currently utilized in the publicly available version of the USC geocoder include the freely available 2008 versions of the U.S. Census Bureau TIGER/Lines (U.S. Census Bureau 2009b), Places, ZCTA5, ZCTA3, County Subregion, and County files (U.S. Census Bureau 2009a), the 2009 version of the Los Angeles County Assessor's Parcel files (Los Angeles County Assessor's Office 2009), and the 2009 version of the USPS TIGER/ZIP5+4 files (U.S. Postal Service 2009c). The parcel and Census files were obtained in ESRI shapefile format, while the TIGER/ZIP5+4 files were converted to shapefiles from the original flat text files. Each of these reference data layers are stored as spatial data within the Microsoft SQL Server 2008 relational database management system (RDBMS). The shapefiles were read and converted into SQLGeography spatial data types using a set of custom written conversion tools based on the Reimers shapefile reading library (Reimers 2009). Each of these data sources is stored within separate databases in the RDBMS, organized by state to improve efficiency of data access. The reference data files that contain full postal addresses (TIGER/Lines and parcel data) are indexed using a composite covering index across the Soundex value of the street name, the Soundex value of the city name, and the USPS ZIP code. The reference data files that contain only city portions of the postal addresses (Places, Counties, etc.) are indexed using a covering index across the Soundex value of the city name, while the USPS ZIP code and USPS ZIP+4 reference data are indexed using covering indexes on the USPS ZIP code and USPS ZIP+4 fields, respectively. The experiments performed in this research use the ESRI Street Map North America street reference data files (Environmental Systems Research Institute 2009d), which are not available in the online version of the USC geocoder and are covered in more detail later.

3.3.2 Address Normalization

The address parsing and normalization component implemented in the USC geocoder is a non-USPS CASS certified (U.S. Postal Service 2009a) deterministic token-based context-aware state machine. Parsing and normalization are applied to the street address portion of an address including the secondary unit and can recognize P.O. boxes and other delivery route address types, e.g., Rural Routes. Input addresses are standardized to the specific address format used by each reference data file such as the USPS Publication 28 specification (U.S. Postal Service 2009b) or the Federal Geographic Data Committee (FGDC)/Urban and Regional Information Systems Association (URISA) Draft Street Address Standard (Urban and Regional Information Systems Association 2009).

In our approach, tokens are identified by first scrubbing non-alphanumeric characters and then splitting the input string based on white space separation. Proceeding from left to right, each token of the input street address is identified as belonging to one or more of a set of lexical types derived from a grammar composed of the components of an FGDC/URISA address (which is a superset of those found in USPS Publication 28). To accomplish this, our approach uses a sliding window of size three around each token to represent the context of tokens that precede and follow the token in question. In this window, a set of state transition rules representing the boundaries between address components is utilized to determine the ultimate final typing of each input token based on what has proceeded the present token and what follows it. Once the lexical type of the token is identified, the normalized value from the specific address standard used in the reference dataset is substituted in the parsed and normalized output value.

3.3.3 Feature Matching Algorithm

The feature matching component is implemented using a loose-matching strategy that returns the maximum number of candidate reference features possible. In our approach, only the street name, city, and first three digits of the USPS ZIP code attributes are used as blocking attributes in the generated SQL query that searches the reference data source for candidate matches. In addition, the street name and city attributes are queried using their Soundex equivalent values to overcome minor spelling mistakes and allow phonetically similar matches.

Depending on the values of the input address, our system generates a set of SQL queries to interrogate a reference source for candidate features. The simplest of these queries is one which is intended for use on a reference data file that only stores one administrative area (e.g., a city) to process an input address that is not a number or numeric abbreviation (Figure 3.1). Reference datasets of this type are most commonly parcel geometry or address point files maintained by local governments that have no use for other administrative delineation notations such as county, county sub-regions, etc.

```
SELECT * FROM [state]
WHERE
Name_Sdx=@p1
AND
(
        ZIP3=@p2 OR
        City_Sdx=@p3
)
```

Figure 3.1: Feature matching algorithm-generated SQL for a city-only reference sources

In the case of multiple administrative areas, such as in the U.S. Census Bureau TIGER/Line files, we expand the query to include the additional administrative areas in order to return results that include them in the case that the user input or reference dataset contain an alias value (Figure 3.2). In the case of USPS ZIP code files, all but the USPS ZIP code attribute would be dropped.

```
SELECT * FROM [state]
WHERE
Name_Sdx=@p1
AND
(
        ZIP3=@p2 OR
        City_Sdx=@p3 OR
        ConCity_Sdx=@p4 OR
        Mcd_Sdx=@p5 OR
        CountySub_Sdx=@p6 OR
        County_Sdx=@p7
)
```

Figure 3.2: Feature algorithm-generated SQL for reference source with multiple administrative areas

Our approach intentionally returns a large set of candidate features with street and city names (and possibly other administrative alias values) that are phonetically equivalent to the input street name or have a USPS ZIP code that begins with the same three digits. Each member of this set of candidates, which undoubtedly includes a large number of false positives, is scored using the feature match scoring approach detailed in Section 3.4. The candidate reference feature with the highest score is used for feature interpolation as detailed in the next section.

This approach has two main benefits. First, it reduces the required index storage space to efficiently answer queries. This is important because the geographic reference data files used for geocoding are quite large e.g., ~143 GB for the U.S. Census Bureau TIGER/Line files. Specifically, when nationwide coverage is required within a geocoding platform, the startup costs can quickly become prohibitive when enterprise quality disk space, such as a storage area network (SAN), is required.

Second, it enables rapid query execution within the RDBMS because the number of query filters requiring evaluation is kept to a minimum which is important when considering the huge number of reference features that must be searched, e.g., ~3 million street segments in U.S. Census Bureau TIGER/Line files for the state of California and ~2.7 million parcels for LA County. Essentially, our approach chooses to perform only the most rudimentary blocking in the RDBMS, allowing us to develop and evaluate a host of match scoring algorithms that operate on the widest possible superset of candidate features in memory.

3.3.4 Feature Interpolation Algorithm

The feature interpolation component of the USC geocoder supports address range, uniform lot, and actual lot linear interpolation (Bakshi et al. 2004) for street segments using either a user-configurable dropback value or defaulting to a static 10 m dropback in the case the user did not supply a value. Areal interpolation is performed on the native SQLGeography spatial data types using their built-in OGC STCentroid() implementation for all areal unit reference sources, e.g., parcels, cities.

3.4 Feature Match Scoring

The feature match scoring algorithm implemented in the USC geocoder is a deterministic process that uses a per-attribute penalty scheme to compute an overall match score for each candidate reference feature. In this process, each attribute is assigned a relative penalty weight by the user which is proportionally applied based on the level of incorrectness between the reference feature address attribute and the input data address attribute. The penalties for each applicable address attribute are calculated and summarized for the address as a whole to compute an overall non-match penalty to associate with the reference feature.

This method of calculating a match score is similar to the processes utilized in other state-of-the-art geocoding platforms that calculate and return a match probability score such as the ESRI ArcGIS Address Locator (Environmental Systems Research Institute 2009c). While similar to the general approach of building a composite overall score based on the penalties associated with individual address attributes, our method differs from that taken by the ESRI ArcGIS Address Locator in two important ways. First, the ESRI attribute weights are probabilistically-based where ours are empirically-based. Second, our approach employs a novel method to match and compute a weighted penalty score for "nearby" reference features in the case that an exact match is not found. This occurs when an input address number is outside of the address ranges of all street segments in the reference data source, typically indicating that the address ranges or the reference features in the reference dataset are missing information. In these cases, the ESRI ArcGIS Address Locator (Environmental Systems Research Institute 2009c), will return an unmatchable result, because as the documentation states "The single number must be within the interval (including the endpoints) to be considered a valid match. Otherwise it is a disagreement " (Environmental Systems Research Institute 2009a, p. 81). In contrast, our approach will return a candidate reference feature that is nearby and within a user-configurable buffer zone. This method can be applied as a general scoring algorithm within any geocoding engine that returns the attributes of the reference feature matched in the case of geocoders that do not support nearby matching e.g., the popular ESRI Address Locator (Environmental Systems Research Institute 2009c) and the Manifold Geocoding Database (Manifold Net Ltd. 2009) desktop engines, or preferably in systems that return multiple matches and require the user to pick the best one, i.e., the Google, Microsoft Bing, and Yahoo! Geocoding systems (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b).

It is a well known fact that geographic identifiers in the form of both toponym and code references are dynamic, changing both their spatial footprints and the names by which they are known over time (Beyer et al. 2008; Krieger, Waterman et al. 2002; Goldberg, Wilson, and Knoblock 2008). For example, over time new USPS ZIP codes are created, old USPS ZIP codes are split or merged, and their boundaries change to facilitate more efficient mail delivery. Likewise, the city name component of the input address can change as development occurs or as small communities succeed from or merge with larger communities. These inconsistencies can also be found in the other components of a postal address such as the address number, directionals, etc. These discrepancies are particularly common along the borders between administrative areas, although they are also found between different neighborhoods with different address numbering schemes within the same administrative area (Fonda-Bonardi 1994). For instance, it is common to encounter inconsistent address ranges between contiguous street segments of the same street as one travels from neighborhood to neighborhood or crosses administrative boundaries, a phenomenon that often troubles automobile drivers unfamiliar with a new area.

In addition to this spatial and temporal dynamism across all components of postal addresses, major discrepancies often occur between the geographic scale of knowledge by which these objects are referred to in practice versus how they are recorded in official listings and databases (Goldberg, Wilson, and Knoblock 2008). A user may submit an address with an unofficial and/or non-recognized community and/or neighborhood name in place of the official version known to the originator of the reference datasets available to the geocoding system. For example, the address "10121 Tabor St, Palms, Ca 90034" indicates the city as "Palms" which is a well-known neighborhood in the City of Los Angeles. Although "Palms" would be immediately recognizable and locatable to most people who live in the West Los Angeles region as the area south of the 10 Freeway expanding outward from the 405 Freeway toward Culver City, it would most likely not be on record as the city attribute within a large national-scale reference dataset of street segments. Instead, the official city of record, "Los Angeles", would be associated with the street segment. Similarly, the city name component is often entered using locally known acronyms or abbreviations such as "NYC, NY" and "NY, NY" for the full official name "New York City, NY" or "S. Hampton, PA" for "Southampton, PA".

The implication of these trends is that a geocoding platform must be flexible in its scoring approach to the components of a postal address, taking special care to remedy these mismatches when appropriate, while at the same time recognizing when such a strategy would be foolhardy. In the following subsection we describe the details of our match scoring approach. The level of detail presented is provided for two reasons. First, by describing the complete technical approach taken, other researchers and engineers will have the ability to employ our approach to achieve similar results. Second, our detailed description provides transparency to the end user as to how our results are achieved. Because the match scoring strategy used by a geocoding engine is ultimately tasked with determining the most correct output from a set of possible candidates, it is central to the accuracy and reliability of the output. In order to have confidence in the data produced by a geocoder, the end users need to know how the scores for a particular geocode were calculated, why one candidate was chosen as the output, what the alternatives were, and why they were not chosen. Commercially available geocoding systems do not often reveal these details, so the user is left to assume that the system operated correctly and chose the best possible output. Our work presented herein seeks to remove this veil of assumption forced on the user, instead allowing them to judge for themselves if the output data from the USC geocoder is of sufficient quality to meet the needs of a particular application or study. For convenience, the variables used throughout this section are listed along with their descriptions in Table 3.1.

3.4.1 Relative Attribute Weighting

The first step in our match scoring process is to compute a relative penalty weight to be assigned to each of the address components. In this scheme, each of the $i = 1 \ldots n$ address components a_i (e.g., street name, suffix, city, USPS ZIP code) are assigned a weight w_i by the user on a scale of 0 – 100.

Table 3.1: Descriptions of the variables used for calculating a match score

Variable	Description
a, a_i	An input address and a single address component, respectively
f, f_i	A reference feature candidate and single address component, respectively
$a_r, a_p, a_n, a_q, a_t, a_c, a_z, a_s$	The street number, pre-direction, name, post-directional, suffix, city, USPS ZIP code, and state address components of an input address, respectively
$f_r, f_p, f_n, f_q, f_t, f_c, f_z, f_s$	The street number, pre-direction, name, post-directional, suffix, city, USPS ZIP code, and state address components of a reference feature, respectively
p_i	The penalty weight associated with a single address component
$w_i, \dot{w}_i$	The full and proportional penalty weight associated with a non-match for a single address component a_i, respectively
w_c, w_s, w_z	The full penalty weight associated with the city, state, and USPS ZIP code address components, respectively
r_i	The relative weight of a single address component a_i as computed in relation to all other address component weights assigned by the user
$S(f, a)$	The match score function that computes a match score between a reference feature f and an input address a
$Max\big(S(f, a)\big)$	The maximum match score that can be computed between a particular reference feature f and an input address a
S_{min}	The user-defined minimum match score that must be attained for a reference feature to be considered a viable candidate
R_M, R_m, R_l	The major, minor, and local geographic regions of the reference feature or input address, respectively
R_{Ma}, R_{Mf}	The major geographic region of the input address and reference, respectively
R_{ma}, R_{mf}	The minor geographic region of the input address and reference, respectively
R_{la}, R_{lf}	The local geographic region of the input address and reference, respectively
$Edit(a_i, f_i)$	The edit distance function that computes the edit distance between an input address component a_i and reference feature address component f_i
L_1, L_2, L_3	The city address component geographic match levels
M_1, M_2, M_3	The match scores for the city address component at each of the geographic match levels
$[M], \overline{M}$	The set of match scores for the city address component at each of the geographic match levels and the highest of this set, respectively
P_s, P_e	The starting and ending nodes of a street segment, respectively
N_s, N_e	The numbers associated with the start and end of a street segment address range, respectively
d	The smaller of the numeric distances between the input address number and the starting and ending address numbers of a reference street segment
B_d, B_M	The number of blocks an input address number is away from a reference feature and the user-defined maximum number of blocks, respectively
$\lvert b \rvert, \lvert b_f \rvert, \lvert \bar{b} \rvert$	The number of addresses on any block, the number of addresses on a particular reference feature f, and the average number of addresses on a block within a geographic region, respectively

This weight w_i indicates the penalty that should be applied in the case of a non-match between the input address datum and the candidate reference feature. Each per-component weight, w_i, is normalized into a relative overall weight, r_i, by dividing the individual weight, w_i, by the total weight assigned across all address components by the user (Equation 3.1). These relative weights indicate the relative importance the user has assigned to each of the address components. If all components are assigned the same weight by the user, each component will receive an equal relative weighting.

$$r_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{3.1}$$

3.4.2 Relative weight scoring strategy

There are many factors to consider when deciding on the best method to compute an overall score $S(f, a)$ for a candidate feature f against input address a, out of the total maximum allowable score, T, which we will always set to be equal to 100. Most notably, the minimum candidate score S_{min} defines the confidence cutoff beyond which candidate features are rejected from consideration. This parameter affects the overall match rate of the geocoder. Setting this parameter too low will produce a high number of false-positive matches (reducing the precision), while setting it too high will produce a high number of false-negatives (reducing the recall).

In addition to the minimum candidate score, S_{min}, different strategies can be employed to assign the relative weights to each of the address components used in the feature match scoring process. The strategy taken in our system uses a hierarchical approach to define the minimum geographic relationships that must exist for the match score calculated for a particular reference feature input address combination $S(f, a)$ to be scored above the minimum match score S_{min}. These relationships are defined in terms of the major geographic region R_M, the minor geographic region R_m, and the local region R_l. The major geographic region is the single top level geographic indicator, the state attribute, e.g. California. The minor geographic region is the set of multiple geographic region indicators, the city and USPS ZIP code attributes, e.g. Los Angeles and 90089, respectively. The local geographic region is the set of attributes of the street address, e.g., the street number, pre-directional, name, suffix, etc. The relationship between these regions is a hierarchical one as shown in Figure 3.3.

The triangle shape of this hierarchy signifies the proportional amount of error allowed within each of the levels in order for a candidate feature to be scored above the minimum candidate score S_{min}. In particular, the candidate feature must agree with the input address on the major region R_M, while there is some room for error in the minor region R_m, i.e. the city or USPS ZIP can be wrong, and even more room still in the local region R_l. where, for example, the street suffix and pre-directional could be wrong.

The geographic relationships implied between different levels are used to enforce a set of constraints that guide the selection of the weights for each individual address attribute. In particular, if the major region of the input address R_{Ma} and that of the reference features R_{Mf} do not agree, the maximum match score calculated $Max\big(S(f,a)\big)$ must be below the minimum match score S_{min} (Equation 3.2a).

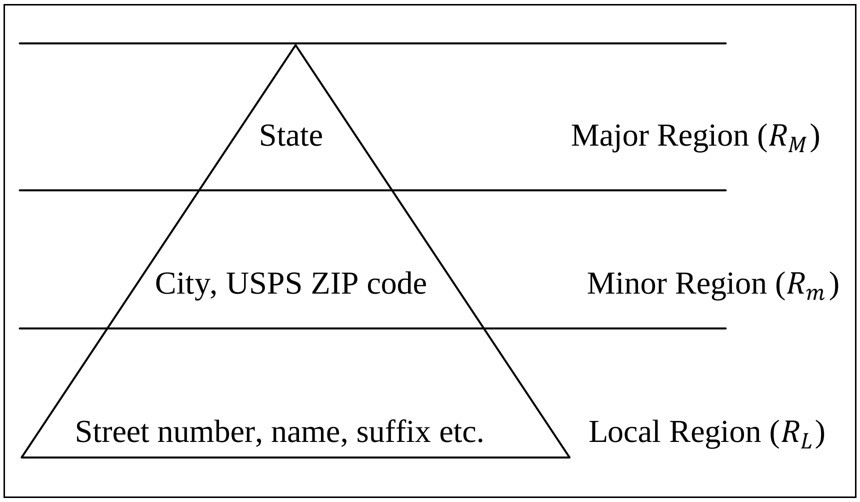


Figure 3.3: Hierarchical geographic region relationships

If R_M agree but none of the set of minor regions attributes between the input address, R_{ma}, and reference feature, R_{mf}, agree, $Max\big(S(f,a)\big)$ must also be below S_{min} (Equation 3.2b). If R_M agree and at least one of the set of minor regions attributes of the input address, R_{ma}, and reference feature, R_{mf}, sets agrees, $Max\big(S(f,a)\big)$ can then be set above S_{min} (Equation 3.2c).

These constraints are used to assign penalty weight scores w_i for each of the $i = 1 \dots n$ address attribute a_i. In particular, these constraints bound the minimum weights to be assigned the state, city, and USPS ZIP code attributes, w_s, w_c, and w_z respectively in terms of S_{min} and the total possible score $T = 100$. Together, the set of constraints from Equation 3.2 and minimum bounds from Equation 3.3 force the input address and reference feature to agree on the state attribute as well as either the city or USPS ZIP code attributes in order for a match score to be calculated as above the minimum match score.

$$Max\left(S(f,a)\right) = \begin{cases} < S_{min}\ ,\ R_{Ma} \neq R_{Mf} & \text{(3.2a)} \\ < S_{min}\ ,\left(R_{Ma} = R_{Mf}\right) \wedge \left(R_{ma} \cap R_{mf} = \emptyset\right) & \text{(3.2b)} \\ \geq S_{min}\ ,\left(R_{Ma} = R_{Mf}\right) \wedge \left(R_{ma} \cap R_{mf} \neq \emptyset\right) & \text{(3.2c)} \end{cases}$$

Equation 3.3a shows that the weight of the state w_s must be greater than the difference between the total possible score ($T = 100$) and S_{min}. Likewise, Equation 3.3b shows that the sum of the weights of the city and USPS ZIP code must be above the difference between the total possible score ($T = 100$) and S_{min}. This means that candidate features that are not in the correct state and city or USPS ZIP will always be scored below the minimum match score and eliminated from consideration as a valid output.

$$w_s \geq T - S_{min} \tag{3.3a}$$

$$w_z + w_c \geq T - S_{min} \tag{3.3b}$$

As an example, if the user chooses the minimum candidate score $S_{min} = 71$ then we know that the minimum penalty for the state attribute is $w_s \geq 30$ and the minimum penalties for the city and USPS ZIP code attributes together are also $w_z + w_c \geq 30$. The actual values for $w_z + w_c$ may vary individually, but as a group they must be greater than or equal to 30.

3.4.3 Per-component non-match penalty score computation

Each of the $i = 1 \dots n$ address components of the input address, a_i, and reference feature address, f_i are compared to compute a per-component penalty score, p_i. Different strategies are used to compute these penalties based on the type of address component under consideration. A brief description of the approach taken for each address component is listed in Table 3.2. A detailed derivation for each address component is outlined in the following sections.

Table 3.2: Method and description of the penalty scoring approach taken for each address component

Address component (a_i)	Method	Description
Pre-directional, Post-directional, Suffix	Character equivalence	If input matches reference return zero penalty, otherwise return full penalty
USPS ZIP code	Similarity weighted	If input matches reference return zero penalty, otherwise return a proportion of full penalty based on the index of the first difference character between the two
City	Similarity weighted	If input matches reference return zero penalty, otherwise return a proportion of the full penalty based on the edit distance between the closest textually matching region identifier
Address number	Similarity weighted	If input is within reference return zero penalty, otherwise return a proportion of full penalty based on the distance to the input number

3.4.3.1 Pre- and post-directionals and street suffixes

The non-match penalty score for the pre- and post-directional address attributes as well as the street suffix are all computed in the same manner which proceeds as follows. The directional and suffix values of the input feature, a_i, and reference feature, f_i, are first normalized into their USPS Publication 28 standard values (U.S. Postal Service 2009b). This results in a single character such as ‘N’ for ‘North’ and a set of characters such as ‘BLVD’ for ‘Boulevard’ for the directional and suffix attributes, respectively. These normalized values are then compared using a case-insensitive equivalence comparator to determine if they are the same. In the cases of equivalence or both the input and reference being empty, a penalty of zero is reported for the attribute in question (Equation 3.4a).

In the case of non-equivalence some penalty weight, p_i, must be calculated and reported because the user-entered information differs from that maintained by the reference feature. There are two scenarios where this can occur: (1) either the input or reference attribute is empty while the other is not; or (2) the attributes are both non-empty and are actually different values. The first of these again results from two possible scenarios: (1) either the user included the attribute and it is not present on the reference data feature; or (2) the user omitted the attribute from their input and the attribute is present on the reference data feature.

The former means that the user provided more information than what is known about the reference feature – the input datum is over-specified. This situation could occur if the attribute set describing the reference feature is not complete and the user has the correct attribute value, if the attribute set describing the reference feature is not complete and the user has the incorrect attribute value, or if the attribute set describing the reference feature is complete and the user has the incorrect attribute value. Which of these three instances has occurred is not knowable using a single reference data source because each reference feature represents all knowledge known about the world. Essentially, the value the user entered cannot be proven to be completely incorrect (which would result in the full penalty weight being returned), nor can it be proven to be completely correct (which would result in a penalty of zero). Thus, the penalty is returned as one third of the full penalty weight to give the benefit of the doubt to the user instead of the reference feature because it is more likely that the user has complete and up-to-date information about the real-world attributes of the feature than the reference dataset (Equation 3.4b).

In contrast, if the user omitted an attribute that is present in the reference data, it means that the user provided less detailed information than what is known about the reference street address – the input datum is under-specified and this is an error of omission on the part of the user. This situation has the potential to result in ambiguous feature matches if all that differs between two or more potential candidate features is the attribute value which was not supplied by the user. Therefore, two thirds of the penalty weight is returned to give the benefit of the doubt to the reference feature attribute value (Equation 3.4c).

In the case where neither the input nor reference feature address attribute is empty, instead the attribute values simply disagree, the full penalty weight is returned (Equation 3.4d). The final output penalty for the attribute, p_i, is then calculated as the proportion of the weight to apply, $\dot{w}_i$, multiplied by the full weight assigned to the attribute, w_i (Equation 3.4e).

$$\dot{w}_i = \begin{cases} 0, \quad a_i = f_i & (3.4a) \\ \frac{1}{3}, (a_i \neq f_i) \wedge (a_i = \emptyset) \wedge (f_i \neq \emptyset) & (3.4b) \\ \frac{2}{3}, (a_i \neq f_i) \wedge (a_i \neq \emptyset) \wedge (f_i = \emptyset) & (3.4c) \\ 1, (a_i \neq f_i) \wedge (a_i \neq \emptyset) \wedge (f_i \neq \emptyset) & (3.4d) \end{cases}$$

$$p_i = w_i * \dot{w}_i \qquad (3.4e)$$

3.4.3.2 USPS ZIP Code

As noted, USPS ZIP codes are temporally and spatially dynamic (Beyer et al. 2008; Krieger, Waterman et al. 2002) and thus must be treated in a special manner when computing a penalty score. A desirable characteristic of a penalty score computed for a USPS ZIP code component is that it should represent a degree of similarity between the input and reference feature values in terms of spatial proximity. Nearby USPS ZIP codes should be returned with less of a penalty than those that are further away. However, it is not a simple matter to determine the spatial proximity of USPS ZIP codes based on their numeric code values alone. Although neighboring USPS ZIP codes typically run in sequence, there are indeed instances where the values of neighboring USPS ZIP codes are far removed from each other, such as 90089 (USC) being surrounded by USPS ZIP codes 90007, 90011, and 90012.

Therefore, in our implementation, the penalty score for non-matching USPS ZIP code components of the address is proportionally weighted using a linearly decreasing function based on the position of the first non-matching character. The penalty applied is inversely proportional to the index of the first non-matching character (Equation 3.5a). Given an input and reference feature USPS ZIP codes a_z and f_z of length $l = 5$, the proportion of the output penalty to return p_i is defined as the index of the first non-matching character c_i divided by the length of the USPS ZIP code character string, l (Equation 3.5b). This strategy results in a decay function that returns the full penalty for USPS ZIP codes that disagree on the first character, and reduces by one-fifth (i.e. 0.2) for each consecutive matching character.

$$p_i \propto \frac{1}{c_i} \tag{3.5a}$$

$$p_i = \frac{l - c_i}{l} \tag{3.5b}$$

This decay function was chosen because it works well for the Los Angeles region, but another function could be more apt in different areas with other spatial and non-spatial patterns of USPS ZIP codes and numeric values. Therefore, our implementation allows other decay functions to be created and incorporated in place of our linear decay function, enabling other scientists and researchers using our geocoding platform to experiment with other strategies to make optimal decisions given their data and geographic region.

3.4.3.3 City name

To accommodate the difficulties inherent in the mismatch between the scale of geographic knowledge within an input address versus that stored in the reference data files, commonly used national-scale reference data sources such as the U.S. Census Bureau TIGER/Line files (U.S. Census Bureau 2009b) often include several administrative delineations along with a street segment in addition to the official place name within which it resides such as the minor civil division or the county sub-region. Therefore, our implementation computes scores for each of the multiple administrative areas associated with the reference feature. Using the U.S. Census Bureau TIGER/Line files as an example, the five possible city name candidates are ranked into the three-level hierarchy $L_1 - L_3$ shown in Figure 3.4.

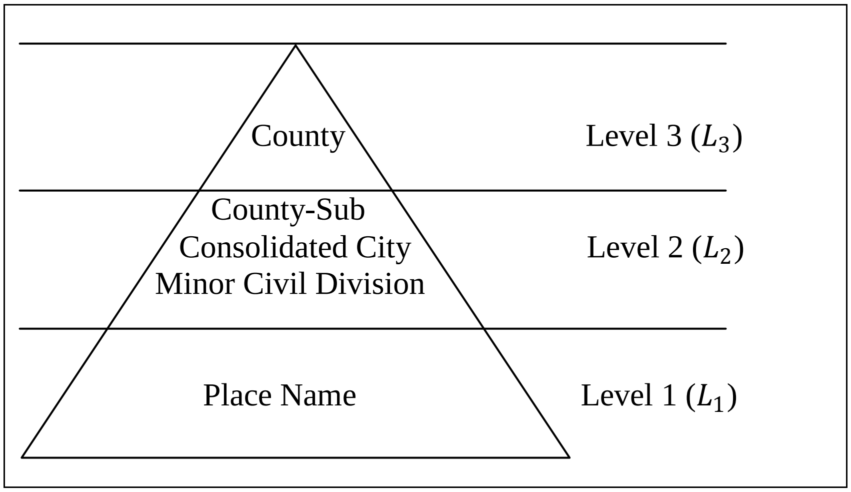


Figure 3.4: City name match-scoring hierarchy levels

The level at which a match is found in this hierarchy is used to proportionally weight the match penalty depending on the agreement or disagreement between a more accurate level. This approach is similar to the geographic hierarchy levels used in the significance- and context-based assignment approach to identifying toponyms proposed by Li (Li 2007).

The first step in our approach is to attempt a case-insensitive string comparison against the L_1 reference feature value (place name), and the value of the input address city. If the two match exactly, a penalty of zero is returned. If they disagree, both are normalized to the full expansion of common abbreviations for directionals, states, and other common substitutions, e.g., “S” becoming “South” in “S. Hampton, PA”, “NY” becoming “New York City” in “NY, NY”, and “LA” becoming “Los Angeles” in “LA, CA”, respectively. The case-insensitive string comparison is then attempted again using these normalized versions of the input and reference feature city values. If an exact match is found, a penalty of zero is returned.

If an exact match is not found, single token strings and multi-token strings are handled differently. In both instances we use an implementation of the Levenshtein distance (Navarro 2001) to compute the edit distance between the input city, a_c, and the feature city, f_c. The edit distance of the city values, $Edit(a_c, f_c)$, is a measure of similarity between a_c, and f_c, in terms of the number of operations that must be performed to make the two strings equivalent (i.e. characters that must be added or removed from the two strings) (Navarro 2001).

This edit distance is used to derive $\dot{w}_i$, the proportion of the full penalty weight w_i to apply by dividing the edit distance by the length of the shorter of the two strings, the input or reference city value (Equation 3.6).

$$\dot{w}_i = \frac{Edit(a_c, f_c)}{Min(|a_c|, |f_c|)} \tag{3.6}$$

The single-token case where both the input and reference feature city each consist of just a single word is the simpler of the two. Here, the edit-distance based proportional penalty between each of the words, $\dot{w}_i$, is calculated and used directly. The case of multi-token non-exact matching city values in either the input or reference feature represents a more complicated case that must be dealt with differently. This can occur in one of three instances: (1) the user enters fewer words than the reference feature contains; (2) the user enters more words than the reference feature contains; or (3) the user enters the same number of words as the reference feature, but they do not match. Examples of these cases are listed in Table 3.3, which shows that in some of these instances a match scoring algorithm should return a low penalty indicating a close match even though the number and/or content of the words were different i.e. the example cases marked types 1 and 2, while in others even though the words are similar, the result should be a non-match, i.e. cases marked type 3.

Table 3.3: City non-match types and examples

Type	Input Error	Input City	Reference City	Match
1	Missing words	Niagara, NY	Niagara Falls, NY	Yes
1	Missing words	New York, NY	New York City, NY	Yes
2	Extraneous words	Alhambra City, Ca	Alhambra, Ca	Yes
2	Extraneous words	South Los Angeles, Ca	Los Angeles, Ca	Yes
3	Wrong words	Philadelphia, PA	Pittsburgh, PA	No
3	Wrong words	Los Alamitos, Ca	Los Angeles, Ca	No

Our implementation takes a deterministic approach to address this problem. We use the smaller of the two sets of words in either the input or reference feature to search within the other and use the cumulative edit distance as in the single token case (Equation

3.6). Our approach essentially looks for the longest common substring between the two city values and returns a value representing the proportion of the substrings that match.

We feel the tradeoff in simplicity here is justified because the relative weight scoring strategy proposed in Section 3.4.2 requires either the city or USPS ZIP code to match in order for a match score to be returned above the minimum candidate score threshold. This approach utilizes the scores of the city name and USPS ZIP code in conjunction to ensure that the candidate street segment is within the correct minor geographic region. Again, our platform is extensible, so more sophisticated techniques could be created and incorporated such as those relying on probabilistically-based pointwise mutual information inherent in statistically evaluated bi-grams (Church et al. 1991), as recently applied to toponym resolution (Wang et al. 2006).

Using this approach, each of the $L_1 - L_3$ city attributes of the candidate reference feature in Figure 3.4 are scored against the city value associated with the input address, resulting in $[M]$, which is the set of match scores from each of the three levels, M_1, M_2, and M_3. These sets differ in size, $|M_1| = 1$, $|M_2| = 3$, and $|M_3| = 1$, because from Figure 3.4, M_2 contains three attributes, consolidated city, county sub-region, and minor civil division, while M_1 and M_3 each contain just one. Within M_2, we only need to consider the highest match score of the set which will be represented by $\mathrm{M}_2^{'}$ (Equation 3.7). The maximum match score of the set of three scores in $[M]$, is then defined as $\bar{M}$ (Equation 3.8). The set of conditions listed in Equation 3.9 are used to determine the final output penalty, based on the level of the highest match score as well as if there is disagreement with a lower level, which is a common approach taken to preserve and exploit the geographic relationships between regions (Li 2007).

$$M_2^{'} = Max(M_2) \tag{3.7}$$

$$\bar{M} = Max(M_1, Max(M_2^{'}, M_3)\ ) \tag{3.8}$$

If the highest match score is M_1, the place name, then M_1 is returned as the proportion of the total penalty to apply (Equation 3.9a). If the highest match score is M_2, the county sub-region, minor civil division, or consolidated city and $M_1 = \emptyset$ because the L_1 attribute was empty, M_2 is returned (Equation 3.9b). If the highest score is M_2, but M_1 is defined, return either $1/3$ the full penalty or M_2, whichever is greater (Equation 3.9c). In this case our algorithm returns a minimum of at least one-third of the full weight penalty because the user entered somewhat incorrect data, but they were still correct at a L_2 match indicating that they had the sub-region correct.

$$\dot{w}_i = \begin{cases} 1 - M_1, \bar{M} = M_1 & (3.9a) \\ 1 - M'_2, \bar{M} = M'_2\ \wedge\ (M_1 = \emptyset) & (3.9b) \\ Min\left(1 - M'_2, \frac{1}{3}\right), \bar{M} = M'_2\ \wedge\ (M_1 \neq \emptyset) & (3.9c) \\ 1 - M_3, \bar{M} = M_3\ \wedge\ (M_1 = \emptyset) \wedge\ (M^{'}{}_2 = \emptyset) & (3.9d) \\ Min\left(1 - M_3, \frac{2}{3}\right), \bar{M} = M_3\ \wedge\ \left((M_1 \neq \emptyset)\ \vee\ (M'_2 \neq \emptyset)\right) & (3.9e) \end{cases}$$

If the highest match score is M_3 (i.e., county) and $M_1 = \emptyset$ or $M_2 = \emptyset$ because both the L_1 and L_2 attributes are empty, then M_3 is returned (Equation 3.9d). If instead the best match is M_3 but either $M_1 = \emptyset$ or $M_2 = \emptyset$, then a minimum of two-thirds of the full penalty weight is returned because the user entered the correct county, but not the correct place name or other sub-region although they were defined for the candidate reference feature.

3.4.3.4 Address number and dynamic nearby feature scoring

In traditional geocoding systems, the address number component of an input address is not used in the match scoring process. Instead, it is either a mandatory requirement of the reference feature for candidate consideration, or it is ignored and not a part of the match scoring procedure. An example of the first case is the ESRI ArcGIS Address Locator (Environmental Systems Research Institute 2009c) where "The single number must be within the interval (including the endpoints) to be considered a valid match" (Environmental Systems Research Institute 2009a, p. 81), or the candidate reference feature will automatically be assigned a match score below the user-specified minimum match score. Examples of the second case include the commonly used online geocoding services provided by Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) that do not provide a quantitative match score at all, and certainly not one that is dependent on similarity of the address number component associated with the input address and the reference feature in question.

Neither of these behaviors is ideal. In the first case, potential candidate reference features that are nearby and could potentially be valid for use in deriving an output are thrown away, while in the second nearby candidates are returned, but there is no indication that the result returned is anything less than an exact match in the most common usage scenario, interactive map visualization. To address these shortcomings in current systems, we have designed an address number match scoring system termed *dynamic nearby candidate range scoring* that is capable of producing match scores for nearby candidate features. Thus, if the closest candidate feature with the highest acceptable match score is a nearby feature, the system will be able to return the nearby feature as well as indicate the status of the match.

The main intuition behind our approach is to weight the match score penalty for the address number component by a factor representing its closeness to the input address. That is, if the address number associated with a reference feature is close to that of the input address, apply less of a penalty to it than to one that is further away. The challenges here are twofold: (1) how can we express closeness, as there are an infinite number of potential addresses that are within *some* distance of the input address?; and (2) what scale should be used to quantify the measure of closeness?

To describe our approach, we will consider the case where an input address query is being tested against a reference dataset containing street segment reference features that have address ranges associated with them, i.e., traditional address range geocoding with street segments such as the U.S Census Bureau TIGER/Lines (U.S. Census Bureau 2009b). Suppose that the feature matching algorithm presented in section 3.3.3 is applied which selects a large candidate set of features because only the Soundex version of the street and city along with the first three digits of the USPS ZIP code are used as query filters. This set will contain many potential candidate street segments with the correct street name and city and/or USPS ZIP code, with these candidate segments varying by the address ranges associated with them. Each of these candidates will then be scored in terms of closeness to the input address as follows.

Because the street segments used in geocoding systems are linear geographic features, they therefore have a starting node P_s and an ending node P_e as well as zero or more interior nodes that describe the curvature of the street. In the majority of cases, the address ranges on these street segments describe a continuous range of addresses, although exceptions do exist where the from-address is equal to the to-address, i.e., the street segment is assigned a single number. In the normal case where the street has a range of addresses, the starting and ending nodes of the segment each have a different number N_s and N_e, respectively, one higher than the other. These numbers will most commonly have the same parity (even or odd) although this is again sometimes not the case in special circumstances where street ranges have been split or merged or there are jurisdictions that have legacy or non-standard parcel numbering systems. Each of the potential candidate features will have one end of the street segment that is closer in number to the number of the input address. The numeric distance between the input address number a_r and the closer of the two address range values associated with each street segment endpoint and is defined as the variable d (Equation 3.10).

$$d = Min\big(Abs(a_r - N_s), Abs(a_r - N_e)\big) \tag{3.10}$$

This variable provides a dimensionless metric by which one potential street segment candidate can be compared against another to determine which of the two is closer to the input address. However, alone, it is not possible to use this distance metric d to compute a meaningful proportional penalty score that can be applied to the street number attribute. This is because we need to know the sum for all potential d to determine the appropriate penalty proportion for any particular d that would describe the closeness of that particular address range in comparison to all other address ranges. Because there is an infinite number of address ranges and an infinite number of actual address divisions within an address range, it is not possible to scale the resulting penalty score for a particular address range as a proportion. For instance, if the input address is "123 Main Street", we know that the 201-299 block of Main Street is closer than the 301-399 Main Street but we cannot determine a proportion of the total distance to either of these blocks because we have no scale by which to compare them.

To overcome this limitation and make d a useful metric, we define the concept of block distance, B_d, which is a measure of the number of blocks the input address number is away from the closer of the two address range end points associated with a reference street segment. However, to know B_d requires that d be expressed as a multiple of a block size $|b|$, i.e. how many addresses are on a block. Any candidate address range that contains the input address number is assigned a penalty score of zero, otherwise the candidate address range is calculated as one plus the number of blocks away because by definition, a candidate address range that does not contain the input address number is at least one block away as calculated in Equation 3.11.

$$B_d = 1 + \frac{d}{|b|} \tag{3.11}$$

The question now becomes: what value should $|b|$ take? One option is to use a single standard average block size to calculate B_d. It could be argued that $|b| = 100$ is a good estimate because most street segments address ranges in the U.S. are of this size. However, when using a single static block size, the actual block size of the address range of the reference feature $|b_f|$ may be either smaller or larger than this value. If $|b_f| < |b|$, some number of consecutive address ranges b_p will all fit within a single average block size (Equation 3.12). Each of these b_p reference features will receive the same proportional closeness score, which is not the desired behavior as the goal is to proportionally score address ranges that are closer with a lower proportion of the total penalty, and thus a higher match score. Note that the actual value of b_p is only roughly equivalent to the size of the actual block because each of the consecutive blocks may be a different size.

$$b_p \sim \frac{|b_f|}{|b|} \tag{3.12}$$

To overcome this problem, an algorithm for scoring nearby features needs to capture the fact that block sizes may vary across space. Even within small distances of each other, two neighboring blocks may have widely varying $|b_f|$ values. As a concrete example, consider the relationship between any intersecting street and avenue in Manhattan, New York. Although a street and an avenue may intersect and therefore share the same physical corner, the size of the address ranges per block are far smaller along the avenues than they are along the streets due to the address numbering scheme used throughout the majority of the area of New York City.

Therefore, to account for the variation of block sizes across space, our system calculates and uses a dynamic block size metric, $\left|\bar{b}\right|$, to determine the block distance for any candidate block. Because our feature matching algorithm is loose and does not filter candidate features by address range, the candidate feature set it returns should contain all street segments that have the same name and USPS ZIP code. This set provides a large representative sample of all address range sizes for the street of interest within the region of interest. If we group this set of candidate reference features by street name and USPS ZIP code, the average block size $\left|\bar{b}\right|$ can be calculated per street name and USPS ZIP code. Thus, $\left|\bar{b}\right|$ becomes a spatially varying quantity describing the average block size for a particular named street in a particular region which can be used to determine the block distance, B_d, of a candidate reference feature from the input address number (Equation 3.13).

$$B_d = 1 + \frac{d}{\left|\bar{b}\right|} \tag{3.13}$$

At this point, B_d represents a geographically relevant measure of the number of blocks a candidate feature is away from the input address number. However, again, this value alone cannot be used to calculate a proportional penalty value to associate with a candidate feature because we have no denominator from which to derive a proportion of the total penalty weight. Essentially, what is missing from our penalty calculation is a sense of a geographic scope within which we wish to limit our candidate features. To fill this void, we define the maximum block distance, B_M, to be the cutoff point beyond which candidate features are assigned the full penalty weight. This variable allows the user to express the search radius for valid candidate reference features. This value can be used to derive the proportion $\dot{w}_i$ of the full penalty weight w_i to associate with each candidate reference feature as the total output penalty p_i for the attribute (from Equation 3.4e). Our present implementation uses a linear decay function to express a reduction of confidence in a candidate feature as the block distance increases (Equation 3.14). This is similar to the notion used in calculating the character index-based penalty for the USPS ZIP code portion of the street address as described in Section 3.4.3.2.

$$\dot{w}_i = \frac{B_d}{B_M} \qquad (3.14)$$

Again, this function could be replaced with another decay function to weight nearby candidate features with a higher or lower penalty and change the rate at which match scores decline as candidate blocks get further and further away from the input address number. It should be noted that in our current linear model, the maximum block distance B_M controls the rate of the decline in confidence. If B_M is small $(B_M < 5)$, the full penalty will be applied after B_d calculates a candidate block is more than four blocks away. As B_M is increased, the penalty for outward moving blocks uniformly decreases, as does the variability in penalty between adjacent blocks meaning that the penalty score difference between adjacent blocks will get smaller and smaller.

3.5 Chapter 3 Experimental Evaluation

The two main goals of our approach are to: (1) determine if nearby match scoring can overcome address range attribute problems of reference data and improve match rates in address range geocoders; and (2) ensure that the output of such an approach is consistent with higher accuracy geocoders that do not suffer from such reference source errors. To evaluate either of these goals, we first need to establish a baseline comparison between the USC geocoder and one or more commercially available geocoding systems to show that the USC geocoder performs on-par with other state-of-the-art systems. Usage limitations prevent geocoding the entire 22,948 records used as our dataset with any of the online geocoders Google, Microsoft Bing, or Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b), so we will perform this initial evaluation over the complete dataset using the USC and ESRI Address Locator (Environmental Systems Research Institute 2009b) geocoders. The same reference data files will be used in both systems, meaning that the geocodes for cases where both systems successfully produce output geocodes should be very closely located to one another. Therefore, as a measure of comparability, we will evaluate the distances between the spatial locations produced by these systems. This spatial consistency evaluation between the two systems will determine if the USC geocoder is a comparable platform to the ESRI Address Locator.

Having done so, we can then proceed to evaluate our two specific research questions: (1) does allowing nearby matches improve match rates?; and (2) does using nearby references to produce an output geocode result in data of equivalent or better accuracy when compared to the current practice of returning the centroid of a randomly chosen nearby street segment?

The first question is intended to evaluate the level of match score improvement possible though the use of nearby candidate features. To do so, we will compare our approach against the ESRI Address Locator geocoding platform to determine the number of cases where our approach will select and return a nearby street segment with a sufficiently high match score whereas a traditional address range geocoder would return a non-match. However, simply improving the match rate of a geocoding engine alone is not a sufficient condition for proving that one geocoding algorithm is better than another because the improvements may be due to the introduction of false positives. Therefore, our second question is intended to evaluate the spatial accuracy of our results against geocoders with access to higher quality reference datasets, and therefore a higher likelihood of reducing these problem cases. To do so, we will compare the spatial locations produced using our approach versus those of the three online geocoders Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b). If the USC geocode is located close to that of one or more of the online geocoders, it tells us that the USC output has a high chance of being correct.

3.5.1 Data sources

The following sections describe the data sources used in our analysis. In particular, we describe: (1) the sample address data used as the test set; (2) the setup of the ESRI Address Locator and the USC geocoder including the reference date files used; and (3) the three online geocoding systems used.

3.5.1.1 Sample address data

The sample input address data used for our evaluation is derived from the National Provider Identification (NPI) file, a nationwide set of 3.9 million addresses of Medicare and Medicaid payment recipients representing hospitals, clinics, and doctors offices (U.S. Department of Health and Human Services 2004). In particular, our analysis focuses on the portion of the records within this file that are located within Los Angeles (LA) County, based on the USPS ZIP code of the practice location associated with an NPI record, $n = 96{,}062$. This dataset is commonly used throughout health-related research, so our evaluation herein serves to offer an heretofore missing report on the experience of geocoding these data. After removing duplicates identified by selecting only unique combinations of street address, city, state, and USPS ZIP code, our sample LA County dataset was reduced to $n = 22{,}984$ records.

3.5.1.2 ESRI StreetMap North America Address Locator

Testing our hypotheses that nearby candidate scoring improves geocode match rates only requires the comparison of our approach against an address range interpolation geocoder. Therefore, our experimental setup uses a single street reference data file instead of creating a multi-level geocoder which uses several layers of reference data to try to identify matches at lower levels of geographic resolution, e.g., USPS ZIP code centroids.

To simulate the most common usage of a typical desktop geocoding system used in academia, industry, and government (Rull et al. 2009; Omer et al. 2008; Blondin et al. 2007; Wagner et al. 2009; Macintyre et al. 2007; Ries et al. 2009; Wrobel et al. 2008; Ruiz et al. 2007; Pezzoli et al. 2007), the Address Locator (Environmental Systems Research Institute 2009c) portion of the ESRI ArcGIS 9.3 platform (Environmental Systems Research Institute 2009b) was used. The ESRI StreetMap North America (Environmental Systems Research Institute 2009d) local roads dataset was chosen as the reference data layer because it is widely used throughout research reports in conjunction with the ESRI Street Map North America Address Locator (Environmental Systems Research Institute 2009d) that is shipped with the data. This combination is commonly used by researchers to investigate spatially-based research questions because of its ease of use and reported levels of high accuracy, e.g. (Wrobel et al. 2008; Ruiz et al. 2007; Gupta et al. 2009).

Within the address locator, the default match score settings were used which include 80% spelling sensitivity, 10% maximum candidate score, 60% minimum match score. The “match candidates if tie” option was turned off (turned on in the default settings) because we did not want to accept ambiguous matches as valid outputs.

3.5.1.3 The USC geocoder

To build a USC geocoder implementation comparable with the ESRI StreetMap North America Address Locator (Environmental Systems Research Institute 2009d), the same ESRI StreetMap North America (Environmental Systems Research Institute 2009d) local roads dataset used as the base for the ESRI StreetMap North America Address Locator was imported as a reference data layer. This dataset is a dime-type reference dataset (i.e., two-sided) and maintains five alias street names for each street segment record along with left and right city names and USPS ZIP codes. Internally within the USC geocoder, this dataset is converted into a nickel-type (i.e., one-sided) reference dataset by expanding each of the alias names into a unique street segment record that maintains both the left and right city name and USPS ZIP code of the original segment to handle border cases where a street segment has different USPS ZIP codes or city names on the left and right side. The attribute weight assignments for each of the address components as well as the maximum acceptable number of blocks and minimum match score parameters used are listed in Table 3.4. These weights were chosen through an empirical iterative process which tuned the process for our regions of interest.

Table 3.4: USC geocoder parameters

Parameter	Value
Minimum match score	88
Weight pre-directional	7
Weight name	45
Weight post-directional	5
Weight suffix	10
Weight city	20
Weight USPS ZIP code	25
Weight number (when outside range)	15
Weight number parity (when on wrong side)	10
Maximum blocks away	5

3.5.1.4 Online commercial geocoding APIs

The three online geocoding systems used for our evaluation are the Google, Microsoft Bing, and Yahoo! APIs (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b). Although the terms of use and/or commercial cost of these services often prevent their use in large-scale geocoding attempts, the interactive one-off geocoding capabilities and map visualization tools provided by these services are commonly used by health researchers to investigate and correct problem addresses that do not geocode properly with other freely available or low-cost geocoding systems (Goldberg, Wilson, Knoblock et al. 2008).

For obvious commercial reasons, little is published about the inner workings of these systems. Only slightly more details are available about the specific data sources used by these geocoding sites, although it is known that the deep pockets and strategic business partnerships maintained by these organizations often means that the reference data sources available to their respective geocoding systems are far more complete and accurate than those available to individual researchers, organizations, and institutions, resulting in high match rates and high levels of spatial accuracy. For example, it has been reported that the Microsoft Bing geocoding engine uses data from the three major data providers TeleAtlas, Navteq, and Map Data Sciences (Pendleton 2008), while Yahoo! appears to rely exclusively on Navteq (Yahoo! Inc. 2009a) and Google has dropped all commercial data providers in favor of their own internally created street networks (Lookingbill 2009).

Each of these services provides interactive geocoding using a browser-based interface on their respective mapping sites. In addition, these services provide API versions of the geocoding services that may be called programmatically. The geocode accuracy metadata reported by these services vary across vendors and are not directly comparable. No quantitative match scores are returned by any of these systems. Instead, the extent of geocoding accuracy reported by these systems is most commonly a match type code that indicates the type of reference geographic feature that was matched such as address, roof centerline, etc. The Microsoft Bing system provides somewhat more information in the form of a confidence score of "Low", "Medium", and "High". The Yahoo! system is the only system that reports a non-exact match, returning an information message along with the result indicating that the exact input address was not found and that the match returned is the closest one available, marked by our system as a nearby result.

Although the terminology differs between them, all three systems purport to return matches at the street, USPS ZIP code, city, and state level, in addition to parcel, building rooftop centerline, and street intersection. However, only the Microsoft Bing geocoder returned either parcel or building accuracy in our tests. In addition, all three systems return multiple matches for a single request, so it is up to the user to determine which matching geocode to use. For our experiments, we always chose the highest level accuracy result based on the type of reference feature reportedly matched to in the following order (from most accurate to least): building roof centerline, parcel centroid, street address interpolation, street segment centroid, street intersection, USPS ZIP code centroid, and city centroid. Only single matches at the highest level are considered valid successful results. If multiple matches were reported at the highest level of accuracy, we mark the match as ambiguous and non-successful.

3.5.2 Results

The following subsections detail the results achieved by our system in comparison to the ESRI, Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b; Environmental Systems Research Institute 2009c) geocoding platforms. We first report the overall match rates achieved by the USC and ESRI geocoders as an indication of completeness and comparability of the two geocoding systems. As part of this analysis, we examine several of the classes of cases where the match type between the two systems differ to reveal intricate differences between the approaches taken as well as the level of improvement that can be gained by using our nearby feature matching approach. We next evaluate the spatial relationships between the USC and ESRI geocodes that were both successfully geocoded to determine if the USC geocoder performs on-par with the ESRI geocoder. We then turn our attention to the comparison of the USC geocoder versus the three online geocoding systems to determine if the improvements in match rates achieved by the USC geocoder over the ESRI geocoder are in fact valid plausible output locations. To do so, we evaluate the two cases that produced these improvements; those due to our nearby feature matching approach and those due to our general overall matching approach without requiring nearby match scoring.

3.5.2.1 USC versus ESRI match rates and types

The overall match rates for non-ambiguous street-level matches between the ESRI and USC geocoder were both quite high, 95% and 96.6%, respectively, indicating a high level of input dataset accuracy. When considering both exact and nearby matches, the USC geocoder match rate increases to 97.7%. The match types and respective counts returned by the ESRI geocoder are shown in Table 3.5, while the match types and respective counts returned by the USC geocoder are shown in Table 3.6, which also lists the type of unmatchable record (this information is not available for the ESRI geocoder).

Table 3.5: ESRI match types and counts

Match type	Count	Percent of total
Matched	21,836	95%
Tie	68	0.3 %
Unmatchable	1,080	4.7%

Table 3.6: USC geocoder match types and counts

Match type	Count	Percent of total
Matched	22,198	96.6%
Tie	145	0.6%
Nearby	243	1.1%
Unmatchable - Below minimum score - No candidates - Exception occurred	398 292 41 65	1.7% 1.3% 0.2% 0.3%

Of the 398 unmatchable records reported by the USC geocoder, the 292 records listed as “below minimum score” indicate records where candidates were available but their score did not meet or exceed the minimum match score of 88%. The 41 with “no candidates” indicate records for which no reference features could be identified in the reference data source. The 65 where an “exception occurred” indicate records where the address parser used within the USC geocoder was not capable of parsing the address number portion of the input address. These cases occurred where there was no address number or extraneous text preceding the address number portion of the input data such as in “Ronald Reagan Hospital, Los Angeles, Ca 90095” and “UCLA Medical Center 757 Westwood Plaza, Los Angeles, Ca 90095”. The comparison of USC match type versus ESRI match type listed in Table 3.7 reveals a high correlation between the match types of the two geocoding engines.

Table 3.7: USC versus ESRI match type comparison

	ESRI Matched	ESRI Tie	ESRI Unmatched	Total
USC Matched	21,797	36	365	22,198
USC Tie	6	29	110	145
USC Nearby	22	0	221	243
USC < Minimum	8	2	282	292
USC No candidates	0	0	41	41
USC Exception	3	1	61	65
Total	21,836	68	1080	22,984

Of the eight cases where the ERSI geocoder returned a valid match and the USC geocoder returned one with less than a minimum match score, four are due to reference data errors which indicate that the ESRI StreetMap North America Address Locator was built with a slightly different reference dataset than the ESRI StreetMap North America reference file used in the USC geocoder.

In the first, “8716 Cord Ave, Pico Rivera Ca 90660”, the street name in the reference data files is “CR” indicating that it was normalized as a “CORD” = “County Road”. This error is not present in the ESRI StreetMap North America Address Locator which therefore obtains a perfect match between input street name and reference feature street name. Similarly, the second “14510 Baldwin Park Towne Ctr, Baldwin Park Ca 91706” matches to a street segment in the reference data file with the name “Baldwin” and the suffix “Park”, although the StreetMap North America Address Locator returns a street segment named “Baldwin Park Towne” and suffix “Ctr”. In this case, the USC geocoder parses the input data into the same name and suffix as the ESRI StreetMap North America Address Locator (“Baldwin Park Towne” and “Ctr”) and matches to the same street segment, but due to the degree of non-similarity between the parsed name and reference feature name, the score returned is less than the minimum score threshold. In the case of the third, “14445 Olive View Dr, Sylmar Ca 90402”, the street segment matched to in the reference data file is a completely different city “Los Angeles, Ca 91342” than that returned from the ESRI StreetMap North America Address Locator “Sylmar, Ca 91342”, neither of which have the input USPS ZIP code value. With both city and USPS ZIP code associated with the reference feature having highly dissimilar values from the input address, the USC geocoder is not able to identify this street segment as a candidate because there is no indication that it is in the correct region. However, because the output city from the ESRI StreetMap North America Address Locator lists the correct city “Sylmar”, it becomes clear that the city name used during the creation of the ESRI StreetMap North America Address Locator was a different (correct) value, unlike what is available in the ESRI StreetMap North America reference data. The fourth case, “923 W Carson St, Torrance Ca 90024” has the same problem, being located in “West Carson Ca 90502”, according to the respective street segment located in the ESRI StreetMap North America reference data.

In the four remaining cases, the USC geocoder fails to match because the reference feature selected has a match score less than the minimum score due to typographical errors in input data. Three of these four are because of incorrect Soundex values being calculated and used to search for candidates in the reference dataset (Table 3.8). In the first two, the placement of a single incorrect character and the number of vowels present in each street name word significantly change the calculated Soundex value.

Table 3.8: Cases where ESRI geocoder matches and USC geocoder does not match due to Soundex errors resulting from input data typographical errors

Input address	Incorrect name	Correct name	Error Soundex	Correct Soundex
11033 East Rosecraws Ave	Rosecraws	Rosecrans	R262	R226
3333 Sypark Drive	Sypark	Skypark	S216	S162
1200 N. State St. Ct	State Street	State	S323	S330

In the third, the extraneous street suffix, “Ct”, forces the real suffix, “St”, to become part of the street name “State Street Ct” which then produces an incorrect Soundex value. In all three cases, the erroneous Soundex values calculated prohibit the correct street segments from being selected as candidate features. In the fourth and final case, the street number and name are collapsed into a single token, “14600Sherman Way Suit200 Van Nuys 91405”, which results in a parsed address without a street name.

All six records that the ESRI geocoder was able to match but resulted in ties within the USC geocoder are due to the same reference data discrepancies. In particular the city and USPS ZIP code combinations of the reference features selected by the ESRI geocoder do not match the values within the ESRI StreetMap North America reference files for the first four while the misspelling of the street name of the fifth caused a non-match (Table 3.9).

Table 3.9: Cases where ESRI geocoder matches and USC geocoder returns ties due to reference dataset discrepancies

Input address	Reference
3300 East South Street, Long Beach Ca 90280	Lakewood 90712
1000 W Carson St, Torrance Ca 90509	West Carson 90502
1000 W. Carson St, Torrance Ca 90505	West Carson 90502
1000 Carson St, Torrance Ca 90509	West Carson 90502
5831 Overhill Dr, Los Angeles Ca 90016	View Park 90043
2001 Whittier Blvd, La Habra Ca 90631	Whitter Blvd

A comparison of the 22 records that resulted in ESRI matches and USC geocoder nearby matches reveals a major difference between the match scoring strategies in the two geocoding platforms. The records that result in this situation are shown in Table 3.10 which displays the distance between the two output geocodes, the complete input address, the street pre-directional, name, and suffix of the reference feature matched by the ESRI geocoding engine, and the address range, pre-directional, name, and suffix of the reference feature matched by the USC geocoder. The city and USPS ZIP code values are omitted from the USC and ESRI address columns because in all cases, each is equivalent to those associated with the input address. The address number is omitted from the ESRI address column because, by definition, the input address must be within the address range of the street segment for the ESRI geocoder to mark it as a match. As Table 3.10 shows, the ESRI geocoder chooses a street segment with either the wrong directional, suffix, or both in 20 of 22 cases, while the USC geocoder chooses the closest street segment with the correct directional and/or suffix.

Table 3.10: ESRI street-level exact matches versus USC nearby matches

Id	Distance (m)	Input Address	ESRI Address	USC Address	Error
1	1,845	767 S Sunset Ave West Covina 91790	N Sunset Ave	789-769 S Sunset Ave	Dir
2	1,487	430 W 97th Street Los Angeles 90003	E 97th St	400-408 W 97th St	Dir
3	926	319 W Tudor St Covina 91722	E Tudor St	225-299 W Tudor St	Dir
4	858	7750 Carson Blvd Long Beach 90808	Carson St	7698-7660 Carson Blvd	Suffix
5	545	400 S Sepulveda Blvd Ste 210 Manhattan Beach 90266	N Sepulveda Blvd	398-366 S Sepulveda Blvd	Dir
6	418	1330 South Fullerton Road Rowland Heights 91748	Fullerton Rd	1370-1352 S Fullerton Rd	Dir
7	412	309 E 2nd St Pomona 91766	W 2nd St	201-299 E 2nd St	Dir
8	326	11275 1/2 Washington Pl Culver City 90230	Washington Blvd	11251-11265 Washington Pl	Suffix
9	118	409 E Merced Ave West Covina 91790	Merced Pl	301-399 E Merced Ave	Dir
10	108	421 E Merced Ave West Covina 91790	Merced Pl	423-455 E Merced Ave	Dir /Suffix
11	102	410 E Merced Ave West Covina 91790	Merced Pl	300-398 E Merced Ave	Dir /Suffix
12	92	2618 Los Coyotes Diagonal Long Beach 90815	N Los Coyotes Diagonal	2598-2500 Los Coyotes Diagonal	Dir
13	84	2101 N Hillhurst Ave Los Angeles 90027	Hillhurst Ave	2129-2159 N Hillhurst Ave	Dir
14	26	1220 S Golden West Ave Arcadia 91007	N Golden West Ave	1100-1198 S Golden West Ave	Dir
15	22	15500 S Normandie Ave Ste B Gardena 90247	Normandie Way	15408-15498 S Normandie Ave	Dir
16	22	1757 N Lake Ave Pasadena 91104	Lake Ave	1653-1749 N Lake Ave	Dir
17	19	15506 S Normandie Ave Gardena 90247	Normandie Way	15408-15498 S Normandie Ave	Dir
18	19	15508 S Normandie Ave Gardena 90247	Normandie Way	15408-15498 S Normandie Ave	Dir
19	9	10418 East Valley Blvd El Monte 91731	E Valley Mall	10426-10458 Valley Blvd	Suffix
20	9	10418 Valley Blvd El Monte 91731	10418 Valley Mall	10426-10458 Valley Blvd	Suffix
21	403	1106 N La Cienega Blvd W Hollywood 90069	N La Cienegra Blvd	1098-1096 N La Cienega Blvd	Spelling
22	403	1106 N La Cienega Blvd West Hollywood 90069	N La Cienegra Blvd	1098-1096 N La Cienega Blvd	Spelling

In the remaining two cases (#21 and 22 in Table 3.10), the USC geocoder was again stifled by minor misspellings in the reference feature street segment name. A manual review of each of the 20 differing cases was performed that revealed that the USC geocoder had made the correct decision in each case. Four of the records (#13, 16, 19, and 20) were geocoded to the same street segment by the USC and ESRI geocoder. In these cases, the addresses values of either the pre-directional (#13, 16) or suffix (#19, 20) were incorrectly listed in the reference data, therefore the USC geocoder returned the code of nearby instead of exact, correctly indicating that the match was not exact while still returning the correct result. In the remaining 16 cases, the ESRI geocoder selected the wrong street segment because the address number of the input address was just outside the address ranges listed on the reference data street segments. Investigating the segments next to the one where the address most likely falls reveals that in all cases, if one assumes a consistent numbering scheme across contiguous blocks, the USC geocoder would place the output geocode in the correct location. In addition, the true location of each address was attempted to be verified by calling the phone numbers associated with the business listed with the NPI record. Only 17 out of 20 records had phone numbers, and of these, only eight calls were successful, but all revealed that the USC geocoder chose the correct output location.

The distance between the geocodes produced by the USC and ERSI geocoders are shown in Table 3.11 which lists the minimum, maximum, average and standard deviation for the entire dataset, including outliers. These data reveal that on the average, the two geocoders produce geocodes that are within 70 m of each other, while 67% of the data are within 73 m of each other. The distribution of these positional offsets shown in Figure 3.5 reveals that the vast majority (21,736 out of 21,793=99.7%) of the data are within 500 m of each other, with most of these actually being within 100 m (17,023 out of 21,793=78.1%).

Table 3.11: Distance between USC and ESRI matched geocodes

Minimum (m)	Maximum (m)	Average (m)	Standard dev (m)
2.2	1,995	70	73

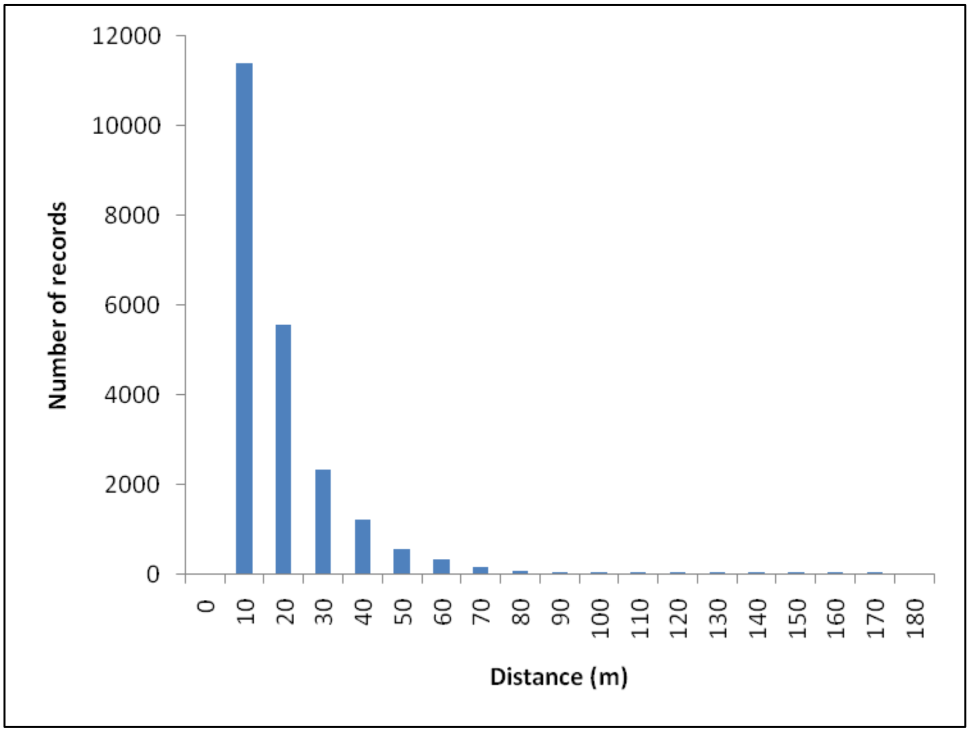


Figure 3.5: Distance between USC and ESRI geocodes

The 57 records that differed in spatial location by more than 500 m were manually investigated which revealed that in all but one case, the USC geocoder and the ESRI StreetMap North America Address Locator selected the same street segment, with the spatial offsets resulting from differences in the interpolation percentage taken along the street segment. The one case where the two geocoding platforms disagreed on street segment was for the input address “100 Carson Mall, Carson Ca 90745”. Here, the ESRI StreetMap North America Address Locator selected a reference feature with correct name and incorrect suffix in the correct USPS ZIP code “100 Carson St, Carson Ca 90745”, whereas the USC geocoder selected a feature with the correct suffix in a neighboring USPS ZIP code “100 Carson Mall, Carson Ca 90746”.

An online check of the true address of this record (NPI=1265580716) revealed that the USC geocoder made the correct choice because the facility is located within the South Bay Pavilion shopping center which resides on the street segment chosen by the USC geocoder.

3.5.2.2 USC versus online geocoding systems

Both the Microsoft Bing and Yahoo! geocoders were able to successfully geocode all addresses associated with nearby output from the USC geocoder, while the Google geocoder failed to return a match for three of the input queries and incorrectly returned two as being located in some other state. However, no geocoder was able to correctly match all of the 243 input addresses to a street-level or better geocode, with the Google, Microsoft Bing, and Yahoo! geocoders each only matching 201 (83%), 218 (90%), and 192 (79%) respectively, although each was successful in a high percentage of the total nearby cases.

Of the nearby cases that were successfully processed by the online geocoders, 153 records were matched to street-level accuracy or better across all geocoders (i.e., exact street level, building, or parcel), representing the subset of data with the highest level of spatial location concordance among the three geocoders. Of these, the geocode with the minimum distance to the USC geocoder was within 125 m in all cases which tells us that the USC geocoder produced an accurate output.

The distances between the output of each online geocoder and the USC geocoder for the 243 cases resulting in nearby matches from the USC geocoder are displayed in Table 3.12 which lists the minimum, maximum, average and standard deviation of great circle distance between the USC geocoder against each of these platforms as well as just for the subset of cases included in the 153 concordant matches across all online geocoders. Also shown are the same statistics calculated for the average of the three distances from each geocoder (Google distance + Microsoft Bing Distance + Yahoo! Distance / 3) as well as just the single geocode with the minimum distance to the USC geocode. A more detailed view of the same results for each geocoder broken up by match types and output geography type are shown in the Tables 3.13 through 3.15.

Table 3.12: Distance between USC and online geocoders for all match types across all geocoders

Geocoder	N	Min (m)	Max (km)	Mean (km)	St dev (km)
Google all records	238	3	43.2	0.8	3.7
Google concordant records	153	3	2.84	0.19	0.29
Yahoo! all records	243	4	18.6	0.5	2.0
Yahoo! concordant records	153	4	1.01	0.14	0.18
Bing all records	243	5	4.7	0.34	0.6
Bing concordant records	153	6	1.01	0.15	0.18
Average all records	238	8	14.7	0.5	1.5
Average concordant records	153	8	1.00	0.16	0.19
Minimum all records	238	3	2.4	0.2	0.3
Minimum concordant records	153	3	0.98	0.12	0.16

Table 3.13: Bing geocoder match types and distance by level of geography for all nearby records

Match type	Geography type	Number n=243	Min (m)	Max (km)	Mean (km)	St dev (km)
Exact	Building	22	26.4	3.0	0.1	0.1
Exact	Parcel	61	19.2	1.0	0.2	0.2
Exact	Segment	135	6.3	2.9	0.2	0.3
Exact	Street centroid	11	4.7	4.5	0.9	1.3
Ambiguous	Segment	9	33.8	1.1	0.4	0.3
Ambiguous	Street centroid	5	1085.7	4.7	2.9	1.37

Table 3.14: Yahoo! geocoder match types and distance by level of geography for all nearby records

Match type	Geography type	Number n=243	Min (km)	Max (km)	Mean (km)	St dev (km)
Exact	Segment	192	0.004	7.8	0.2	0.6
Nearby	Segment	45	0.005	5.5	0.4	0.8
Ambiguous	Segment	3	1.1	8.7	3.7	4.3
Nearby	City	3	13.2	18.6	16.8	3.1

Table 3.15: Google geocoder match types and distance by level of geography for all nearby records

Match type	Geography type	Number n=243	Min (m)	Max (km)	Mean (km)	St dev (km)
Exact	Segment	201	2.8	2.8	0.2	0.3
Exact	Street centroid	12	42.9	3.9	1.6	1.3
Ambiguous	Segment	12	7.7	5.5	0.9	1.5
Ambiguous	Street centroid	9	902.7	27.1	4.4	8.5
Exact	City	4	833.1	43.1	20.4	17.4

In addition to the 243 cases where the match rate improvement achieved by the USC geocoder over the ESRI geocoder was due to the use of the nearby matching strategy, 365 additional cases were geocoded by the USC geocoder that were missed by the ESRI geocoder due to the other aspects of the matching strategy employed in the USC geocoder. Of these 365 records, only the Microsoft Bing geocoder was capable of geocoding all addresses matched by the USC geocoder, and only 295 (81%), 298 (82%), 204 (56%) were successfully geocoded to street-level or better accuracy (i.e., exact street level, building, or parcel) by the Google, Microsoft Bing, and Yahoo! geocoders, respectively.

However, of these 365 records, 355 (97%) were successfully geocoded to street-level accuracy or better by at least one of the three online geocoders with the average distance to the geocode being 135m (min=2 m, max=5.5 km, std=374 m), but of these only 153 records were matched to street-level accuracy or better across all geocoders. In these concordant cases, the geocode with the minimum distance to the USC geocoder was within 500 m in 147 of the cases (96%) and 133 (87%) are within 250 m.

Table 3.16 lists the minimum, maximum, average and standard deviation of great circle distance between the USC geocoder against each of these platforms as well as just for the subset of cases included in the 153 concordant matches across all online geocoders. Also shown are the same statistics calculated for the average of the three distance from each geocoder (Google distance + Microsoft Bing Distance + Yahoo! Distance / 3) as well as just the single geocode with the minimum distance to the USC geocode. These same metrics broken up by match type and output geography type are shown for each of the online geocoders in Tables 3.17 through 3.19. A total of nine records were successfully geocoded by the USC geocoder to street level when all three of the online geocoders failed (Table 3.20).

Table 3.16: Distance between USC matched and online geocoders for street-level matches across all geocoders

Geocoder	N	Min (km)	Max (km)	Mean (km)	St dev (km)
Bing all records	298	0.007	658.147	3.458	41.551
Bing concordant records	153	0.008	20.652	0.389	1.777
Yahoo! all records	204	0.002	5.537	0.241	0.617
Yahoo! concordant records	153	0.003	5.537	0.272	0.694
Google all records	295	0.005	2203.54	8.04	128.298
Google concordant records	153	0.005	2203.54	14.696	178.124
Average all records	355	0.011	4829.888	99.389	576.978
Average concordant records	153	0.012	734.533	5.119	59.363
Minimum all records	355	0.002	5.537	0.135	0.374
Minimum concordant records	153	0.012	734.533	5.119	59.363

Table 3.17: Bing geocoder match types and distance by level of geography for nearby records matched by all three online geocoders

Match type	Geography type	N	Min (km)	Max (km)	Mean (km)	St dev (km)
Exact	Building Centroid	54	0.021	0.356	0.097	0.076
Exact	Parcel	101	0.02	5.59	0.222	0.563
Exact	Street Segment	143	0.007	658.147	7.013	59.885
Exact	Street Centroid	10	0.422	29.198	6.605	8.692
Exact	USPS ZIP	20	0.112	37.206	4.22	10.404
Ambiguous	Street Centroid	8	0.138	30.439	15.729	11.774
Ambiguous	Street Segment	28	0.009	11.952	0.625	2.249

Table 3.18: Yahoo! geocoder match types and distance by level of geography for nearby records matched by all three online geocoders

Match type	Geography type	N	Min (km)	Max (km)	Mean (km)	St dev (km)
Exact	Street Segment	204	0.002	5.537	0.241	0.617
Ambiguous	Street Segment	3	0.008	0.063	0.038	0.028
Nearby	Street Centroid	1	1.199	1.199	1.199	-
Nearby	Street Segment	115	0.002	11.85	0.316	1.27
Nearby	USPS ZIP	1	27.253	27.253	27.253	-
Nearby	City	35	0.292	36.874	6.722	9.452
Unmatchable	-	5	-	-	-	-

Table 3.19: Google geocoder match types and distance by level of geography for nearby records matched by all three online geocoders

Match type	Geography type	N	Min (km)	Max (km)	Mean (km)	St dev (km)
Exact	Street Segment	295	0.005	2203.54	8.04	128.298
Exact	Street Centroid	9	0.097	3438.916	973.149	1476.533
Exact	City	27	0.081	26.8	7.107	7.462
Ambiguous	Street Segment	24	0.015	4160.677	607.998	1375.782
Ambiguous	Street Centroid	7	2.319	3456.465	509.92	1299.576
Ambiguous	Intersection	1	6.775	6.775	6.775	-
Unmatchable	-	1	-	-	-	-

Phone calls revealed that the USC geocoder geocoded the input address to the correct location in all but the second case. The street address associated with this erroneous record in the NPI database incorrectly lists this address as being on “E Avenue” with a unit of “K6”, while in reality, the address is located on “E Avenue K6”. Therefore, the USC geocoder was not able to geocode the address to the correct physical location.

Table 3.20: Addresses USC could geocode to street level that the online geocoders could not

Id	Address	USC Feature
1	341E. Center Drive Anaheim Ca 90660	317-399 E Center St Anaheim CA 92805
2	349-A E. Ave. Lancaster Ca 93535	225-201 E Ave Lancaster CA 93534
3	1000 W Carson St Carson Ca 90509	1012-862 W Carson St West Carson CA 90502
4	1329 N Lubre Ave Inglewood Ca 90302	1403-1309 N La Brea Ave Inglewood CA 90302
5	10833 Le Conte Ave UCLA Medical Ctr Los Angeles Ca 90095	10949-10801 Le Conte Pl Los Angeles CA 90095
6	6290 E PCH Long Beach Ca 90803	6264-6352 1 Long Beach CA 90803
7	3737 Martin L. KkngBlvd Lynwood Ca 90262	3735-3741 Martin Luther King Jr Blvd Lynwood CA 90262
8	501 D Valley Drive Palos Verdes Ca 90274	659-501 Deep Valley Dr Rolling Hills Estate CA 90274
9	25825 S Vermont Parkview Bldg Harbor City Ca 90710	25771-25983 Vermont Ave Los Angeles CA 90710

3.5.3 Discussion

Our results illuminate several key findings about both the USC geocoder in general and our nearby matching approach in particular. First, the USC geocoder and ESRI geocoders are both quite capable of geocoding a very high percentage of the NPI file for the Los Angeles region. Our analysis of the distance between the locations produced by both geocoding systems shows that the USC geocoder generates similar results to the ESRI geocoder. This result is not unexpected because both geocoders utilize the same ESRI Street Map North America reference data files. However, this result does show that the USC geocoder is capable of performing on-par with the industry standard, which thereby justifies its use to investigate further technical geocoding improvements such as the nearby match scoring method we proposed.

Second, the comparison of match rates between the USC and ESRI geocoders reveals that the USC geocoder not only competes with, but actually outperforms the ESRI geocoder, even when only considering normal address range geocoding alone, i.e., without employing the non-nearby matching strategy. In all instances where the USC geocoder obtains a match without using nearby match and the ESRI geocoder does not, the high degree of spatial correlation between the output locations of the online geocoders and that produced by the USC geocoder provides evidence that these improvements in match rate are true positives whose locations are accurately placed. Likewise, when using the nearby matching strategy, the USC geocoder produces geocodes that are quite close to those produced by the more accurate reference data sources available to the online geocoding systems. Taken together, these results show that our nearby matching strategy both improves match rates and produces accurate output, thereby overcoming limitations of the attributes available in the reference data files.

Third, the record-by-record review of the ESRI exact matches and the USC geocoder nearby matches reveals that the ESRI results that differ on directional and/or suffix are essentially false positives. In these cases, the ESRI geocoder fails to obtain an exact match, instead returning a similar, yet incorrect, match. Essentially, our evaluation shows that the ERSI geocoder is prone to giving a false positive rather than returning a failure because it scores the missing/wrong pre-directional or suffix higher than a correct, nearby segment. The major problem with the ESRI approach is that these errors occur on the street segment types where it matters, i.e., those that have both a north and a south segment which will be linearly far away from each other as the numbers increase: 10x, 20x, 30x being two, four, and six blocks away from the truth, whereas the nearby will always be on the closest one (within the range) or marked as a non-match. This behavior of the ESRI geocoder has the potential to introduce spatial error into geocoded spatial datasets. Within this test dataset, the average distance between the USC and ESRI output geocodes (excluding the misspelling cases) is 372 m, a relatively small distance given the fact that streets with different directionals are often across town from one other, so the potential harm that could be introduced by the ESRI false positives is minimal. However, one quarter of the records are more than 500 m apart and two of them are over 1 km apart so the potential introduction of spatial error may become a more serious problem depending on the peculiarities of a specific input dataset.

In contrast, the USC geocoder does not attempt to return the similar incorrect match, opting instead to return a match as close as it can find to the exact correct match. This behavior is consistent with that of researchers who often perform manual review on these non-100% matches cases. Here, the normal protocol is to choose a location somewhere on the nearest street segment (Goldberg, Wilson, Knoblock et al. 2008). The nearby feature match scoring approach developed in the present work results in the same outputs, without the need for manual intervention.

Finally, all of the cases that were successfully geocoded by the ESRI geocoder and/or any of the online geocoders that failed within the USC geocoder were due to errors in either the reference or input data. This highlights one of the key limitations of a deterministic match scoring system such as that described here. In our approach, the Soundex values of names are used to block for candidate features in the reference dataset. The Soundex approach is highly vulnerable to the number and placement of vowels in a word and as such, cases do occur where this results in an incorrect blocking value being used to select reference features. The edit distance function used to determine a weighted similarity score is likewise vulnerable to the number and placement of non-matching characters between the input and reference address words. In a deterministic approach, little can be done to overcome these types of errors without simultaneously and dramatically increasing the rates of false positives. In the present investigation, there were only a few instances of these types of errors which did not significantly impact the overall quality of the USC geocoder's output for the input dataset as a whole.

When compared to the online geocoders used in this investigation, several other interesting facts come to light as well. In the case of an unmatchable exact result most relevant to our approach, each of the online geocoding systems attempted to return “nearby” matches. In the most typical use-case, via the online map interface, this occurs without informing the user because these systems do not report any type of quantitative match score along with the output geocode results. Instead, they simply return a point and the user is left to assume that the data he/she just entered perfectly matched the output location. In API mode, where the user queries the web services programmatically using function calls, this situation improves in that the user is returned a set of possible candidate results. However, these are raw data and the user is left to determine which of these candidates is the most correct as well as what level of confidence should be placed in it. Our approach could be applied in conjunction with these services to calculate candidate scores for each of the candidates.

Perhaps most troubling is the description from the Manifold Systems Manifold Geocoding Database, another popular commercial desktop geocoding application, “If an address cannot be found in the available street address ranges for a specific street the command will choose a point near the middle of the street segment” (Manifold Net Ltd. 2009). If, by definition, the input address number is not contained within the address range of any reference features leading to the situation where the street number cannot be found in the available street address ranges, then which of the available street segments does the system use to create a centroid, and how and why was it chosen?

These types of “automatic correction” make two assumptions: (1) the input data is incorrect and, (2) the user wishes the system to automatically “fix” their input address to be valid in comparison with the reference data. These assumptions lead to misleading results and the inclusion of false positives in the output datasets. In many cases the user assumes that since the service returned a geocode for their input query, it was an exact match, especially in the cases where they are not informed otherwise. Our approach overcomes both of these assumptions in that it, first, allows for errors in the reference data by providing the ability to match, score, and return matching reference features with slightly (or highly if the user wishes) erroneous street number attributes, the most common form of reference feature attribute error. Second, our system is still capable of “correcting” erroneous input data, in that it is capable of producing the same output as these geocoding engines (i.e. the geocode produced using a nearby reference feature), but it additionally informs the user that the quality of the output data is not an exact match.

3.6 Chapter 3 Conclusions

In this study we have developed a method for calculating, ranking, and returning match scores for geocoding candidate reference features that are not presently handled by traditional address range geocoding systems. Our approach determines the closest street segment to the requested input address and proportionally scores the candidate based on a notion of closeness defined in terms of the number of blocks within a user-defined maximum search distance.

As part of this process, we have detailed the inner working of the USC geocoding system including the reference data sources used, as well as the feature matching, scoring, and interpolation strategies employed to ensure that other researchers and engineers may replicate our results and/or create similar systems.

To evaluate the benefits of our approach, it was first necessary to show that the USC geocoder performs on-par with other industry standard geocoding systems, namely the ESRI Address Locator based on the ESRI StreetMap North America reference files, a widely used resource in industry, academia, and government. Our results indicate that the USC geocoder achieves this goal, producing the same spatial output as the ESRI geocoder for records which they are both able to geocode. Next, we investigated the two types of instances where the USC geocoder is capable of improving the match rate above and beyond that produced by ESRI; those due to our nearby matching strategy and those due to our deterministic match scoring algorithm. In both cases, the results provided by three independent online geocoding platforms, Google, Microsoft Bing, and Yahoo!, confirmed that the USC geocoder results were actually true positives, not simply incorrectly placed false positive inflating the match rate.

In sum, the combination of our deterministic feature scoring strategy in conjunction with the nearby match candidate scoring achieves the goals set forth in the present research of improving the match rate of geocoded datasets while at the same time producing spatial data that are of high spatial accuracy.

CHAPTER 4: IMPROVING GEOCODE ACCURACY WITH CANDIDATE SELECTION CRITERIA

This chapter will be published as:

Goldberg, D.W., 2010.

Improving Geocode Accuracy with Candidate Selection Criteria.

Transactions in GIS 14 in press.

4.1 Chapter 4 Introduction

Geocoding is the process of converting postal address data into geographic coordinates, i.e., latitude and longitude pairs (Boscoe 2008). To accomplish this task, geocoding systems first use one or more feature matching algorithms to attempt to correlate an input address with a geographic object present in a reference dataset such as a street segment that contains the address or the parcel to which the address belongs. Once the most likely candidate reference feature is identified, a geocoder will use a feature interpolation algorithm to calculate an approximate output position along or within the reference feature if necessary, as in the case of linear or areal reference features such as street segments and parcel geometries, respectively. In the case of point-based reference data sources such as address points derived from field surveys with GPS devices, interpolation is not necessary and the reference feature location is returned directly.

Most geocoding systems in use today include multiple implementations, representations, and/or versions of reference datasets to improve performance based on characteristics or assumptions that can be made about a particular input dataset, a geographic region, a period of time, or the availability of a beneficial data source. These performance improvements are usually measured in terms of increases to match rates, the number of records it is able to match, or the spatial accuracy of the output of the system, the distance between the calculated and the true location. Common combinations found within the commercial online geocoding systems Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) include building centroids, parcel geometries, street segments, and centroids for USPS ZIP codes, cities, counties, and states. In this arrangement, if a building centroid is found for a particular address it will be returned. If not, the systems will revert to attempting to locate a parcel for the address, then the street segment, then the USPS ZIP code, etc., until a match is finally found and used to interpolate an output which is returned to the user. This practice ensures the match rate of the system remains high – *some* geocode output will be returned for every input address – but the spatial resolution of features within the reference data layers quickly degrades, resulting in lower and lower utility for the end user as a match is attempted but not found in subsequent layers – e.g., how useful is a county centroid geocode?

Clearly, using multiple reference datasets can result in a large set of candidate geocode values that could be returned – potentially one from within each reference dataset – of which the end user is generally only concerned with what the system declares to be the single "best" result. Therefore, any geocoding system that maintains multiple reference data layers must choose and return a single output geocode based on some criterion, be it an arbitrarily-chosen sequence created by the developer, a scientifically-based ordering drawn from a set of experiments, or a discipline-specific set of conventions. The trouble here is that although every geocoding platform which includes multiple reference datasets clearly must make these decisions, the reasoning behind these choices in terms of how and why a particular output was chosen above all others is not typically reported along with the result, nor are the alternatives that were not chosen and may have been useful in their own right.

While in some cases it may be argued that investigating the way a geocode result is chosen from among the possible levels is a moot point because data that are not geocoded to parcel or street level should just be thrown away, this is not the standard practice in certain research fields. For instance, in the health sciences where geocoding often creates the underlying data used in studies determining the link between location and a certain health outcome or activity pattern (Sui 2007; Rushton et al. 2006), numerous reports have shown that this practice would introduce serious bias which could invalidate results and/or conclusions drawn from the remaining data (e.g., Oliver et al. 2009; Zandbergen et al. 2007; Krieger, Waterman et al. 2002). Furthermore, in the U.S. many states mandate that disease registries geocode their case data for use in disease surveillance activity, such as the California Cancer Registry which is required by California law to geocode a certain portion of their incidence data. Finally, our prior work (Goldberg, Wilson, Knoblock et al. 2008; Goldberg 2008) as well as that of others (Henry et al. 2008; Boscoe et al. 2002; Shi 2007; Rushton et al. 2006) investigated match rates and quality types reported in numerous recent health studies and determined that geocodes of less than street level accuracy will present a consistent challenge for some time to come.

That said, no research has presented an evaluation of the methods and/or choices that researchers and engineers may use to determine the best output from a set of possible candidate geocodes drawn from multi-layer reference datasets. Furthermore, little evidence has been presented to-date which shows one selection strategy as being better than any other. In the present work we propose to formalize the decisions used to select the "best" geocode output from the set of possible candidate outputs as an optimization problem using an objective function we term the *best-match criterion*. This formalization enables us to study the implications of one selection strategy over another in terms of overall geocode accuracy, used herein to mean the distance between the computed location and ground truth. With this theoretical basis in hand, we contextualize and examine the shortcomings of the current state-of-the-art best-match criterion, which we call the *hierarchy-based criterion*, and propose three alternative strategies, termed the *uncertainty-, gravitationally-, and topologically-based criteria* that improve on the status quo.

The remainder of this chapter is organized as follows. In Section 4.2 we discuss related work. In Section 4.3, we provide the background needed for the development of the concept of a best-match criterion, fit the current practices into this model, and provide the rationale and theoretical foundation for our three proposed best-match criteria. In Section 4.4 we evaluate the performance of our methods on a national dataset of ground truth data. We end with conclusions and future directions in Section 4.5.

4.2 Chapter 4 Related Work

The existing research related to the present work can be broadly grouped into two categories. The first investigates the causes, characteristics, and effects of geocode accuracy, while the second proposes technical advances to improve geocode accuracy. The majority of existing research in the first category has attempted to understand and explain the foundations of the geocoding process (Boscoe 2008), especially with regard to empirical analyses of resulting spatial error (Strickland et al. 2007) introduced from varying geocoding methods (Whitsel et al. 2006; Zhan et al. 2006) including reference datasets (Beyer et al. 2008; Gatrell 1989; Ratcliffe 2001; Schootman et al. 2007; Zandbergen 2008a), matching algorithms (Levine et al. 1998), and/or interpolation algorithms (Cayo et al. 2003; Dearwent et al. 2001; Gilboa et al. 2006). These studies have also investigated how the spatial accuracy and/or match rates may be correlated with one or more components of the geocoding process (Martin et al. 1999), input data (Hurley et al. 2003; Krieger, Waterman et al. 2002), or underlying geography (Zimmerman et al. 2007; Zimmerman et al. 2010), and how the accuracy of geocoded data affects subsequent research (Krieger et al. 2001; Mazumdar et al. 2008; Oliver et al. 2009; Zandbergen 2007; Zandbergen et al. 2007). The present work builds on this previous research and adds to the existing body of empirical evidence of quantitative geocode accuracy assessments that can be used to guide data suitability decisions.

The second class seeks to use novel techniques to improve specific aspects of the geocoding process including the matching algorithms (Christen et al. 2005; Christen et al. 2004; Churches et al. 2002) and interpolation algorithms (Bakshi et al. 2004). The works in this category most directly related to that presented herein are the Point Radius Method proposed by Wieczorek et al. (Wieczorek et al. 2004) and the Geocoding Quality Index proposed by Davis et al. (Davis Jr. et al. 2007; Davis Jr. et al. 2003). Both of these techniques attempt to derive a qualitative certainty measure for the output based on characteristics of the input data. The uncertainty-based approach presented herein uses a similar approach to determine the highest quality output possible.

4.3 The Formalization of a Best-Match Criterion

The majority of commercially-available geocoding systems in use today maintain multiple reference data layers. Most commonly, each of these datasets represent a different class of geographic object, each at a different geographic scale (Figure 4.1), or alternatively, multiple representations of the same set of geographic objects created through different means, obtained from different sources, or representative of different periods in time. For example, most freely-available commercial geocoders utilize the U.S. Census Bureau TIGER/Line files (U.S. Census Bureau 2009b) for representing street networks, and Cartographic Boundary files for representing city, minor civil division, ZCTA, county, and state boundaries (U.S. Census Bureau 2009a). Proprietary geocoding platforms typically include county-based parcel boundary files, one or more alternative commercial street network files, and approximations of USPS ZIP code boundaries.

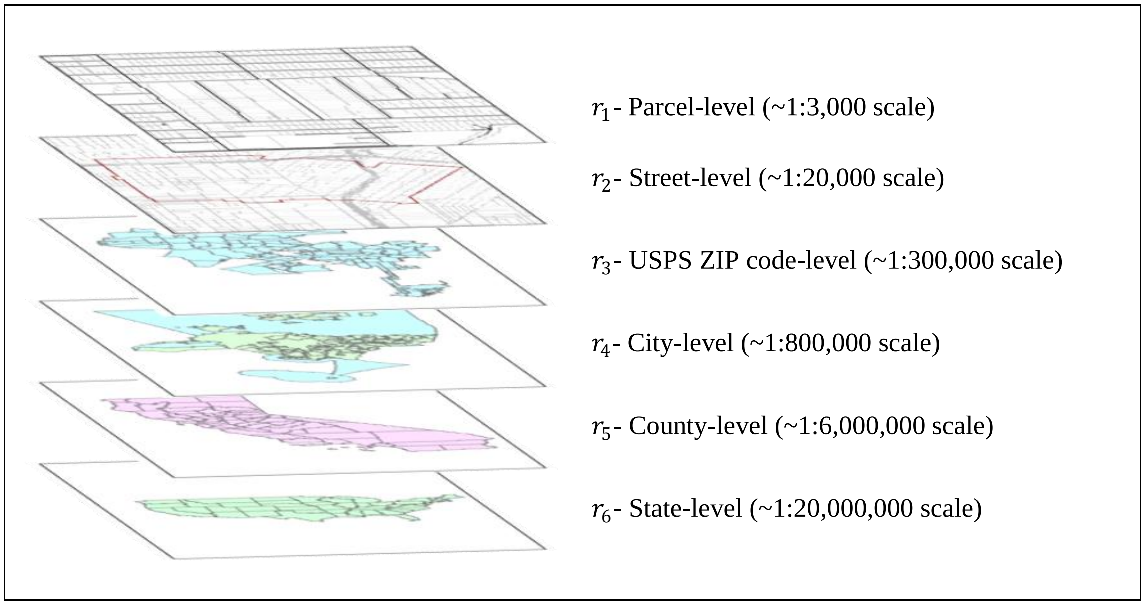

r_1- Parcel-level (~1:3,000 scale)
r_2- Street-level (~1:20,000 scale)
r_3- USPS ZIP code-level (~1:300,000 scale)
r_4- City-level (~1:800,000 scale)
r_5- County-level (~1:6,000,000 scale)
r_6- State-level (~1:20,000,000 scale)

Figure 4.1: Graphical depiction of a six layer reference data source set commonly used in geocoding systems

Given n reference datasets available to a geocoding system, we define $[R]$ to be the set of reference data layers available to the geocoding system, with each $r_i \in [R]$ representing the i^{th} reference data layer in this set, for $i = 1 \dots n$. To service a query, the feature matching process of the geocoder queries each $r_i \in [R]$ to produce zero or more potential reference features, f_{ij}, representing the j^{th} candidate reference feature from the i^{th} reference data layer, for $j = 0 \dots m$ (Equation 4.1). $[F_i]$ represents the set of all j reference features matched in r_i (Equation 4.2).

$$f_{ij} = FeatureMatch(a, r_i) \tag{4.1}$$

$$[F_i] = \bigcup_{j=0}^{m} f_{ij} \tag{4.2}$$

The number of features resulting from the feature matching process using reference layer r_i, $|[F_i]|$, is equal to zero when no reference feature was capable of being found for an input address with a sufficiently acceptable match score. $|[F_i]| = 1$ when a single, non-ambiguous viable candidate feature is produced. $|[F_i]| > 1$ when multiple viable candidate features are matched indicating an ambiguous match between an input address and multiple reference features. Of these three cases, we will only consider the $|[F_i]| = 1$ instance because no action need be taken when $|[F_i]| = 0$, and ambiguous results are typically treated as failures in production geocoding scenarios. Therefore, we assume that each $r_i \in [R]$ provides one candidate reference feature in each $[F_i]$ which is then used by an interpolation algorithm to produce g_i, the candidate output geocode produced by geocoding the input address, a, within the i^{th} reference data layer, r_i (Equation 4.3). $[G]$ then defines the complete set of potential output geocodes (Equation 4.4).

$$g_i = Interpolate(a, [F_i]) \tag{4.3}$$

$$[G] = \bigcup_{i=0}^{n} g_i \tag{4.4}$$

Although each $g_i \in [G]$ could be considered a viable geocode output, end users are usually only interested in the single "best" geocode, which we define as $\dot{g}$. The challenge for a geocoding system is to use some criterion to choose and return $\dot{g}$ from all $g_i \in [G]$. To differentiate between possible methods for making these decisions and facilitate comparisons between their performance, we define $BestMatch([G])$ to be the criterion function used to make this decision (Equation 4.5).

$$\dot{g} = BestMatch([G]) \tag{4.5}$$

In the following subsections, we describe four such best-match criterion functions for selecting $\dot{g}$. In the first, we contextualize the existing state-of-the-art practices used by commercially available geocoding systems into our best-match criterion framework, termed the *hierarchy-based criterion.* In the second, we describe an alternative best-match criterion, termed the *uncertainty-based criterion*, which utilizes the uncertainty associated with each candidate geocode to choose $\dot{g}$. In the third, termed the *gravitationally-based criterion*, we extend the uncertainty-based approach to exploit the uncertainty present in multiple reference features. In the fourth, termed the *topologically-based criterion*, we extend the gravitationally-based approach to utilize the spatial relationships between multiple candidate reference features as well as the uncertainty present in each.

4.3.1 The hierarchy-based approach

The concept of a best-match criterion is directly applicable to the approach taken within currently available commercial geocoding services which use a static ordering to choose $\dot{g}$, e.g., the ESRI Address Locator, Google, Microsoft Bing, and Yahoo! (Environmental Systems Research Institute 2009b; Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b), which we term the *hierarchy-based criterion*. In these systems, the reference data layers are first placed into a qualitative, and in many cases, arbitrary hierarchy. Next, as described by the ESRI documentation and typical of this approach, “The order in which the address locators are listed ... determines the order in which they are used in the geocoding process. The address locator listed first will be used first, and so on.” (Environmental Systems Research Institute 2009c). One such explicit ordering is the North American Association of Central Cancer Registries’ (NAACCR) GIS Coordinate Quality codes (Table 4.1) which is routinely used by the health research community as the de facto standard for geocode quality reporting and the determination of suitability for inclusion in scientific research and practice.

Table 4.1: NAACCR GIS Coordinate Quality Codes

Code	Description
1	GPS
2	Parcel centroid
3	Complete street address
4	Street intersection
5	Mid-point on street segment
6	USPS ZIP5+4 centroid
7	USPS ZIP5+2 centroid
8	Assigned manually
9	USPS ZIP5 centroid
10	USPS ZIP5 centroid of P.O. box or RR
11	City centroid
12	County centroid

This hierarchy-based method is quite simplistic and assumes global relationships of relative accuracy between reference data layers based on qualitative rules-of-thumb associated with the various reference datasets. These schemes give false prudence to the idea that a higher rank geocode is more accurate than a lower rank counterpart, e.g., choosing a USPS ZIP code centroid will always be more accurate than a corresponding city centroid for the same input address. However, reference datasets are often composed of geographic objects defined for use as administrative areas (i.e., cities) or to aid in efficient mail delivery (i.e., USPS ZIP codes), and as such they exhibit spatial heterogeneity from region to region in terms of size, shape, and distribution which invalidates these types of assumptions.

For example, USPS ZIP codes shrink as the number of mail recipients increases, resulting in compact geometries in dense urban areas and large sprawling ones in rural areas. These shapes and areas can vary dramatically within even small regions, as depicted in Figure 4.2 which shows the difference in areas for three USPS ZIP codes in Los Angeles County. Administrative regions often exhibit the inverse behavior, becoming smaller and spread further apart as population density decreases.

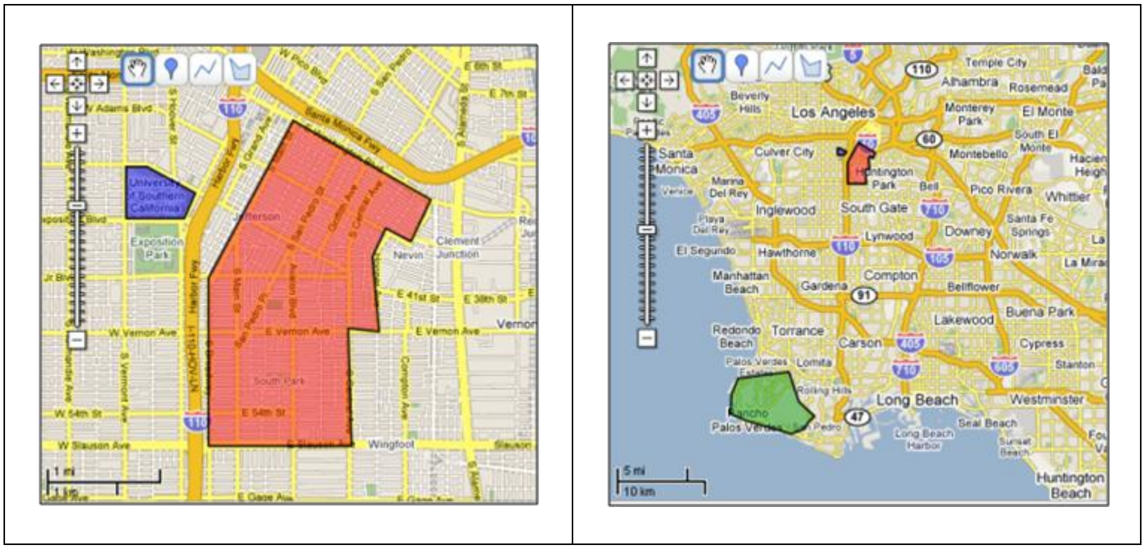


Figure 4.2: USPS ZIP codes 90089 (blue) and 90011 (red) at ~1:10,000 scale (left) and 90089, 90011, and 90275 (green) at ~1:300,000 scale (right)

The presence of these variations within and between reference datasets reveals that a single static hierarchy-based approach to optimal candidate selection will be unable to consistently return the best geocode output, $\dot{g}$, in all circumstances.

This approach also ignores possible local variations in the underlying geographic features within and across reference data sources that could be exploited to return more precise geocodes. Our new approaches developed herein endeavor to make use of this information to identify the best geocodes.

4.3.2 An uncertainty-based approach

The primary motivations behind the first strategy, termed the *uncertainty-based criterion*, are to (1) minimize the maximum uncertainty associated with any geocoded output location; and (2) choose $\dot{g}$ from within $[G]$ based on quantifiable measures of uncertainty associated with each $g_i \in [G]$. To this end, we define u_i to be a quantitative discrete value representing the uncertainty associated with $g_i \in [G]$ as computed by some uncertainty function $u()$ (Equation 4.6).

$$u_i = u(g_i) \tag{4.6}$$

We recognize that, ideally, one would choose a $u()$ function that accounts for all stages of the geocoding process (i.e., input data completeness/correctness, feature matching, feature interpolation, etc.) to derive u_i. However, for our present proof-of-concept implementation, we utilize the intuition that matching an input address to a reference feature with a smaller area and calculating its centroid is more likely to result in a geocode with greater accuracy because there is less uncertainty as to where, within or along the reference feature, the true geocode resides. Therefore, in this research we use the spatial area of the reference feature as a first approximation of spatial uncertainty because it is a useful metric for representing the number of equally probable output locations within a reference feature. This measure is based on local quantitative information directly related to the geocode rather than relying on global rules-of-thumb as in the hierarchy-based approach.

To determine a discrete uncertainty value u_i for any candidate geocode, we must derive a probability representing how likely it is that a geocode would fall at any particular location within a reference feature. This must be normalized across all reference datasets so that uncertainty values from different reference data sources are comparable. To do so we can overlay a grid, Q, of cells with a width and height of q on top of r_i, the reference feature responsible for producing a particular $g_i \in [G]$, as depicted in Figure 4.3. For simplicity of notation, we define β to be the resulting number of cells in the grid in r_i (Equation 4.7).

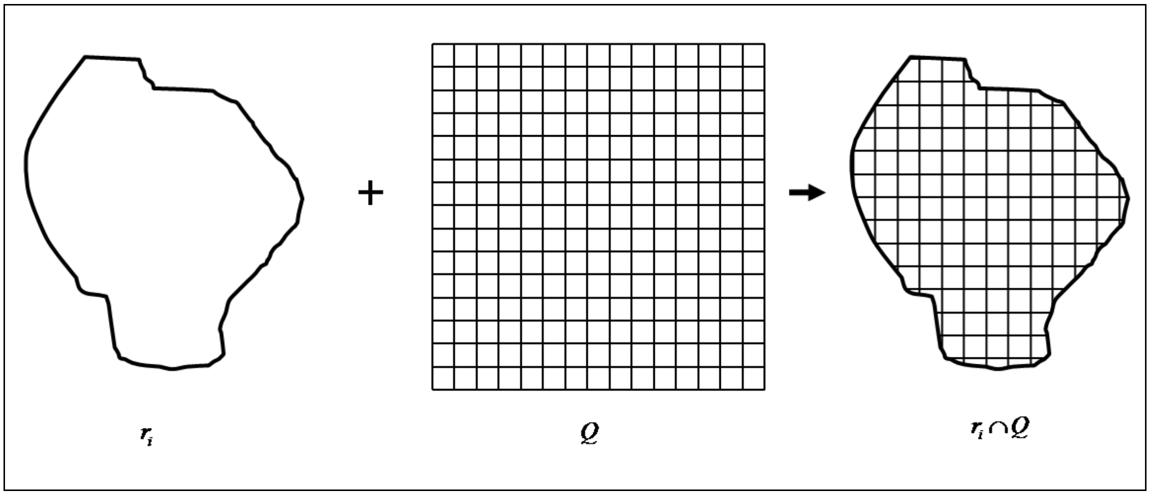


Figure 4.3: Creation of an uncertainty surface using a uniform grid overlay

$$\beta = \left(\frac{Area(q)}{Area(r_i)}\right)^2 \tag{4.7}$$

We next define $p_i = p(g_i)$ to be the probability that an output location is correct at any location within r_i. Because every location in r_i is an equi-probable location for a geocode, p_i is equal to the probability of choosing a cell in r_i at random (Equation 4.8). Therefore, the uncertainty associated with a candidate geocode g_i, $u_i = u(g_i)$, becomes the likelihood that the true geocode should have been placed in any other location within the reference feature (Equation 4.9).

$$p_i = p(g_i) = \frac{1}{\beta} \tag{4.8}$$

$$u_i = 1 - p_i \tag{4.9}$$

The β, p_i and u_i values for three idealized reference features are displayed in Figure 4.4 which shows that p_i is uniformly distributed and that $u_i \propto Area(r_i)$, which meets our intuition that larger reference features should result in geocodes with less certainty. We should note that for areal unit-based reference features such as cities and USPS ZIP code boundaries, this approach can be applied directly. For linear- and point-based features however, some technique must to be applied to determine an appropriate areal unit-based approximation of the feature for use calculating u_i. In the case of linear-based features such as street segments, this can be done by buffering the feature with the size of the dropback typically applied (10 m).

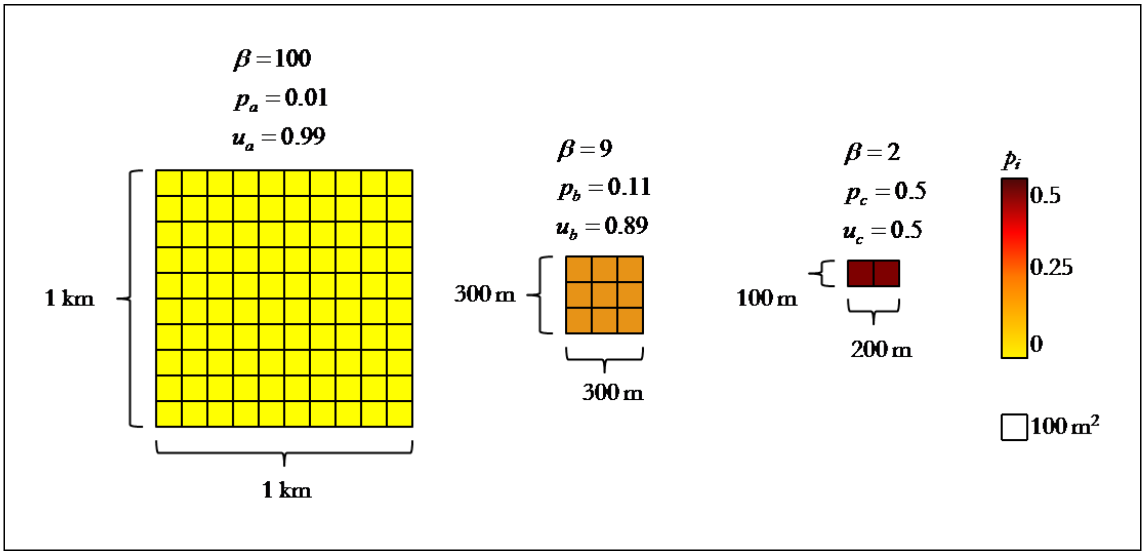


Figure 4.4: Uncertainty assignment for three idealized reference features

We see this as a reasonable approach because once the correct street segment has been selected, the actual location of the true geocode could in fact be anywhere along either side of the street, accounting for the fact that address ranges and parities are often misreported in street segment reference data files. In the case of point-based features, a similar buffering approach can be taken, but the size of the buffer should be chosen to reflect the type of reference feature, i.e., address points should be buffered with a smaller distance than USPS ZIP code centroids or even state centroids.

To implement the selection process used in this method, the reference data sources are first sorted from low to high based on their index from the hierarchy-based approach. Next, u_i is calculated for each of the g_i candidate geocodes. Finally, the candidate geocode with the minimum g_i value is chosen and returned as output. If two candidate geocodes both have the same area, this method defaults to choosing the one with the lower position in the hierarchy-based approach.

4.3.3 A gravitationally-based approach

While the previous method uses quantitative metrics drawn from the spatial characteristics of the reference features to determine the best output, it is naïve in that the returned geocode only uses information from a single reference feature, effectively ignoring any information available from any other reference feature. While it makes intuitive sense to place the output geocode within (or near to) the most specific of the reference features available, i.e., that with the least uncertainty where uncertainty is expressed in terms of the area of the feature as in the previous method, exactly where within or nearby this feature is unknown because by our definition of u_i, the uncertainty associated with any random location within the reference feature r_i is as likely as any other location within r_i.

Instead of simply picking a random location or the location that minimizes the maximum error (i.e., the centroid), we can use the candidates selected from within each of the other reference layers to help guide the selection of the output location. To accomplish this, we can extend the previous method to express each $g_i \in [G]$ as single points having mass m_i proportional to its probability of correctness p_i. This means that smaller area reference features will have larger masses, e.g., a USPS ZIP code feature will have a far higher mass than that of the state feature to which it belongs. Because the uncertainty for all reference features is determined using the same grid, the mass will scale proportionally with the inverse of the area of the feature.

Conceptually, these masses can be thought of as rings around the centroids of each feature having a radius proportional to the mass, although we will consider them as points during the calculation of the ultimate output geocode location g_i. The idealized reference features r_a and r_b from Figure 4.4 are displayed in Figure 4.5 which shows red dashed circles (m_a, m_b) representing the masses around the red dots (c_a, c_b) representing the centroids of each. In this model the location of the output geocode g_i can be located using a standard gravitational center of mass calculation for any number of objects (Stewart 1995, pp 506). This approach requires that we define the moment of the system about the x- and y-axis as in Equation 4.10, from which we can determine the coordinates of the centroid $g_i = (\bar{x}, \bar{y})$ as in Equation 4.11 where $m = \sum m_i$ is the total mass of the system.

$$M_y = \sum_{i=1}^{n} m_i x_i \, , M_x = \sum_{i=1}^{n} m_i y_i \tag{4.10}$$

$$\bar{x} = \frac{M_y}{m}, \bar{y} = \frac{M_x}{m} \tag{4.11}$$

This approach has the benefit that it takes into account the placement and uncertainty associated with the alternative candidate geocodes from other reference data layers rather than just the one selected as having the minimum uncertainty due to it having the smallest area. Therefore, the other features act as gravitational pulls, moving the output geocode toward their own center of mass with a force proportional to the inverse of their uncertainty and the distance between them.

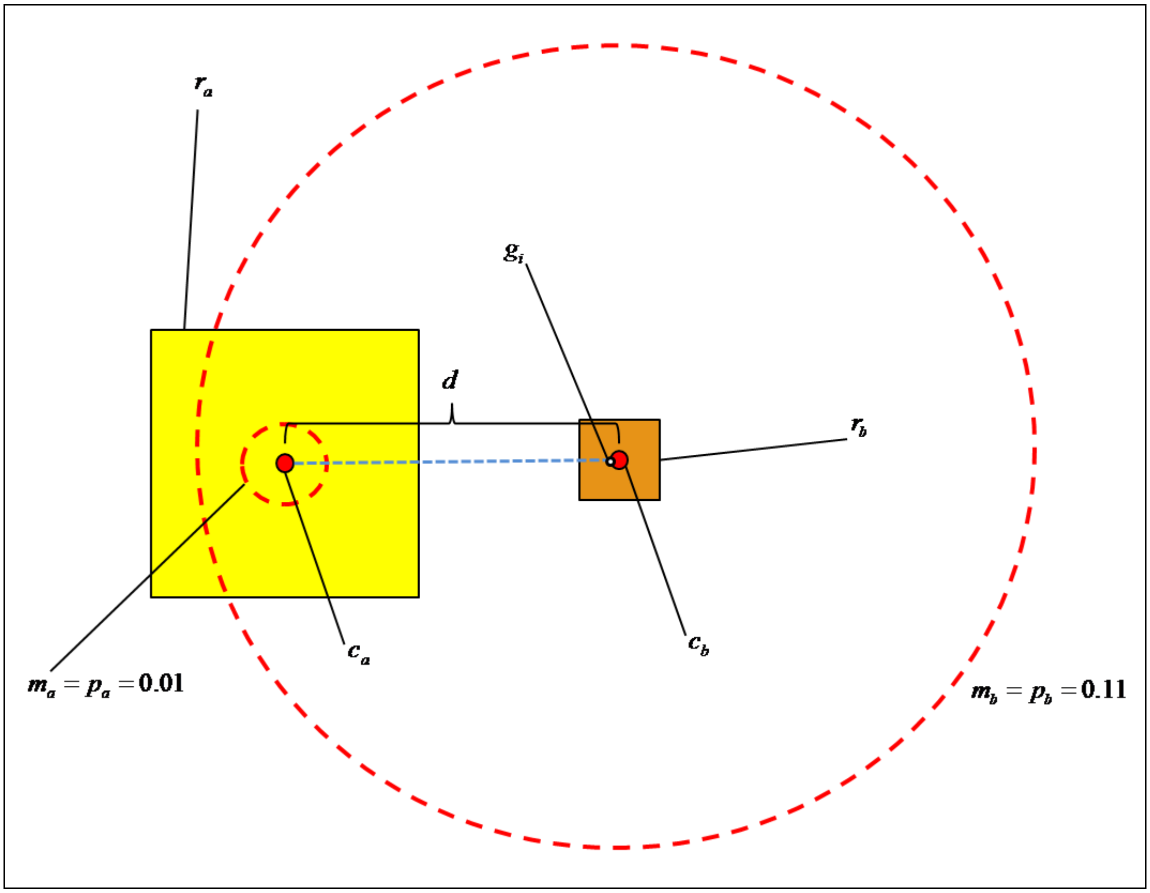


Figure 4.5: The probability of two reference features as areas (dashed red lines) along with their centroids (red dots) and the gravitational center of mass between the two (small white dot)

Because the mass value associated with each reference feature is proportional to the inverse of its area, very large area reference features will have little gravitational pull on the smaller, i.e., more specific, reference features. This is a desirable characteristic because it would not be beneficial for the centroid of the state feature to have a large effect on where the placement of the USPS ZIP code centroid is located.

However, this approach is not without its weaknesses. Consider the case when the smaller of the two reference features is topologically contained within the other. In this situation, it is possible that the centroid of the larger will pull the center of mass of the two away from the true output location if it is located on the opposite side of the smaller reference feature. An example of this is displayed in Figure 4.6 which depicts the center of mass of the large feature (c_a) drawing the output geocode (g_i) away from the true output location ($\bar{g}_i$). From this figure, we see that if the true output location was located anywhere in the shaded region of the smaller reference feature (r_b), the gravitational model would produce a result more erroneous than just using the centroid of the smaller feature (c_b) alone.

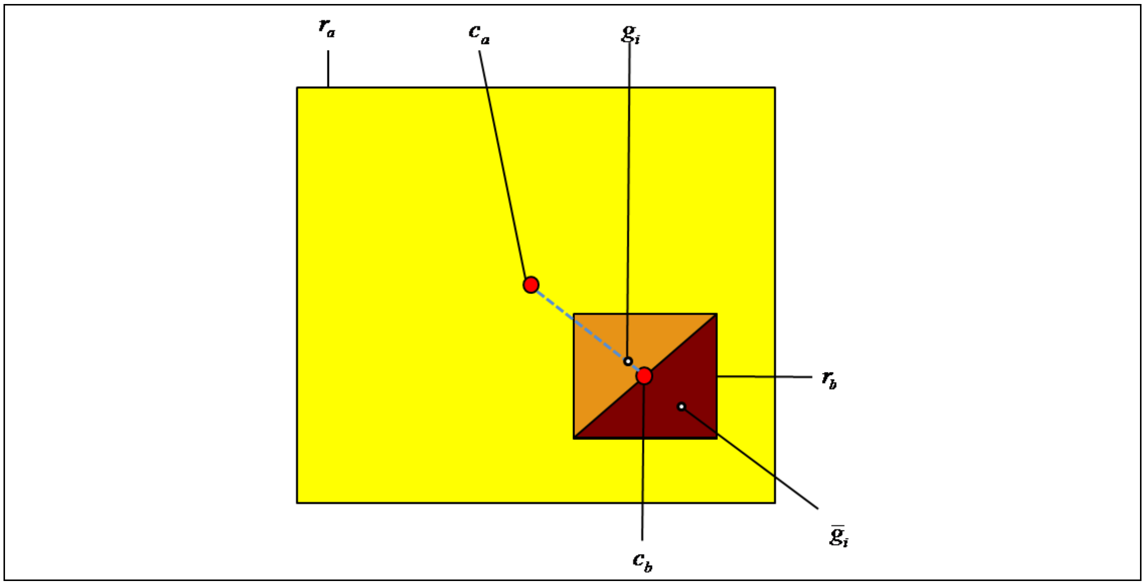


Figure 4.6: Gravitational model applied to two non-equivalently-sized reference features resulting in an output location ($\boldsymbol{g_i}$) pulled away from the true output location ($\bar{\boldsymbol{g}}_i$) because the smaller reference feature is contained within the larger

4.3.4 A topologically-based approach

The gravitationally-based method just presented may improve the result over the uncertainty-based method in some but not all instances because it considers multiple reference features when determining the final output location. Some of the shortcomings previously identified stem from to the fact that it does not consider the spatial relationships between each of the $g_i \in [G]$. Because each of the candidate geocodes and the reference features from which they are drawn are spatial in nature, they are implicitly related to each other by the eight topological predicates derived from the 9-intersection model (Egenhofer 1991). These relationships can be used to drawn spatial inferences given the uncertainties associated with each reference feature which in turn can be used to guide the estimation of the most probable output location.

We will only consider four of the eight relations in the discussion that follows – "disjoint" (Figure 4.7a), "touches" (Figure 4.7b), "overlaps" (Figure 4.7c), and "contains" (Figure 4.7d). The "crosses" operator is excluded because we will force the two reference features under consideration to always be of the same dimension (polygon) as described earlier when linear- and point-based features are buffered. The "intersects" operator is excluded because our treatment of "overlap" will handle it. The "within" operator will be excluded because we will always assign the smaller of the two objects to be the second, thereby eliminating its occurrence in preference to the "contains" operator. Finally, the "equals" operator will be excluded because it is a trivial case in that the output from both reference features results in the same location which would be returned and warrants no special consideration.

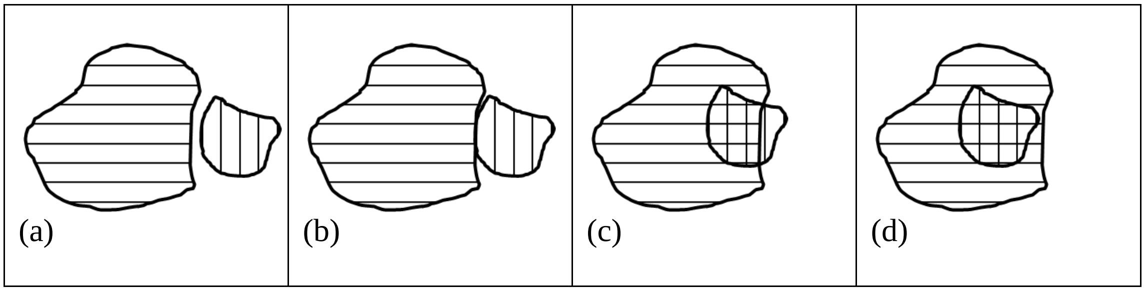


Figure 4.7: Topological relationships between two objects: (a) disjoint; (b) touch; (c) overlap; and (d) contain

When considering each reference feature independently using only the spatial area as the selection criterion (i.e., the uncertainty-based method), inter-feature relationships are not utilized. When considering only the masses of the features (i.e., the gravitationally-based method) inter-feature relationships are utilized, but the resulting output location may actually be worse than just using an uncertainty-based approach alone. To minimize this risk, we can augment the gravitationally-based method to utilize the topological relationships between two reference features. In particular, we can utilize the presence or absence of an overlapping region between the two reference features to derive a better probability estimation which we can then use to compute an output location with the minimal amount of possible uncertainty. If an overlapping region does not exist, it means that the one of the two reference features is incorrect because the true output location can only be located within one or the other. In these cases we will snap the output location to the point on the boundary of the reference feature with less uncertainty (i.e., the smaller of the two) that is the closest to the other reference feature because this represents the best approximation given the conflicting information at hand.

However, if an overlapping region does exist, this area provides an additional basis of support, narrowing the potential candidate output locations within the larger reference feature to just the area where we have reason to believe the output geocode is likely to reside, i.e., within the smaller of the reference features. More specifically, if two reference features overlap, some portion of the same area will be present in both reference features because the area is a discrete space that exists only once. This overlapping region must have two uncertainly values associated with it, one from each of the reference features. These two uncertainties are independent; the uncertainty associated with one does not depend on the uncertainty of the other. Placing the output anywhere within the smaller feature does not conflict with placing the output geocode location at any other location within the larger feature because all locations with the boundary of the larger are equally likely, including those within the smaller feature. In effect, the smaller feature provides the boundary of a spatial filter that creates the smallest possible area within which we have the strongest evidence that the true output geocode most likely resides. Therefore, the potential uncertainty associated with the output of choosing any location within the overlapping region must be smaller than the uncertainty of choosing any location within either of the reference features independently.

To accomplish this, we define r_c to be the area resulting from the intersection of two reference features r_a and r_b of equal degree, i.e., $r_a^\circ = r_b^\circ$, as in Equation 4.12. As in the uncertainty-based approach, we first overlay grid Q and then assign a p_i to each of the cells in r_c according to Equation 4.8. The two idealized reference features from our previous figures are shown intersecting to produce r_c in Figure 4.8.

$$r_c = r_a \cap r_b \tag{4.12}$$

The region r_c represents the area where the uncertainty associated with an output geocode should be minimized because it is the location that is independently confirmed by two reference features. Therefore, our goal is to calculate a weighted mass for the centroid of r_c which captures ω, the degree of topological agreement between features r_a and r_b, while also accounting for the independent degree of belief provided by each.

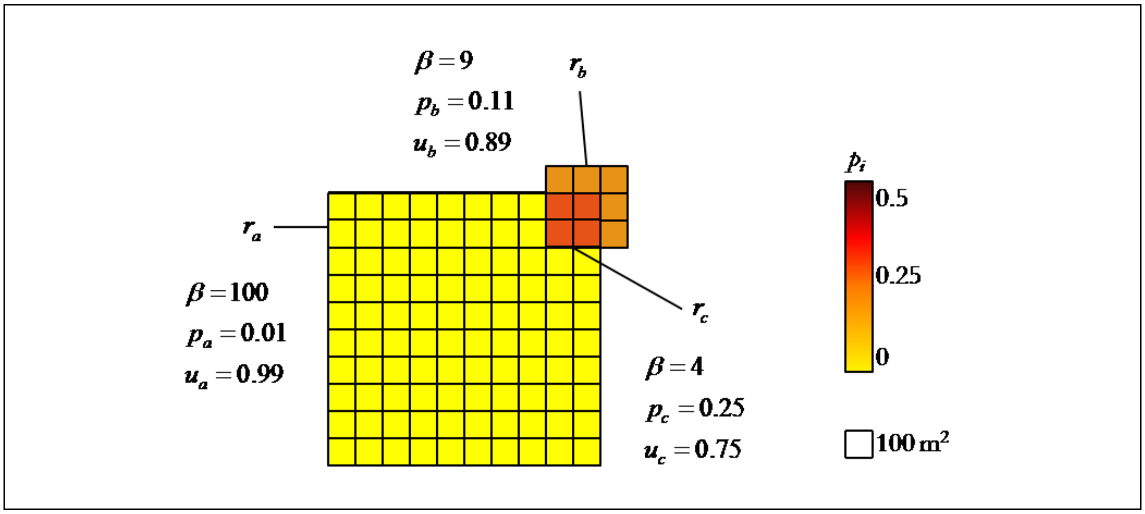


Figure 4.8: The intersection of two idealized reference features resulting in an overlap area ($\boldsymbol{r_c}$) whose grid cells are assigned probability values ($\boldsymbol{p_c}$) using the uncertainty-based approach

To express ω, we first define r_u to be the union of the two reference features r_a and r_b (Equation 4.13). Having done so, we then create a probability surface over r_u with the same grid Q used for r_a and r_b. This surface represents $p_u = p(r_u)$, the probability that placing an output location anywhere at random within the union of the reference features known to the system is correct as demonstrated for our idealized reference features in Figure 4.9. Using raster algebra, we can compute the non-uniformly distributed probability surface for the joint surface probabilities for reference features $p(r_a \cap r_b)$ as the sum of the two independent probabilities normalized by the probability of the system as a whole p_u (Equation 4.14).

$$r_u = r_a \cup r_b \cup r_c \tag{4.13}$$

$$p(r_a \cap r_b) \;=\; p(r_a) + \; p(r_b) - \; p(r_u) \qquad (4.14)$$

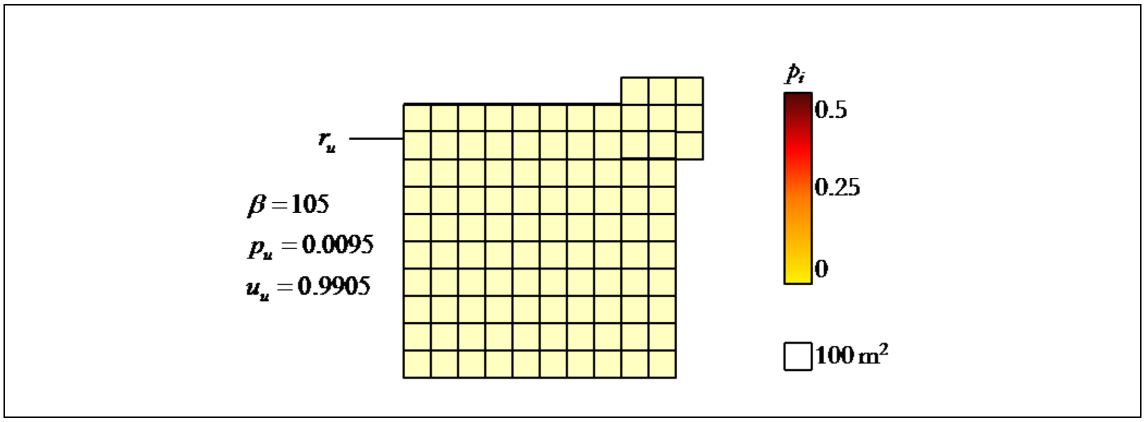


Figure 4.9: The probability surface for the union of two overlapping idealized reference features

This process is displayed in Figure 4.10 which illustrates that r_c, the area representing the overlap between r_a and r_b, is assigned a higher probability than anywhere within either reference feature independently, whereas the area not contained in r_c is assigned a probability lower than any corresponding location within r_a and r_b.

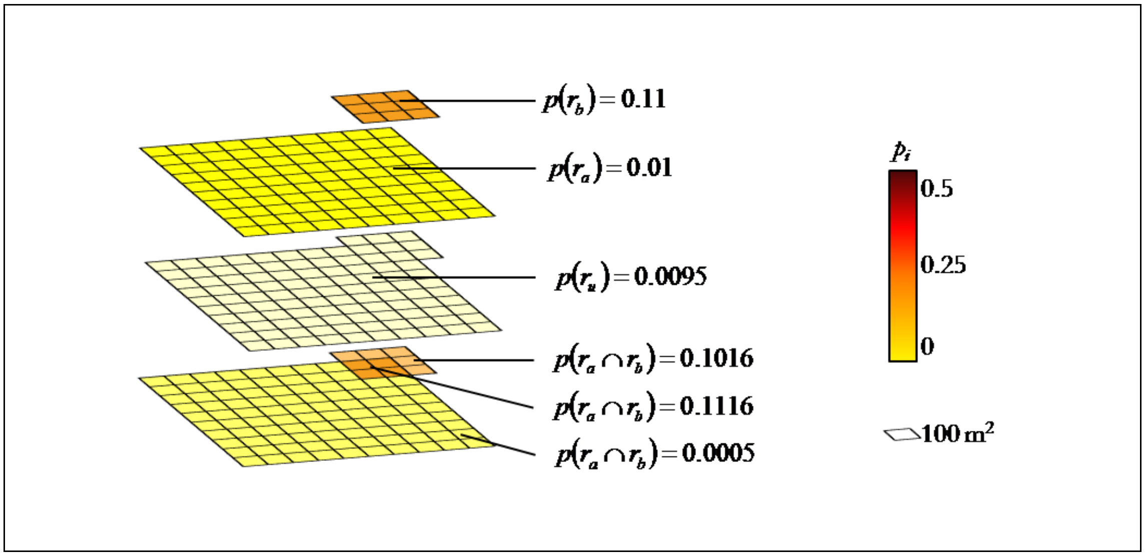

$p(r_b) = 0.11$
$p(r_a) = 0.01$
$p(r_u) = 0.0095$
$p(r_a \cap r_b) = 0.1016$
$p(r_a \cap r_b) = 0.1116$
$p(r_a \cap r_b) = 0.0005$
p_i
0.5
0.25
0
100 m²

Figure 4.10: Computation of the joint probability surface resulting from two independent reference features which overlap

Finally, we define ω, as the ratio of the area of r_c to the area of the whole system (Equation 4.15). Note that had we assigned ω using the sum of both r_a and r_b in the denominator, the area in r_c would have been double counted. Using ω, we can then compute m_c, the weighted mass value for the centroid of r_c that represents the amount of overlapping area of two reference features (Equation 4.16).

$$\omega = \left(\frac{Area(r_c)}{Area(r_u)}\right) \tag{4.15}$$

$$m_c = \omega r_c \tag{4.16}$$

The centroid of r_c along with its mass m_c can then be included as an additional object in our gravitationally-based method to compute the location within the system as a whole that has the highest likelihood of being correct based on the reference features available to the algorithm and the spatial relationships between them. While still not able to completely eliminate the potential for incorrect placement due to the gravitational pull of the larger reference feature as shown in Figure 4.6, this effect should be lessened by the addition of the proportionally weighted centroid r_c into the system.

4.4 Chapter 4 Experimental Evaluation

Our evaluation will seek to answer three specific questions. First, does the spatial accuracy of geocodes improve if we simply reverse the order of layers in a hierarchy-based approach? Second, does accuracy improve when utilizing an uncertainty-based based approach instead of any type of hierarchy-based approach? Finally, what level of spatial improvement is possible when using either the gravitationally- and/or topologically-based approach over the uncertainty-based approach?

To answer these questions, we performed an analysis of the spatial error resulting from each method in terms of distance to the true location. To do so, we first geocoded two national datasets for which ground truth locations were available using the hierarchy-, reversed hierarchy-, and uncertainty-based methods and calculated the distance from truth for each method to determine the level of spatial improvement. Using the results of these runs, we determined the applicable records for the gravitationally- and topologically-based methods based on whether or not a candidate geocode could be found in more than one reference data layer, a requirement for using these approaches as they both depend on at least two reference features. These records were then processed using all approaches.

Thus, each test dataset was processed 10 times; first using U.S. Census Bureau TIGER/Lines, ZCTA, and Place reference data sources with each of the five best-match criteria, and next with only the ZCTA and Place files along with each of the five methods. The first set of five runs should provide information about the improvements possible in a typical geocoding scenario where a high percentage of addresses geocode to the street-address level, leaving a small portion geocoded to USPS ZIP code- or city-level which will be applicable to the different best-match methods. The second set should shed light on what would happen in the case of an extremely low street-address level match rate where the majority of records geocode to either USPS ZIP code- or street address-level accuracy.

4.4.1 Data sources and methods

The two address dataset that we used were a listing of 2,093 Best Western hotels and 1,649 Target stores in the U.S. coming from all 50 U.S. states as well as Puerto Rico. These datasets were obtained from a ‘point of interest’ (POI) website (POIfriend Inc. 2009), and are intended to be imported into a GPS unit for navigation to these locations. According to the source, these lists contains the “official” locations of Best Western and Target locations obtained through GPS readings produced by each company as well as the full postal address. Because these companies have a vested interest in making sure that people can find these locations, we can assume that these locations are described at a fairly high level of accuracy. Additionally, 10% of the datasets were manually reviewed using a previously developed interface (Goldberg, Wilson, Knoblock et al. 2008), and in all cases was found to be positioned within 10 m of one or more of the buildings. These datasets are currently available for free and can be obtained by any researcher who wishes to extend or replicate our work.

The geocoder used for these experiments was built by the University of Southern California GIS Research Laboratory as a research platform for implementing and testing novel geocoding techniques and data sources. The version used for this research (2.94) is available online at https://webgis.usc.edu, for which the implementation details and a performance evaluation against several commercial geocoding systems including ESRI, Google, Microsoft Bing, and Yahoo! (Environmental Systems Research Institute 2009b; Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) can be found in our prior work (Goldberg et al. 2009; Swift et al. 2008) as well as **Chapter 3**.

For our experiments in this chapter, the reference data used include the freely available 2008 versions of the U.S. Census Bureau TIGER/Lines (U.S. Census Bureau 2009b), Places, ZCTA5 files (U.S. Census Bureau 2009a). The settings used are the same as those listed in the experimental evaluation in **Chapter 3**.

The hierarchy-based approach was implemented according to the NAACCR standard for all applicable reference datasets (Hofferkamp et al. 2008). The reversed hierarchy-based approach was also based on the NAACCR standard, except the USPS ZIP code and city priorities were reversed. For the uncertainty-, gravitationally-, and topologically-based approaches, the OCG STArea() and STIntersection() function implementations provided by the Microsoft spatial data types were used to determine the area of uncertainty for each reference feature and any overlapping regions between features, respectively (Microsoft Corporation 2009b). A 10 m Q grid was used to determine the number of grid cells within each reference feature.

4.4.2 Results

Overall, the USC geocoder used in our experiments had a match rate of 99.9%. Two Best Western and one Target address were unmatchable: "Routes 340 & 772 9 Queen Rd, Intercourse, Pa, 17534", "Route 322 & Us Route 1, Concordville, Pa, 19331", and "P.O. box 985, Oaks, Pa, 19456". These addresses are intersection- and Post Office Box-types which are not currently supported for street-level geocoding in the USC geocoder. Upon investigation it was found that neither the city names nor the USPS ZIP codes associated with these records could be located in the U.S. Census Bureau TIGER/Lines, Place, or ZCTA files. These records were excluded from further analysis.

Consistent with other research reports (Cayo et al. 2003; Zandbergen 2008a, 2008b; Zimmerman et al. 2010; Zimmerman 2008; Zimmerman et al. 2007), we define the spatial accuracy of a geocode as being its distance to a ground truth value representing the real location that should have been returned, i.e., a geocode that is closer to the ground truth location is more accurate than one which is further away. That said, when using the U.S. Census Bureau TIGER/Lines reference source to geocode the records, 85.6% and 83.9% of the Best Western and Target input data were successfully geocoded to the street-level accuracy with average spatial errors of 0.484 and 0.477 km, respectively. The remaining 14.4% and 16.1% of the data resulted in city and/or USPS ZIP code-level matches, which came from 47 of the 50 U.S. states plus Puerto Rico across both input datasets.

The spatial errors resulting from each of the five best-match methods for each dataset independently as well as combined are displayed in Tables 4.2 and 4.3. The difference between the two tables is that Table 4.2 includes all records from both input datasets, while Table 4.3 only includes the records for which both a ZCTA and Place feature could be located for a record, i.e., just the records that have the potential to change between the different best-match methods as they require two or more reference features. The results shown in Table 4.2 indicate the level of impact one could expect from each best-match method when run over each of the datasets as a whole, whereas those in Table 4.3 reveal the level of impact on just the affected records.

Table 4.2: Spatial error by best match method across all records in sample datasets with (a) TIGER/Lines, ZCTA, and Place reference data layers; and (b) just ZCTA and Place reference data layers

Sample Dataset	Layers	n % of total	Hierarchy min, max, mean, std (km)	Hierarchy Reversed min, max, mean, std (km)	Uncertainty min, max, mean, std (km)	Gravitational min, max, mean, std (km)	Topological min, max, mean, std (km)
Best Westerns	(a)	2091 99.9%	0.001, 49.11, 1.428, 3.884	0.001, 39.552, 0.975, 2.48	0.001, 39.552, 0.872, 2.15	0.001, 39.552, 0.869, 2.141	0.001, 39.552, 0.866, 2.137
Target Stores	(a)	1648 99.9%	0.029, 29.441, 1.1, 2.228	0.029, 30.505, 1.215, 2.684	0.029, 19.164, 1.001, 1.87	0.029, 19.164, 0.958, 1.751	0.029, 19.164, 0.951, 1.74
Combined	(a)	3739 99.9%	0.001, 49.11, 1.283, 3.263	0.001, 39.552, 1.081, 2.574	0.001, 39.552, 0.929, 2.032	0.001, 39.552, 0.908, 1.979	0.001, 39.552, 0.904, 1.972
Best Westerns	(b)	2091 99.9%	0.01, 69.574, 5.825, 6.245	0.077, 39.552, 4.039, 4.336	0.01, 39.552, 2.951, 2.946	0.049, 39.552, 2.858, 2.858	0.028, 39.552, 2.812, 2.813
Target Stores	(b)	1648 99.9%	0.037, 29.441, 3.657, 3.037	0.085, 33.54, 4.897, 4.4	0.037, 27.055, 3.098, 2.351	0.065, 27.055, 2.867, 2.024	0.034, 27.055, 2.833, 2.074
Combined	(b)	3739 99.9%	0.01, 69.574, 4.87, 5.2	0.077, 39.552, 4.417, 4.385	0.01, 39.552, 3.016, 2.701	0.049, 39.552, 2.862, 2.525	0.028, 39.552, 2.821, 2.514

The number of records listed in Table 4.2 is equivalent to the number of records matchable in each input dataset, while the number in Table 4.3 is the amount which resulted in either a ZCTA or Place match, where both were possible. Both tables display the spatial error that results when the U.S. Census Bureau TIGER/Line, Place, and ZCTA files are used versus when only using the Place and ZCTA files.

Table 4.3: Spatial error by best match method across records with ZCTA and Place matches in sample datasets with (a) TIGER/Lines, ZCTA, and Place reference data layers; and (b) just ZCTA and Place reference data layers

Sample Dataset	Layers	n % of total	Hierarchy min, max, mean, std (km)	Hierarchy Reversed min, max, mean, std (km)	Uncertainty min, max, mean, std (km)	Gravitational min, max, mean, std (km)	Topological min, max, mean, std (km)
Best Westerns	(a)	255 12.2%	0.32, 49.11, 7.445, 7.537	0.077, 31.757, 3.737, 3.999	0.077, 19.984, 2.888, 2.731	0.067, 19.99, 2.864, 2.694	0.028, 19.987, 2.842, 2.681
Target Stores	(a)	222 13.5%	0.127, 29.441, 4.422, 3.509	0.359, 30.505, 5.274, 4.763	0.359, 14.976, 3.685, 2.415	0.32, 10.927, 3.367, 2.06	0.293, 11.191, 3.319, 2.048
Combined	(a)	477 12.8%	0.127, 49.11, 6.038, 6.19	0.077, 31.757, 4.452, 4.433	0.077, 19.984, 3.259, 2.617	0.067, 19.99, 3.098, 2.43	0.028, 19.987, 3.064, 2.417
Best Westerns	(b)	1881 89.9%	0.01, 69.574, 5.876, 6.317	0.077, 37.484, 3.89, 4.154	0.01, 28.312, 2.681, 2.343	0.049, 28.208, 2.577, 2.206	0.028, 28.26, 2.526, 2.134
Target Stores	(b)	1448 87.8%	0.037, 29.441, 3.693, 3.019	0.085, 33.54, 5.102, 4.505	0.037, 24.876, 3.057, 2.216	0.065, 17.222, 2.796, 1.804	0.034, 24.876, 2.756, 1.865
Combined	(b)	3329 89%	0.01, 69.574, 4.927, 5.26	0.077, 37.484, 4.418, 4.351	0.01, 28.312, 2.845, 2.296	0.049, 28.208, 2.672, 2.043	0.028, 28.26, 2.626, 2.024

The spatial error for applicable records broken down by topological relationship are shown when using U.S. Census Bureau TIGER/Lines, ZCTA, and Place reference files (Table 4.4) as well as when just using ZCTA and place (Table 4.5).

Table 4.4: Spatial error by best match method and topological relation across 3,329 records in combined dataset resulting from geocoding with ZCTA and Place reference files

Relation	n % of total	Hierarchy min, max, mean, std (km)	Hierarchy Reversed min, max, mean, std (km)	Uncertainty min, max, mean, std (km)	Gravitational min, max, mean, std (km)	Topological min, max, mean, std (km)
Contains	852 25.6%	0.01, 69.574, 5.224, 6.235	0.077, 6.144, 3.912, 4.543	0.01, 28.312, 2.189, 2.137	0.067, 28.208, 2.178, 2.121	0.028, 28.26, 2.18, 2.127
Disjoint	54 1.6%	0.348, 63.947, 4.831, 8.728	0.971, 7.484, 13.492, 6.944	0.411, 24.876, 5.351, 4.711	0.577, 20.262, 4.871, 3.861	0.411, 24.876, 5.351, 4.711
Overlaps	2,417 72.6%	0.053, 56.442, 4.828, 4.765	0.085, 33.54, 4.378, 3.961	0.09, 19.235, 3.014, 2.194	0.049, 19.165, 2.791, 1.897	0.034, 19.085, 2.716, 1.812
Touches	6 0.2%	0.48, 4.738, 3.235, 1.567	5.007, 7.132, 10.363, 4.52	2.422, 10.982, 5.031, 3.041	1.842, 9.506, 5.084, 2.95	2.079, 9.154, 5.384, 2.912

The magnitude of spatial error was non-normally distributed for each of the best-match methods. Consistent with the analysis of spatial error presented in similar studies (Zimmerman et al. 2010; Zandbergen 2008b), a log transformation was used to normalize the distribution of error. Figure 4.11 shows (a) the original, and (b) log-transformed error distributions for the applicable records when using U.S. Census Bureau TIGER/Lines, ZCTA, and Place reference layers, which was typical of all data resulting from all methods. Student's t-test was then performed on the transformed error values to determine the statistical significance of the improvements resulting from each of the methods (Zimmerman et al. 2010; Zandbergen 2008b).

Table 4.5: Spatial error by best match method and topological relation across 477 records in combined dataset resulting from geocoding with TIGER/Lines, ZCTA, and Place reference files

Relation	n % of total	Hierarchy	Hierarchy Reversed	Uncertainty	Gravitational	Topological
		min, max, mean, std (km)	min, max, mean, std (km)	min, max, mean, std (km)	min, max, mean, std (km)	min, max, mean, std (km)
Contains	128 26.8%	0.333, 43.584, 7.358, 7.79	0.077, 27.796, 3.725, 4.448	0.077, 19.984, 2.747, 3.06	0.067, 19.99, 2.748, 3.052	0.028, 19.987, 2.745, 3.055
Disjoint	3 0.6%	4.063, 9.278, 7.03, 2.681	5.11, 11.191, 8.128, 3.041	5.11, 11.191, 8.128, 3.041	2.592, 10.867, 5.707, 4.5	5.11, 11.191, 8.128, 3.041
Overlaps	344 72.1%	0.127, 49.11, 5.552, 5.451	0.207, 31.757, 4.647, 4.381	0.207, 16.135, 3.386, 2.342	0.221, 16.08, 3.184, 2.096	0.207, 16.192, 3.114, 2.052
Touches	2 0.4%	2.422, 4.738, 3.58, 1.637	10.982, 3.139, 12.061, 1.525	2.422, 10.982, 6.702, 6.053	4.085, 9.506, 6.795, 3.833	5.057, 9.154, 7.105, 2.897

The mean spatial error reduction observed when choosing one matching strategy over another is shown in Table 4.6 along with the number of records and p value resulting from Student's one-tailed t-test with $\alpha = 0.05$. The data in this table are organized by reference data sources used and portion of data analyzed. In particular, the "a" cases are the results when the U.S. Census Bureau TIGER/Lines, ZCTA, and Place reference files were used, while the "b" cases are results when only the ZCTA and Place files were used.

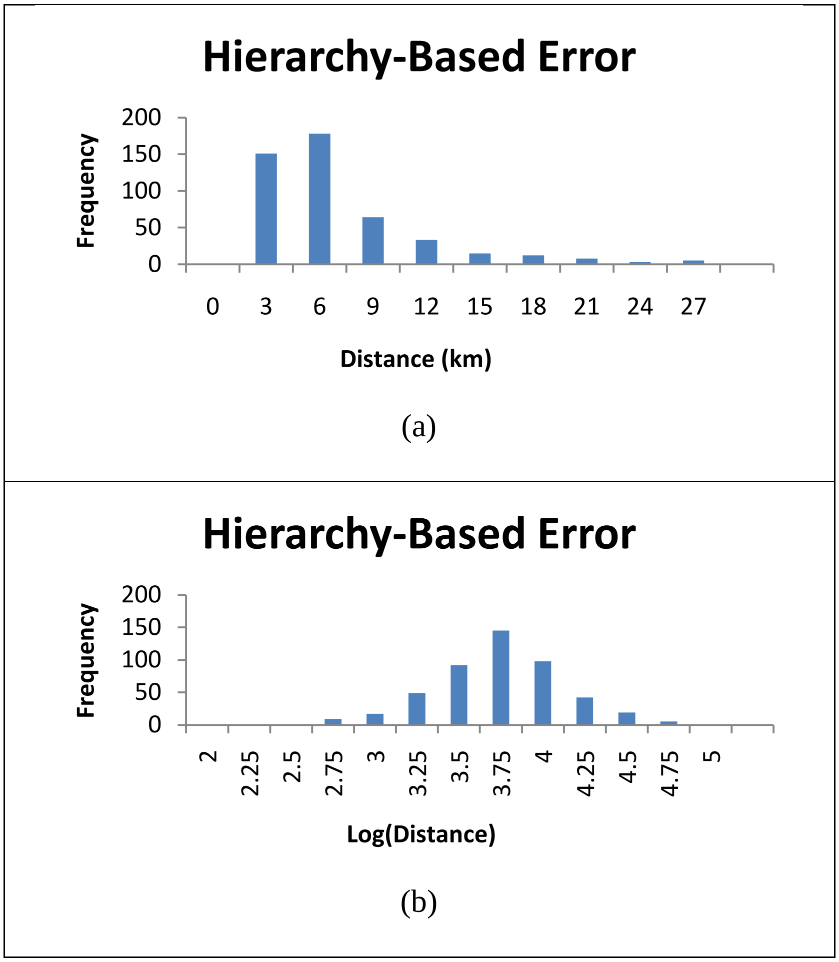


Figure 4.11: Comparison of (a) spatial error distribution; and (b) log of spatial error distribution for applicable records using TIGER/Lines, ZCTA, and Place reference data layers

Similarly, the “1” cases are the results over the entire combined input dataset, the “2” cases are a subset of records for whom both a USPS ZIP code- and city-level geocode was obtainable (those where all methods are applicable), and the “3” cases are the subset of “2” that topologically overlap.

Table 4.6: Mean error reduction observed between best-match methods using (a) TIGER/Lines, ZCTA, and Place and; (b) ZCTA and Place across all combined records for (1) all combined records; (2) only combined records with a ZCTA and Place; and (3) only combined records with a ZCTA and Place that overlap

	Hierarchy reduction (km), n, p	Hierarchy Reversed reduction (km), n, p	Uncertainty reduction (km), n, p	Gravitational reduction (km), n, p
Hi. Rv.	a1: 0.017, 3,739, 0.132 a2: 0.133, 477 ,<0.001 a3: 0.069, 343, 0.007 b1: 0.036, 3,739, <0.001 b2: 0.006, 3,329, 0.213 b3: 0.029, 2,417, 0.003			
Un.	a1: 0.029, 3,739, 0.03 a2: 0.222, 477, <0.001 a3: 0.163, 343, <0.001 b1: 0.167, 3,739, <0.001 b2: 0.187, 3,329, <0.001 b3: 0.155, 2,417, <0.001	a1: 0.011, 3,739, 0.222 a2: 0.089, 477, <0.001 a3: 0.094, 343, <0.001 b1: 0.13, 3,739, <0.001 b2: 0.146, 3,329, <0.001 b3: 0.126, 2,417, <0.001		
Gr.	a1: 0.031, 3,739, 0.022 a2: 0.239, 477, <0.001 a3: 0.184, 344, <0.001 b1: 0.183, 3,739, <0.001 b2: 0.205, 3,329, <0.001 b3: 0.18, 2,417, <0.001	a1: 0.013, 3,739, 0.184 a2: 0.106, 477, <0.001 a3: 0.115, 344, <0.001 b1: 0.147, 3,739, <0.001 b2: 0.165, 3,329, <0.001 b3: 0.151, 2,417, <0.001	a1: 0.002, 3,739, 0.445 a2: 0.016, 477, 0.222 a3: 0.021, 344, 0.183 b1: 0.017, 3,739, 0.018 b2: 0.019, 3,329, 0.011 b3: 0.025, 2,417, 0.003	
Tp.	a1: 0.031, 3,739, 0.02 a2: 0.244, 477, <0.001 a3: 0.192, 344, <0.001 b1: 0.19, 3,739, <0.001 b2: 0.213, 3,329, <0.001 b3: 0.19, 2,417, <0.001	a1: 0.014, 3,739, 0.173 a2: 0.111, 477, <0.001 a3: 0.123, 344, <0.001 b1: 0.153, 3,739, <0.001 b2: 0.172, 3,329, <0.001 b3: 0.161, 2,417, <0.001	a1: 0.003, 3,739, 0.429 a2: 0.021, 477, 0.162 a3: 0.029, 344, 0.109 b1: 0.023, 3,739, 0.002 b2: 0.026, 3,329, 0.001 b3: 0.035, 2,417, <0.001	a1: 0.001, 3,739, 0.484 a2: 0.005, 477, 0.411 a3: 0.008, 344, 0.371 b1: 0.006, 3,739, 0.213 b2: 0.007, 3,329, 0.191 b3: 0.01, 2,417, 0.13

4.4.3 Discussion

Our ground truth-based evaluation reveals several important factors about the strengths and weaknesses of the different best-match criteria presented herein. First, although each input dataset varied slightly in level of improvement, on average across both datasets, using an uncertainty based approach instead of the hierarchy- or reversed hierarchy-based method reduces the overall spatial error in geocoded datasets. The two left-most columns of data in Table 4.6 as well as the descriptive statistics in Tables 4.2 and 4.3 provide evidence that using an uncertainty-, gravitationally-, or topologically-based approach offers an improvement in the level of spatial accuracy that one can achieve over any form of hierarchy-based approach. The t-test results reveal that these improvements are highly significant when considering a national dataset such as the Best Western and Target locations we evaluate herein (Table 4.6). When considering all records in the combined dataset, spatial error is reduced by 0.354 km (from 1.283 to 0.929 km) when used in conjunction with a street reference data layer, and by 1.854 km (from 4.87 to 3.016 km) when using only ZCTA and Place layers (Table 4.2). These effects become even more pronounced when considering just the records to which the uncertainty-, gravitationally-, and topologically-based methods apply, reducing by 2.779 km and 2.082 km for U.S. Census Bureau TIGER/Lines, ZCTA, and Place and ZCTA and Place reference source combinations, respectively (Table 4.3).

Second, although smaller in magnitude, our results indicate that employing the gravitationally- and topologically-based approaches follow a similar pattern of spatial error reduction as that of the uncertainty-based approach. When considering each dataset as a whole, the gravitationally-based approach improves on the uncertainty-based approach, and topologically-based approach improves further still beyond the gravitationally-based approach, in all cases. However, as shown in Tables 4.4 and 4.5 the benefits of these approaches can clearly be seen to depend on the topological relationship between the set of reference features used in each approach. In the case when only the ZCTA and Place reference data files are used for geocoding (Table 4.4) as well as when U.S. Census Bureau TIGER/Lines, ZCTA, and place are used (Table 4.6), it is clear that the most benefit can be realized from the gravitationally- and topologically-based approaches when the underlying reference features overlap. For all other topological relations (contains, disjoint, and touches), there are instances when both of these methods perform better and/or worse than each other and/or the uncertainty- and hierarchy-based approaches.

Third, from the top left cell of Table 4.6 we see that substituting one hierarchy for another does not consistently improve the resulting spatial accuracy of geocoded data. This approach reduces the error in the Best Western dataset, while increasing it in the Target dataset (Table 4.2). Clearly, although some improvements may be seen in part or all of a national input dataset, these create a degradation of the spatial quality in other areas with different characteristics. This is indicated by the low levels of significance when choosing the reversed hierarchy over the NAACCR hierarchy when considering either the dataset as a whole or just those records that change when using only ZCTA and Place reference files (Table 4.6).

Fourth, although the gravitationally- and topologically-based approaches improve upon the uncertainty-based approach, the levels of improvement one could expect are marginal in some cases. On average, these improvements are typically on the order of tens of meters, which are at most just a small fraction of the total spatial improvement when considering that these errors are typically on the order of several kilometers to begin with. Further, these improvements are seen to be statistically significant ($p < 0.05$) only when considering the geocodes produced using the ZCTA and Place reference files without including the U.S. Census Bureau TIGER/Line files. This could indicate that the majority of cases which fail to match to a street are affected only slightly by either of these approaches, but that in general, these approaches improve the accuracy in a USPS

ZIP code versus city decision. In a real-time application scenario, the computational complexity and additional processing overhead of the topologically-based method in particular may outweigh the benefits in resulting spatial accuracy it offers. However, if processing time is not a consideration and the specific topological relationship between the underlying reference features can be determined, we find no reason not to utilize the topologically-based method over any of the other three for intersecting features. Under these constraints, this method consistently outperforms all other methods we developed herein.

Finally, our comparison of the results achievable when using a street reference data file in addition to a USPS ZIP code and city reference data layer versus just USPS ZIP code and city alone show that the uncertainty-, gravitationally, and topologically-based approaches can substantially reduce the error in both high and low street-level match rate scenarios. This is a particularly important finding because although street reference data sources are continually increasing in accuracy and parcel files are becoming more available every day, many recent reports continue to find high levels of non-street level matches (Zandbergen 2008a; Hibbert et al. 2009; Henry et al. 2008) for which a USPS ZIP code or city would typically be the next selection. So, any techniques that can be used to reduce the amount of spatial error, even at the low end of the geocoding accuracy spectrum, will continue to have a beneficial effect for some time to come.

4.5 Chapter 4 Conclusions

The goal of this chapter was to evaluate if the formalization of a best-match criterion for the selection of a geocode output from a candidate set provides a framework with which researchers may develop alternative criteria that outperform the current static hierarchical selection method used by commercial systems. To this end, we have developed the notion of a best-match criterion to describe the choices used by a geocoder to determine the best output from a set of candidate results. We have placed the current state-of-the-art hierarchy-based approach that uses a static ordering of qualitative codes for determining an appropriate output into this scheme and developed three alternative methods: (1) the uncertainty-based approach, which relies on quantitative characteristics to make a determination; (2) the gravitationally-based approach which exploits the spatial characteristics of multiple reference features; and (3) the topologically-based approach which utilizes the spatial relationships between reference features.

Our evaluation of these four best-match criterion in addition to a modified ordering of the hierarchy-based approach using a nationwide ground truth dataset reveals that the choice of the best-match criterion used in the selection of the output geocode does in fact matter when considering the spatial accuracy of the output location. We have shown that changing the order of layers in a hierarchy-based approach does not consistently improve performance. Further, the use of an uncertainty-based approach will improve the spatial accuracy of resulting geocodes as well as lower the geographic uncertainty associated with them. The application of more sophisticated spatial reasoning processes that make use of the uncertainty inherent when considering multiple reference features in conjunction (the gravitationally-based approach) and the spatial relationships between them (the topologically-based approach) only marginally improve the spatial accuracy beyond the uncertainty-based method alone, and may only be worth the extra complexity in systems requiring the highest levels of spatial accuracy.

The impact of the present study is limited by several factors that we plan to examine in our future work. First, this study only considers uncertainty derived from the geographic area of the underlying reference data feature selected. This approach assumes that no uncertainty is introduced by the other components of the geocoder, which is clearly not the case in most geocoding systems. We plan to develop methods for deriving quantitative descriptions of the geographic uncertainty produced within each component as well as methods for combining them into a single overall value for the geocode. Second, our results indicate that geographic characteristics of a region may influence the quality of a geocode produced from each of the reference datasets. As such, we plan to analyze the effect of predominant geographic characteristics to produce a predictive model for geocode error and uncertainty, thereby utilizing local knowledge for the selection of an optimal geocode. Finally, because many of the reference features used in geocoding are based on administrative or postal delivery regions, it may be possible to further refine our methods by the inclusion of population density as an additional component in uncertainty calculation and output location weighting.

CHAPTER 5: INTELLIGENT TIE-BREAKING FOR AMBIGUOUS GEOCODE CANDIDATES

This chapter will be submitted for publication as:

Goldberg, D.W., Wilson, J.P. Knoblock, C.A., and Cockburn M.G. 2010.

Intelligent Tie-Breaking for Ambiguous Geocode Candidates.

Journal of Spatial Information Science.

5.1 Chapter 5 Introduction

Geocoding is the process of converting postal address data into geographic coordinates, i.e., latitude and longitude pairs. At the highest level, a geocoding system consists of six main components (Boscoe 2008; Goldberg 2008; Rushton et al. 2006): (1) the input data to be geocoded; (2) the address parsing and normalization algorithms that identify and standardize the pieces of the input data; (3) the geographic reference data representing real-world geographic features from which an output geocode is interpolated or returned directly; (4) the feature matching algorithms that link the input data to one or more reference features; (5) the feature interpolation algorithms that identify where along or within a reference feature the output should be located; and (6) the output data.

During the geocoding process, geocoding systems may encounter situations where two or more underlying reference features (street segments, address points, etc.) are returned from a feature matching algorithm operating on some reference data layer (street centerlines, parcel boundaries, etc.), each above the match certainty threshold defined by the user and with equal probability of being correct. This results in a set of potential reference features, any of which could be a valid output. This situation typically results from either problems with the input address or the underlying reference data. An input address may be missing some information that could have disambiguated between the set of candidate matches (incompleteness) or some part of the input data has been entered incorrectly and was therefore dropped during the address relaxation portion of the feature matching algorithm. Similarly, a reference data layer may have incomplete or incorrect attribute values for certain reference features, e.g., missing directional values or overlapping address ranges for reference street segments.

In these cases where multiple reference features are all equally likely and valid matching reference features, typical geocoding systems in use today will take one of four paths: (1) involve the user to manually choose the most likely candidate; (2) use *a priori* knowledge about the attributes of an individual entity whose address is being geocoded to impute the most likely location given a geographic distribution of a particular attribute; (3) arbitrarily choose one reference feature for use in a feature interpolation algorithm to derive an output, e.g. always choose the first match or flip a coin; or (4) indicate an unsuccessful match for the reference layer in question and begin searching for a match in the next available reference data layer, e.g. try a USPS ZIP code-level match after a street-level match. Choosing the first strategy will often result in the highest levels of accuracy, but requires that a human be involved in the process. In geocoding scenarios with small input datasets this may be an acceptable option, but our prior work has shown that the time and effort required for this process may become prohibitive as the size of the input dataset increases (Goldberg, Wilson, Knoblock et al. 2008).

Prior work has shown that the second option of geo-imputation can be quite effective in the health domain for determining an appropriate location of an individual down to the level of census tract (Henry et al. 2008), but requires that specific characteristics be known in advance about the person (or address) in question. In health studies, these data are often available because the address of an individual is just one of the many variables collected about study participants, particularly race and ethnicity which have direct corollaries in the demographic information associated with census tract boundaries. For nearly every other imaginable geocoding scenario outside of a health study, however, such detailed ancillary information about an address is simply not available rendering similar geo-imputation approaches non-applicable.

Therefore, in large-scale geocoding attempts where the scale of the data prohibits manual review and/or appropriate variables are not available for geo-imputation, either the third or fourth option is typically employed. The third strategy of arbitrarily choosing one of the available candidate geocodes, either at random or always in the same order, results in the creation of spatial data with marginal certainty (no greater than 50% since there are at least two viable options) which translates into limited utility for applications requiring high levels of confidence in their underlying spatial data. The fourth option throws away valuable information (a set of semi-certain matches) that could potentially be used to improve the output of other data layers or derive new information from which an output could be determined.

Given these limitations and pitfalls in current geocoding tie-breaking practices, this chapter examines what, if anything can be done to improve the overall spatial accuracy and uncertainty of geocoded datasets which include some subset of ties. To this end, we will attempt to answer three questions related to this problem: (1) How often does this situation actually occur in real world data, i.e. how bad is this problem in practice?; (2) Can information about the set of ambiguous reference features and their local neighborhood be used to produce an interpolated output more accurate than reverting to the next reference data layer; and (3) Do these alternative tie-breaking methods significantly reduce the average spatial error and uncertainty over simply flipping-a-coin to choose one of the available reference features or the centroid of them all?

To answer the first question we will investigate the prevalence of these ambiguous cases using a history of nine million geocoding transactions processed by our online geocoding engine (Goldberg et al. 2009), a reference dataset containing the addresses and parcel centroids for the 4.5 million addresses in Los Angeles County (Los Angeles County Chief Information Office 2009), a geocoded set of 1.7 million health care providers commonly used in health-related research (North American Association of Central Cancer Registries 2009), and two GPS points of interest (POI) files of national commercial retail chain locations. Together this varied set of test data comprises a large and representative sample of real world data a geocoding system could be expected to handle “in the wild”. To investigate the second question we first present an analysis of the types and sources of ambiguity in both input address data and reference data files which are used to develop two novel geocoding methods. The first, termed *dynamic feature composition* utilizes the convex hull of all ambiguous feature matching results to automatically create reference data features for use in feature interpolation. The second, termed, *geo-intelligent tie-breaking*, uses characteristics of the region around each candidate reference feature to select the most likely candidate. We evaluate these approaches against the state-of-the-art ESRI Address Locator geocoding system (Environmental Systems Research Institute 2009b) and tie-handling techniques on GPS ground truth and parcel centroid data to determine the effectiveness of our approach in producing more accurate geocodes, i.e. geocodes that are closer to the ground truth values.

The remainder of this chapter is organized as follows. In Section 5.2 we discuss related work. In Section 5.3 we detail several prototypical situations where ambiguous matches occur. In Section 5.4 we develop a novel geocoding method which uses neighborhood characteristics and candidate geocodes to identify the most likely reference feature in a set of ambiguous matches. In Section 5.5 we evaluate the performance of our method on a nationwide dataset of GPS ground truth data and local parcel centroids. We end with conclusions and future directions in Section 5.6.

5.2 Chapter 5 Related Work

Many studies have identified ambiguous geocoding results as a consistent problem in geocoded data (Boscoe 2008; Frizzelle et al. 2009; Gilboa et al. 2006; Goldberg 2008; Goldberg et al. 2007; Karimi et al. 2004; McElroy et al. 2003; Rushton et al. 2006; Ward et al. 2005; Whitsel et al. 2006; Zandbergen 2008a). However, of these, only a few reports have focused on the underlying causes of these problems (Karimi et al.

2004; Zandbergen 2008a). To the best of the author's knowledge, no work has evaluated the benefits of the different technical methods of handling ambiguous cases other than using a geo-imputation approach based on characteristics of the population to determine the best approximation down to the census tract level using demographic variables (Henry et al. 2008).

Our approach to tie-breaking using characteristics from the region around each ambiguous candidate reference feature can be seen as the inverse of the evidence-based co-occurrence model used in named entity recognition/disambiguation in the natural language processing literature (Li et al. 2003). These approaches look for additional evidence to ground the geographic scope of a text document from within the document itself, whereas our approaches which follow will look for additional evidence in the surrounding geography. To the best of the author's knowledge, no published work has explored utilizing such an approach to determine the most likely candidate from a set of ambiguous reference features.

5.3 Sources of Ambiguity in Geocoded Data

The geocoding literature is rich with reviews of the technical components of the geocoding process which identify numerous reasons why ambiguous geocoding results (ties) occur in practice (Boscoe 2008; Frizzelle et al. 2009; Gilboa et al. 2006; Goldberg 2008; Goldberg et al. 2007; Karimi et al. 2004; McElroy et al. 2003; Rushton et al. 2006; Ward et al. 2005; Whitsel et al. 2006; Zandbergen 2008a). A common thread throughout these discussions is the existence of incorrectness and incompleteness in both the input address and underlying reference data used by a geocoding system. Several such examples that we will investigate in detail are presented in Table 5.1 which are drawn from (a) a list of Best Western hotel addresses from a ‘point of interest’ (POI) website (POIfriend Inc. 2009); and (b) the National Provider Identification (NPI) file, a nationwide set of 2.9 million addresses of Medicare payment recipients representing hospitals, clinics, and doctors offices (North American Association of Central Cancer Registries 2009). The ambiguous matches present in this table occur when geocoded with the 2008 version of the U.S. Census Bureau’s TIGER/Lines reference data layer (U.S. Census Bureau 2009b).

Although great strides have been made on the part of the U.S. Census Bureau under the Master Address File improvement program (Broome et al. 2003), the U.S. Census Bureau TIGER/Line files still contain incomplete and/or incorrect attribute data for reference features around the country which continue to plague geocoding systems (Frizzelle et al. 2009; Wu et al. 2005). Example (1) in Table 5.1 is indicative of incomplete address range data associated with the U.S. Census Bureau TIGER/Line files. In this particular case, the address in question (701) is exactly between the known address ranges in the reference data file (631-699 and 703-707). It is clear that the input address should belong to one or the other of the two blocks, most likely the block in the 700s, but one cannot say for sure as addressing systems are known to vary widely even within small areas (Fonda-Bonardi 1994).

Table 5.1: Example input data resulting in ambiguous matches from (a) Best Western hotels; and (b) Medicare National Provider Index when geocoding with TIGER/Line reference data

Id	Source	Address	Ambiguous matches	Ambiguity reason	Type of ambiguity
1	(a)	701 N Main Street Colfax Wa 99111-2120	631-699 block of N Main Street 703-707 block of N Main Street	Between known address ranges	Reference data incompleteness
2	(b)	626 W Route 66 Glendora Ca 91740	616-698 block of Route 66 (W) 624-630 block of Route 66 (E)	E/W missing from reference features	Reference data incompleteness
3	(b)	8354 Natalie Lane West Hills Ca 91304	8338-8398 block of Natalie Lane 8336-8498 block of Natalie Lane	Overlapping address ranges	Reference data incorrectness
4	(b)	439 S 97th Street Los Angeles Ca 90003	439 E 97th Street 439 W 97th Street	Incorrect pre-directional	Input data incorrectness
5	(b)	222 Market Street Inglewood Ca 90301	222 N Market Street 222 S Market Street	Missing pre-directional	Input data incompleteness

Therefore, in a traditional geocoding approach, each of these candidates would receive an equal match score resulting in ambiguous candidate geocodes. Similarly, example (2) results in an ambiguous match because the pre-directional values are not included in the attributes of the two particular reference data features resulting from a feature matching attempt. Example (3) displays a third type of reference data error where the address ranges associated with multiple reference features both contain the address in question.

The improvements to the accuracy of spatial geometries and accompanying attribute completeness and correctness contained within commercially available street-segment reference databases are beginning to resolve many of these situations. However, the cost of these data sources is usually prohibitive for individual researchers and all but the largest institutions, commercial platforms, and/or government agencies. For this reason, the U.S. Census Bureau TIGER/Line files continue to be the commonly used reference data source in the most popular freely available geocoding systems including Geocoder.ca (Geolytica Inc. 2010), Geocoder.us (Locative Technologies 2009), JGeocoder (Liang 2008), and the USC WebGIS Geocoding Service (Goldberg et al. 2010). In contrast, examples (4) and (5) in Table 5.1 result in ambiguous candidate geocodes due to errors in the input addresses provided to the geocoding system. In example (4), the pre-directional attribute is present, but is incorrect. In this case, a feature matching algorithm would typically employ an attribute relaxation procedure to iteratively drop attributes until a feature match can be made (Levine et al. 1998).

This would result in the same situation as example (5) once the incorrect value for the pre-directional attribute is dropped. The address in example (5) is missing the pre-directional attribute. Because of this, a feature matching algorithm would be presented with two candidate street segments to choose from. In both cases, the reference data source contains two candidate street segments that are both equally likely to be correct.

When the input data are the problem, no amount of improvement to the underlying reference data will be able to resolve this class of error. Ultimately, a geocoding system can only do so much with the input data it is given; poor input data will typically result in poor geocoding results. In such cases, it is common for the online geocoding service Google Maps (Google Inc. 2009) to attempt to “correct” a user’s input data using term frequencies for likely sets of tokens and historical search results to determine what was most likely meant by the input (Cantador et al. 2008; Cirasella

2007). In these approaches, similar to a web search where the user is asked “did you mean…”, one of the ambiguous geocode results that is calculated as “most likely” is automatically chosen and displayed on the map for the user. In many cases this approach gives the false impression that an address was geocoded successfully because, without looking at the actual address displayed, a user could assume that what they entered is what is being displayed. For casual mapping purposes and driving directions, this could be seen as an acceptable approach but for scientific studies, this approach would normally be frowned upon. In contrast, the Microsoft Bing and Yahoo! Maps mapping interfaces (Microsoft Corporation 2009a; Yahoo! Inc. 2009b) are not so bold as to pick an address for a user. Instead, they inform the user of the ambiguity and present the user with all of the multiple options from which the user must decide the one to use.

5.4 Current Tie-Breaking Strategies

Although each of the five example addresses listed in Table 5.1 would result in the same ambiguous geocode candidate scenario, there are important distinctions between several of them that could be useful for breaking the tie in a more intelligent manner than simply flipping-a-coin. In the following subsections, we will examine the rationale behind and potential error resulting from the currently used techniques of either reverting to a lower level of accuracy or flipping-a-coin. We then present several more intelligent techniques that could be employed to address each particular source of ambiguity.

5.4.1 Reverting to a lower level of accuracy

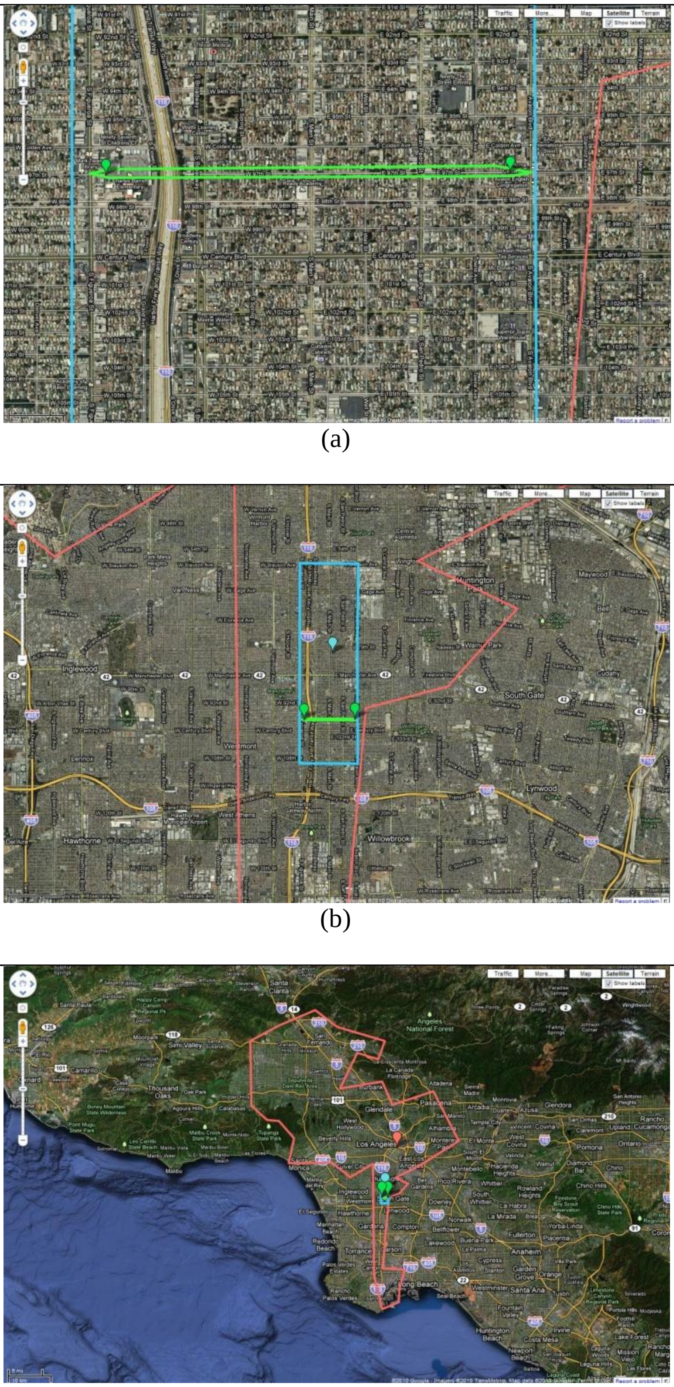
Many geocoding strategies used in the health sciences follow a protocol that excludes ties from being included in study data. In these instances, when geocoding ties occur, researchers follow a set protocol which orders the reference data layers that should be attempted subsequently to find a non-ambiguous geocoding match, as described in **Chapter 4**. The most commonly used of these hierarchies is the North American Association of Central Cancer Registries (NAACCR) GIS Coordinate Quality Codes (Hofferkamp et al. 2008), although it is also common to the approaches used in online geocoding services such as those provided by Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b). Here, when a match is not found at the building, parcel, or street level, a match is next attempted at the USPS ZIP+4 and USPS ZIP code levels before attempting one at the city, county, and state level.

In Figure 5.1, this approach is shown for example address 5 from Table 5.1 with the two area candidate outputs shown in green, an approximation of the USPS ZIP code for 90003 shown in blue, and an approximation of the boundary of the City of Los Angeles shown in red. The boundaries displayed in each color represent the regions within which all locations have an equal likelihood of being the true location. From this figure, it becomes clear that the spatial uncertainty with an output geocode greatly increases when applying this process as one moves through the different levels of the hierarchy. More specifically, if we define the potential spatial uncertainty associated with any particular geocode, g, to be ε, we see that ε greatly increases as we move from ambiguity at the street-level (green boundary) to the non-ambiguous match at the USPS ZIP code- and city-levels, respectively. Similarly, we see that the spatial error increases as the centroid of the regions (which would be the output geocode) get farther and farther away from the region within which we have reason to believe the true location resides.

In this approach, ε can be computed quantitatively as the area of the r_i reference feature responsible for producing the geocode output g_i, for $i = 1 \dots n$ or the number of possible matches (Equation 5.1).

$$\varepsilon = Area(r_i) \tag{5.1}$$

The spatial error resulting from this approach would vary depending on a particular input/reference feature combination, but would have a maximum value in the worst case scenario when the true geocode is positioned at the farthest point in the region from where the computed geocode is located. More generally, the average value of spatial error would equal the average distance from the centroid of the region to each point in the region. Note that this approach is not applicable to an ambiguous match and hence a value for the green area would not be calculated, which is consistent with the reversion to lower level data layers taken in this hierarchical approach.



(a)

(b)

(c)

Figure 5.1: Areas of ambiguity at (a) street level in green; (b) USPS ZIP code-level in blue; and (3) city-level in red

It is widely known that although this approach computes a non-ambiguous output, the utility of the resulting geocodes degrades quickly in terms of what these values can be used for in scientific studies based on the level of spatial error and uncertainty present (Krieger, Waterman et al. 2002; Whitsel et al. 2006; Zandbergen 2007). Therefore, following this approach is typically listed as a limitation within research that employs it and alternative methods to tie-breaking and/or other forms of geo-imputation are typically chosen when applicable (Goldberg, Wilson, Knoblock et al. 2008; Henry et al. 2008).

5.4.2 Flipping-a-coin

As previously noted, geocoding systems in use today that do not revert to lower levels of accuracy will commonly arbitrarily choose one of the available reference features when ambiguous results occur rather than prompt the user for intervention. In the best case when only two candidate reference features are found, this practice results in a maximum of a 50% probability of choosing the correct reference feature. In general, if n ambiguous candidate reference features are found, the probability of a system choosing the correct one becomes $1/n$. However, because these reference features are spatial in nature and are used to produce a spatial output, the actual spatial error resulting from this probability of correctly choosing a reference feature is not so straightforward. In particular, if the two reference features are spatially close to each other, the potential error that results from choosing the wrong one is less than if they were farther apart.

For instance, example 1 in Table 5.1 results in an ambiguous result because the address ranges associated with two contiguous street segments are incompletely described, i.e., missing an address from one or the other. Since these two reference features meet at the same location and have consistent address ranges both in terms of numeric direction (low to high) and parity (even/odd), a person investigating this scenario would most likely choose the one that contains the 700 block addresses because the input address is in the 700 range. Here, the spatial error resulting from this decision is minimal because in the worst case where the address actually belonged on the end of the 600 block, this output location would only be on the order of tens of meters away.

In contrast, if we consider example 5 from Table 5.1, we see that the two candidate features are eight blocks away from each other at a distance of over a mile, so the potential for spatial error is quite a bit larger. From these two examples, it becomes clear that the spatial uncertainty associated with simple flipping-a-coin and randomly choosing one reference feature out of the set of candidates will not be consistent in every case. Instead, the spatial uncertainty resulting from this approach will be tempered by the distance between the two reference features. In particular, the spatial uncertainty, ε, associated with any particular geocode, g, output chosen at random from the set of ambiguous potential candidates, $[G]$, will be the area defined by the convex hull of the each r_i reference features used to produce each g_i candidate geocode in $[G]$ for $i = 1 \ldots n$ (Equation 5.2).

$$\varepsilon = Area\left(ConvexHull\left(\bigcup_{i=1}^{n} r_i\right)\right) \tag{5.2}$$

Note that the convex hull is used here because although we are assuming the reference features were selected correctly during the feature matching algorithm, we are not assuming that the attributes of the reference feature are defined correctly (e.g., address range could be reversed) which therefore means that an output could be located anywhere along or within each r_i reference feature.

We know that in the best case, this coin-flipping strategy will pick the wrong address 50% of the time and the resulting area of uncertainty associated with a geocode will be proportional to the distance between the candidate reference features, r_1 and r_2. Thus, the resulting spatial error for a geocode with two candidate reference features will either be zero (if picked correctly) or d (if picked correctly), where d is the spatial distance between r_1 and r_2. On the average, if both candidates have an equal probability of being correct, this results in a spatial accuracy of $d/2$ for any particular ambiguously matched geocode. However, in the general case with $n > 2$ candidates, the average spatial error resulting from this approach becomes the average distance from the centroid to each of the ambiguous candidate features.

5.4.3 Dynamic feature composition

An alternative to picking one reference feature at random is to dynamically create a new reference feature that includes all ambiguous results by calculating their convex hull and use this new feature for interpolating an output. A workflow for accomplishing this task, which we term *dynamic feature composition*, is displayed in Figure 5.2 for a hypothetical input address and set of ambiguous reference feature results. By definition, each of the ambiguous matches is equally likely, so we know that the convex hull containing them all includes what should be the most likely location for a computed output. However, picking any one at random rather than the centroid of the whole region would result in a higher overall spatial error. This is because, by definition, the centroid is the point within the region that is, on average, the closest to all other points in the region.

Therefore, this point minimizes the potential for spatial error that exists if the true point were on any of the candidate features. Essentially, choosing the centroid guarantees that although the output point will not be correct, the potential for error will be minimized because we have specifically chosen the point that is the closest to all others. This approach is conceptually the same as a commonly used geocoding method that relies on street centroids, and is functionally the same when the ambiguous street segments are contiguous.

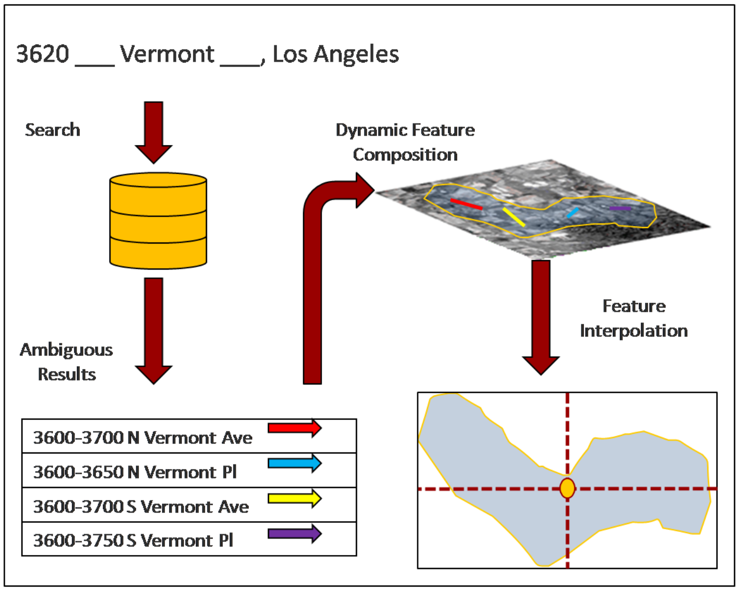


Figure 5.2: Dynamic feature composition workflow

Here, when an address number cannot be located on the street segment but the street segment can be found, a geocoder will return the centroid of that segment as the output, as is done in the online geocoding engines Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b). Both the dynamic feature composition and street centroid methods take the centroid of the reference feature, and the further constraint could be placed on the two contiguous features case to ensure that the output centroid lies somewhere on the actual reference features, as in the single street segment case. When viewed as a specific case of dynamic feature composition, we provide a method for the common street centroid interpolation method to be quantified, compared, and ranked against other geocoding outputs in terms of their spatial uncertainty. This is in line with and improves upon on the current state-of-the-art practices in that it still continues the status quo, while allowing street centroid output to be compared along with other possible output options.

It is important to note the relation between this approach and simply flipping-a-coin to pick one ambiguous match at random as described in the previous section. In particular, the dynamic feature composition method will result in an identical spatial uncertainty value as the case of simply flipping-a-coin, i.e., the area of the region defined by the convex hull of all ambiguous matches from Equation 5.2. However, dynamic feature composition is more likely to reduce the spatial error of a geocode output when the size of the candidate set increases above two. This is because as the likelihood of picking correctly goes down with the increase in number of candidates, the likelihood of increasing the resulting spatial error goes up because there are more options which are equally likely but not necessarily equally separated spatially. Therefore, picking the output location which minimizes the potential for spatial error becomes the better choice as the uncertainty amongst geocode candidates increases.

In either the flip-a-coin or dynamic feature composition methods for choosing an output location based on a set of two or more ambiguous candidate reference features, the goal is to avoid reverting to a lower level geographic feature as in the hieararchy-based approach described in **Chapter 4**. The intuition behind both these approaches is that by limiting the potential set of output locations within the entire area of a large geographic feature, the spatial error and spatial uncertainty of the resulting geocode will be smaller than choosing the centroid of the larger, less-certain feature that would have been reverted to following an alternative NAACCR hierarchy-type approach. In Figure 5.1 presented earlier, the benefits of such an approach are clearly visible. This figure dramatically illustrates that choosing either candidate reference features (green points) as in the flip-a-coin method or the centroid of the convex hull of both (green region) would result in a geocode with far greater spatial accuracy than choosing either the centroid of the USPS ZIP code (blue point) or the centroid of the city (red point) as well as an area of spatial uncertainty which is far smaller. For further discussion and a quantitative evaluation of such an uncertainty-based approach to geocoding error reduction see **Chapter 4**.

5.5 Types of Geo-Intelligence for Tie-Breaking

The discussion thus far has assumed that each of the ambiguous candidates resulting from a feature matching algorithm is equally likely and the best that can be done is to revert to a lower level, flip-a-coin, or use the centroid of all ambiguous reference features. In many instances this is the case, and the methods discussed up to here have been geared toward limiting and/or describing the potential for spatial error and uncertainty when considering ambiguous reference features independently (flip-a-coin) and in concert (dynamic-feature-composition). However, there are many instances where all candidate reference features are not equally likely, although a naïve approach would consider them as such. Depending on the underlying reason that caused the ambiguity, a geocoding system may be able to reason that one of the potential candidates is more likely than any other. If sufficient evidence exists and is available to a geocoding system, it would be beneficial if this information could be incorporated into a reasoning process and taken into consideration when applying a probability for selecting one candidate over another.

Consistent with similar concepts from the image analysis literature, we will consider this additional information as geo-intelligence, where "… [geo-]intelligence refers to geographic information that enables users to effectively perceive, interpret and respond to some specific issue … that is, geoinformation within a specific user context" (Hay et al. 2008, p. 80). In particular, in the subsections which follow, we will examine what, if any, geographic information could be derived and/or utilized about an input address, a single or group of reference candidate features, and/or the area immediately surrounding them to reason about the most likely candidate feature.

The goal in each of these sections is to utilize local information related to the specific input address in question to determine if one of the available ambiguous geocodes could be more likely than any of the alternatives. To do so, we will examine the character of each of the five example ambiguous addresses from Table 5.1. Through this discussion, a set of rules will be developed for handing each class of conflict, and an implementation strategy for realizing a subset of these will be detailed.

In the second example from Table 5.1, the pre-directional attribute is missing from two candidate street segments resulting in two reference features that are equally viable, although they are not spatially close to each other, instead they are on the order of 12 blocks or ~1.2 miles apart. Here, the attribute data for these particular street segments should contain a pre-directional that distinguishes them, but they instead appear to be identical from a feature matching perspective because this lone distinguishing attribute is absent. One approach to resolving this ambiguity would be to look at the attributes present in street segment features in the surrounding geography to determine if any evidence can be gleaned that would provide a basis for choosing one of the candidate features over the other. Although street numbering and addressing systems are vary across relatively small regions such as neighborhoods within a city (Fonda-Bonardi 1994), they are typically consistent within very small regions, and particularly so within a single USPS ZIP code. This is often the case because USPS ZIP codes are specifically designed to aid in mail delivery and are accordingly usually comprised of a single consistent addressing system, and will most often change on the borders of such systems. If we assume this to be true, a geocoding system can investigate the region around each candidate street segment to determine if other street segments in the neighborhood have a character consistent with the discriminating attribute of the input address. In example 2 from Table 5.1, our geocoding system should investigate the area around each of the candidate features to determine if other streets in the area consistently contain the W pre-directional. If so, this candidate should be chosen, which leads to Rule 1.

Rule 1: When an input address contains a directional attribute that is missing from the ambiguous reference features, use the directional character drawn from other reference features in the geographic area surrounding each candidate to pick the candidate that is contained in the region with the most similar directional character.

Example 4 in Table 5.1 results in ambiguity because the input address is incorrectly specified. Here, the pre-directional supplied does not match the pre-directionals available for the corresponding street segments in the reference data layer. While not always possible to determine, the reference features in this particular case can be used to definitively indicate that the input directional is wrong. We know this because all street segments in the area with the same name, suffix, USPS ZIP code, and city as that in the input address all consistently have either an E or W while the input address has an N. In these types of cases where the input address contains a directional indication as do the reference features, but they do not agree in cardinal direction axis (E/W versus N/S), a geocoding system can infer that the input data are wrong but probably provide an indication region within which the most likely output should be located. This incorrect directional information can be used to determine the most likely candidate reference feature based on the overall directional character of the area surrounding the ambiguous candidate reference features, as in Rule 2.

Rule 2: If the directional value associated with the input address is on an incorrect directional axis of the ambiguous candidate reference features, use the directional character drawn from other reference features in the geographic area surrounding each candidate to pick the candidate which is contained in the region with the most similar directional character.

Examples 1 and 2 from Table 5.1, in contrast, both result in ambiguous results because the reference data are incomplete. In the first, the address range attributes are missing the complete range of possible addresses, resulting in two nearby candidate reference features which are closely related to each other both spatially (intersecting segments) and non-spatially (nearly intersecting address ranges). In these cases, we know that the address in question falls between the two known address ranges, meaning that one or the other is incomplete. As discussed in the section about the flip-a-coin tie-breaking strategy, picking the street segment corresponding to the 700 block would be more likely to result in less spatial error than picking the street segment corresponding to the 600 block. In this specific situation, information about the relation of the reference feature and the input address can be used to intelligently break the tie. From this we create Rule 3, which also contains the restriction that the segment matched to must be within a certain number of blocks away.

Rule 3: When an input address falls between the known address ranges of reference features, pick the endpoint of the reference feature segment with the closer address range to that of the input address if the block distance is less than a set block distance, which we define as the variable d.

Example 3 in Table 5.1 shows how ambiguity can stem from an error in the reference data files. In this case, two separate street segments both contain the input address in question. We know that in reality this is impossible because the actual true location could only exist on one or the other, not both. In this specific case, one address range completely encompasses the other. The relationship implicit in this information reveals that one street address range is more specific than the other, most likely because one address range was updated during a street reference file improvement project such as the U.S. Census Bureau TIGER/Line MAF improvement project (Broome et al. 2003), while the other was not. Following this reasoning that portions of an input reference dataset are improved at different times, it would make sense for a geocoding system to infer that the more specific of the two is the correct one, leading to Rule 4.

Rule 4: If the address ranges associated with two ambiguous street segments both contain the input address and one of the reference address ranges completely contains the other, pick the candidate which is contained within the other.

Example 5 in Table 5.1 shows an address that is missing an attribute which results in ambiguous candidate reference features. In this case, the directional attribute would have been required to distinguish between the possible candidates but this information was not provided. Here, the directional reasoning approaches outlined in Rules 1 and 2 would not apply because the input address does not contain enough information to reason with any confidence about what the user intended. Tackling this class of ambiguous data where the user has under-specified the input presents a consistent challenge that will be the subject of future work.

5.6 An Algorithm for Tie-Breaking Using Regional Geo-Intelligence

In the previous section, we have outlined many of the types and sources of information that can be gleaned from the relationships between candidate reference features and an underlying geographic region to aid in determining the most likely candidate from a set of ambiguous results. The goal of this chapter is to investigate the challenges and benefits of formalizing one such approach, termed, *geo-intelligent tie-breaking*, which deals with the most common of these sources of ambiguity, those stemming from missing and/or incorrect directional indicators in either the input address or reference features with the goal of formalizing an implementation approach to address Rules 1 and 2 above. In combination with the implementation of Rule 3 as described in **Chapter 3** and the attribute comparison required to implement Rule 4, our overall approach will serve to address a wide swath of geocoding ambiguities resulting from incorrect input and/or reference data.

To begin, the overarching goal of our algorithm is to use the geographic regions around the potential candidates to provide evidence that one candidate feature is more likely than all of the others. Practically speaking, given r_i ambiguous candidate reference features which together comprise the set $[R]$ total ambiguous features for $i = 1 \dots n$, what we are aiming to do is compute $p(r_i)$, the probability that any particular r_i is actually the correct choice, with more discriminating power than simply assigning each an equivalent probability, i.e., $p(r_i) = 1/n$ for each $r_i \in [R]$. Because we are assuming that the ambiguous candidate reference features in $[R]$ differ only in the value of directional (pre- or post-directional) and that a small region should contain a consistent addressing system, we know that as one moves outward away from each $r_i \in [R]$ along the axis between any two of the reference features it should be possible to compute the overall directional character of the region. Therefore, our method uses a bounding box approach to select a subset of all reference features in the geographic region around each candidate in the direction away from all other candidate features. The directional attributes of the features in these regions can then be examined to determine the overall directional character present in the region. If the predominant directional character of a region matches the directional from the input address and is substantially more significant than that associated with any other region, we would like to assign a higher probability to that corresponding reference feature and a lower probability to all others.

To do so, we will draw a line which connects a single point, (x_i, y_i), on the edge of each r_i with the closest point of every other $r_j \in [R]$ for $j = 1 \dots n$ where $i \neq j$. This will create a series of incoming lines to (x_i, y_i), each with an angle of θ_j. If we define the average of all θ_j angles as $\bar{\theta}_i$, we will have computed the angle closest to the direction pointing the farthest away from all other r_j reference features for each r_i. We will then define b_i to be a bounding box placed around a candidate reference feature r_i, whose four sides all have length l. If we place each b_i on each r_i at (x_i, y_i) such that (x_i, y_i) is located $l/2$ along b_i and the angle of the intersecting side of b_i and r_i is orthogonal to $\bar{\theta}_i$, we will have defined a region around each $r_i \in [R]$ expanding outward a direction furthest away from all other $r_j \in [R]$, where $i \neq j$. This derivation is illustrated in Figure 5.3.

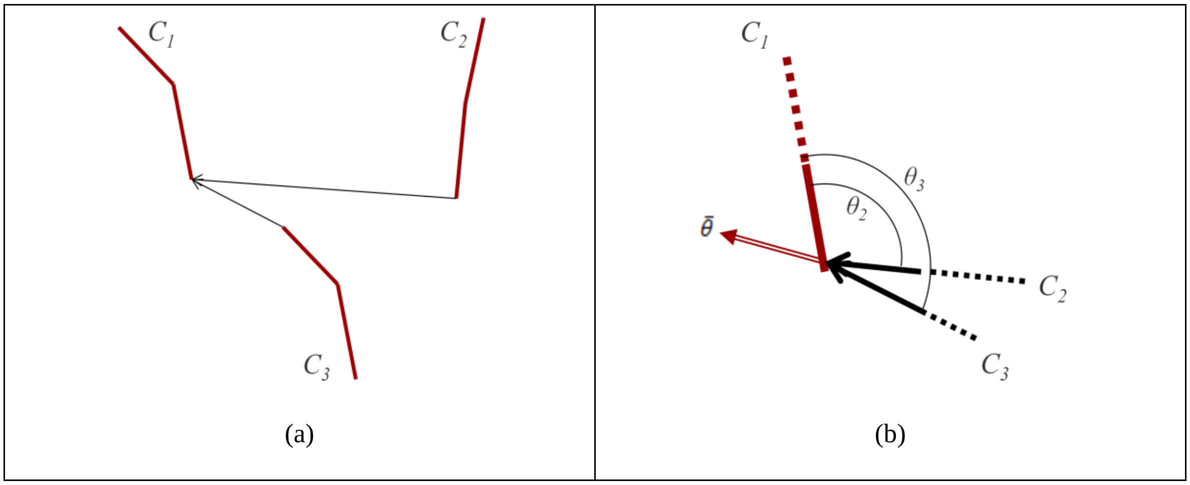


Figure 5.3: Derivation of incoming and outgoing angles from all candidate features; showing (a) a set of candidate features; and (b) the computation of the average incoming angle

With this definition, each b_i can be used as the boundary of a spatial query to obtain all reference features within the area surrounding and close to each $r_i \in [R]$. We will define the results of these spatial queries around each r_i which use b_i as the spatial filter as $[E_i]$. Each individual $e_{ik} \in [E_i]$ element represents a single street segment within the region for $k = 1..m$ with a spatial geometry and a vector of attributes $v_{ik} = \langle a_p \dots a_r \rangle$, where $a_p \dots a_r$ represent the non-spatial address components of the k^{th} street segment within the bounding box region around the i^{th} reference feature for $p = 1 \dots r$. These attributes will be specific to the address format of the reference features but at a minimum will include the address range, pre-directional, name, suffix, post-directional, city, state, and USPS ZIP code.

To use this method, we must first determine the size of each of the b_i bounding boxes. In our approach, we will use the same edge length l for each of the $b_i \in [B]$. In assigning the size of l, we would like to choose the smallest region possible that still discriminates between the predominant directional of the regions to ensure rapid processing of geocodes because the computational time will increase linearly with the number of reference feature street segments in each $[E_i]$ resulting from the spatial query using each b_i. To accomplish this, we will use the intuition that we should stop increasing the size of a bounding box when we begin to encounter reference features that are no longer relevant to the immediate area around a candidate reference feature r_i. These additional reference features from a substantially different region have little bearing on the directional character of the region immediately nearby r_i and may actually introduce false data if the region they are in has a different addressing scheme, e.g., when crossing a city or other administrative boundary.

Within each $[E_i]$, we can condense all v_{ik} vectors into a single vector of vectors, $V_i = \langle a_{kp} \dots a_{kr} \rangle$, where each of the k rows in V_i represents the k^{th} v_{ik}. We will use the USPS ZIP code and city attributes of the reference features as regional characteristics to determine if features being added to $[E_i]$ by the expansion of b_i are within the same region as the input address. If we consider the USPS ZIP code and city attributes to be nominal values, with each distinct value encountered in the USPS ZIP code and city elements of V_i within each $[E_i]$ as a separate class, we can use Shannon's information entropy (H) to measure the diversity of the region (Chao et al. 2003) across the city and USPS ZIP code variables in V_i, H_{ic} and H_{iz}, respectively. To constrain the maximum proportion of nearby areas considered during the evaluation of the predominant directional evaluation for a region, we will define δ to be the diversity threshold such that expansion ends when $H_{ic} + H_{iz} > \delta$. Our algorithm for obtaining $[E_i]$, the set of reference features in the neighborhood around each ambiguous candidate features $r_i \in [R]$, is shown in Figure 5.4.

Using each $[E_i]$ we can determine the major predominant directional character of both the pre- and post-directional values by maintaining a counter for each direction attribute value that is incremented when its value is encountered while scanning through the directional elements of V_i. This will result in a vector containing a count of the occurrences of each of the possible directionals across the region of $[E_i]$ as a whole, $d_i = \langle c_q \dots c_t \rangle$, where $c_q \dots c_t$ represent some descriptive set of directionals for $q = 1 \dots t$, such as the eight cardinal directions N, S, E, W, NE, SE, NW, SW. The occurrence of any particular directional c_q within the region $[E_i]$ can then be expressed as the function $d_i(c_q)$.

If an address numbering scheme in a region is well-behaved, meaning that it follows a typical pattern of assigning directional values in the direction in which they occur, i.e., E is east of W and N is north of S and vice versa, the values in this vector should quickly increase in the one or two cardinal directions that are predominant within the region. Thus, the directionals in d_{ik} with the highest values can be used to assign an overall directional character to the r_i candidate reference feature for which it is defined.

```
Input: a query address $Q$ with address attributes $q = \langle a_u \dots a_v \rangle$; a set of candidate reference features $r_i \in [R]$; the spatial reference data layer $L$ where $[R] \subseteq L$; the maximum diversity value $\delta$; and the bounding box the growth factor $\psi$.

assign $[E] \leftarrow \langle \quad \rangle$. $[E]$ is the set of reference features around each $r_i \in [R]$.
assign $[\theta] \leftarrow \langle \quad \rangle$. $[\theta]$ is the set of incoming angles for each $r_i \in [R]$.
assign $[XY] \leftarrow \langle \quad \rangle$. $[XY]$ is the set containing the point on each $r_i \in [R]$ closest to all $r_j \in [R]$, where $i \neq j$.
assign $l \leftarrow 0$. Current length is initialized to zero.
assign $H_{max} \leftarrow 0$. The maximum diversity value is initialized to zero.
for each reference feature $r_i \in [R]$ do
  $[R'] \leftarrow r_j \in [R]$, where $i \neq j$. $[R']$ is the set of ambiguous candidate reference features excluding $r_i$.
  $\vartheta_i \leftarrow \langle \quad \rangle$. $\vartheta_i$ is the set of incoming angles to $r_i$ from all $r_j \in [R']$.
  $[X_i Y_i] \leftarrow AvgNearestPoint(r_i, [R'])$. $[X_i Y_i]$ is the point on $r_i$ with the minimum average distance to all $r_j \in [R']$.
  for each reference feature $r_j \in [R']$ do
    $(x_j, y_j) \leftarrow NearestPoint([X_i Y_i], r_j)$. $(x_j, y_j)$ is the point on $r_j$ with the minimum distance to $[X_i Y_i]$.
    $\vartheta_{ij} \leftarrow Bearing\left((x_j, y_j), [X_i Y_i]\right)$
  $[\theta_i] \leftarrow Avg(\vartheta_i)$
while $H_{max} \leq \delta$ do
  $l \leftarrow l + \psi$
  for each reference feature $r_i \in [R]$ do
    $b_i \leftarrow BuildBoundingBox([X_i Y_i], [\theta_i], l)$
    $[E_i] \leftarrow Clip(L, b_i)$. $[E_i]$ is the set of reference features from $L$ within $b_i$.
    $h_{iz} \leftarrow Shannon([E_i], q\langle a_Z \rangle)$. $h_{iz}$ is the diversity of USPS ZIP codes values in $[E_i]$.
    $h_{ic} \leftarrow Shannon([E_i], q\langle a_c \rangle)$. $h_{ic}$ is the diversity of city codes values in $[E_i]$.
    if $h_{iz} + h_{ic} > H_{max}$ then
      $H_{max} \leftarrow h_{iz} + h_{ic}$

Output: the set of candidate features $E_i$ around each $r_i \in [R]$ with maximum diversity $H$ closest to $\delta$
```

Figure 5.4: Algorithm for determining bounding box size and obtaining reference features in the neighborhood of each ambiguous candidate reference features

However, by simply knowing the predominant directional values for each r_i, we cannot yet produce a quantitative probability value for a candidate reference features, $p(r_i)$ because, for instance, all of the other candidates may also have the same predominant directionals. What is required is a method capable of discriminating between them and computing a probability based on counts of directional values in one area in comparison to all others.

To derive this, we first note that each of the b_i bounding boxes can be thought of as a member of a single overall geographic area used to create a set of reference features $[\bar{E}]$ resulting from the union of the spatial clip with each b_i. The total number of reference features in $[\bar{E}]$, which we will denote as $|[\bar{E}]|$, will be equal to the sum of the number of reference features in each clipped region $[E_i]$, denoted as $|[E_i]|$ (Equation 5.3).

$$|[\bar{E}]| = \sum_{i=1}^{n} |[E_i]| \tag{5.3}$$

Each $[E_i]$ will have its own vector of computed directional counts, d_{ik} derived from its own set of individual reference features, $e_{ik} \in [E_i]$. The probability that any street picked at random in the areas surrounding each candidate reference feature $r_i \in [R]$ will have a directional value consistent with the input address is the frequency of occurrence of the c_q directional in region r_i out of the total number of streets in all regions, $|[\bar{E}]|$ (Equation 5.4), where $d_i(c_q)$ is the function that returns the count of directional c_q from the directional occurrence vector $d_{ik} = \langle c_q \ldots c_t \rangle$ for region r_i.

$$p(r_i) = \frac{d_i(c_q)}{|[\bar{E}]|} \tag{5.4}$$

In this approach, $p(r_i)$ becomes a quantitative measure describing the relationship between the directional values associated with the input address and those of the street segments found in the region surrounding each $r_i \in [R]$ candidate reference features. The more times the input address directional appears around r_i, the higher the value of

$p(r_i)$, meaning that we can compute and use relevantly weighted probability values when choosing between candidate features. Each $p(r_i)$ can then be normalized to $0 \leq \bar{p}(r_i) \leq 1$ (Equation 5.5) and used to probabilistically pick the candidate reference feature with the highest likelihood.

$$\bar{p}(r_i) = \frac{p(r_i)}{\sum_{i=1}^{n} p(r_i)} \tag{5.5}$$

5.7 Chapter 5 Experimental Evaluation

Our evaluation seeks to answer three specific questions: (1) How often do ambiguous reference features occur which prevent successful geocoding?; (2) What level of spatial error improvement results from the various alternative approaches to dealing with ties?; (3) What level of spatial uncertainty improvement results from the various alternative approaches to dealing with ties? Together, these three questions will shed light on the relative levels of effectiveness one could anticipate by applying any particular tie-handling approach over another.

To determine a measure of how frequently geocode candidate ties occur, we investigated the prevalence of occurrence across four large and varied datasets of real-world data drawn from addresses in administrative government lists, retail commercial lists, and a list of all addresses processed by our online free geocoding system. These addresses were processed by two independent geocoding systems to produce a subset or records which resulted in ambiguous results. This subset was then used to evaluate the performance of each tie-handling approach.

To do so, we processed the ambiguous set of records with each approach and performed a geographic analysis of the spatial error in each method as determined from the GPS locations of the ground truth values and an administrative set of parcel centroids. The flip-a-coin strategy was run five times to ensure an even distribution candidate selection. All other methods were run once.

5.7.1 Data sources and methods

5.7.1.1 Sample address data

To capture a realistic view of the effectiveness of our approach on a broad swath of real-world data, we utilized the four sample address datasets in Table 5.2. To speed up the processing of the sample data, each of these datasets were pre-processed to remove duplicate addresses leaving only one record for each unique street address, city, state, USPS ZIP code combination. Likewise, records with empty street address fields were removed and excluded from our evaluation.

Table 5.2: Sample input datasets with number of records originally and after pre-processing

Source	Original count	Count after pre-processing
USC WebGIS transactions	12,119,850	6,354,666
Medicare NPI file	2,903,156	1,086,196
LA County address points	2,890,639	2,890,639
Best Western hotels	2,074	2,074
Target stores	1,648	1,648
Totals	17,917,367	10,335,223

The first of these data sources was a set of 12.1 million historical transactions processed by the USC geocoding website which is available online at https://webgis.usc.edu and provides free geocoding services using the geocoding engine described in this dissertation. This site requires users who geocode more than 2,500 records to create a user account, and as such has over 2,800 individual users from industry, academia, and government, processing over 40,000 geocoding transactions on a typical day. This dataset represents the broadest view of input address data stemming from highly varying sources and ultimately being used for any number of end applications ranging from health studies to commercial endeavors.

To ensure the widest possible usage amongst potential users, this site provides users the option of not storing their transaction details (i.e., input data query or output results), thereby providing a small degree of confidentially to those who would rather not have their data recorded for confidentiality or other purposes. As shown in Table 5.2, this ability resulted in nearly half of the total queries sent to the system since its inception being removed when de-duplication and/or empty records were cleaned in the pre-processing step. In addition, without any further end-user information, we can assume that these data represent both residential and commercial addresses throughout the U.S.

The second dataset was the National Provider Identification (NPI) file, a nationwide set of 2.9 million addresses of Medicare payment recipients representing hospitals, clinics, and doctors offices (North American Association of Central Cancer Registries 2009). The version of the file used for this study was obtained from the GIS Committee of the North American Association of Central Cancer Registries (NAACCR) which provides a geocoded version of the NPI file for use by health researchers without access to a production scale geocoding system of their own (North American Association of Central Cancer Registries 2009). This dataset represents an official government list compiled from self-reported information from each individual entity that receives Medicare payments. As seen in Table 5.2, this dataset also contained a significant number of duplicate records because each individual doctor and/or service provider is required to be included in order to receive Medicare payments, resulting in repeated records for each of the multiple doctors that may operate at a clinic or other facility with a single postal address. Likewise, the same postal address for different departments within the same clinic or facility would also be repeated. By definition, these data represent only commercial addresses throughout the U.S. Note that these data are not explicitly included in the 12.1 million addresses present in the first reference dataset although it is entirely possible that one or more of the website users may have geocoded some portion of these data.

The third reference dataset was an administrative list obtained from the Chief Information Officer of Los Angeles County containing the full postal address and geocodes for all 2.9 million addresses in Los Angeles County (Los Angeles County Chief Information Office 2009). This list is used by the various Los Angeles County government offices for all geocoding needs and is based on the parcel boundaries that are used for taxation purposes (Los Angeles County Assessor's Office 2009). This list is highly complete and accurate in terms of the number and specificity of addresses included because Los Angeles County has a vested interest in ensuring proper tax assessment and collection, as do the citizens of Los Angeles County responsible for paying these taxes. As such, Table 5.2 shows that there were no addresses that were duplicated or blank within this dataset. By definition, these data represent both residential and commercial addresses. This dataset also included the latitude and longitude coordinates of the parcel associated with each address.

The fourth and fifth address datasets were listings of Best Western hotels and Target store locations in the U.S. These datasets were obtained from a 'point of interest' (POI) website (POIfriend Inc. 2009), and are intended to be imported into a GPS unit for navigation to these locations. According to the source, these lists contain the "official" locations of Best Western and Target locations obtained through GPS readings produced by each company and also include the full postal address. Because these companies each have a vested interest in making sure that people can find these locations, we can assume that these locations are of fairly high accuracy. Additionally, 10% of each dataset was manually reviewed using a previously developed interface (Goldberg, Wilson, Knoblock et al. 2008), and in all cases were found to be positioned within 10 m of the building. These ground truth datasets are currently available for free and can be obtained by any researcher who wishes to extend, replicate, or validate our work.

5.7.1.2 Geocoding Systems

To determine the frequency with which ambiguous geocoding results are encountered, our experiments utilize two geocoding systems; (1) the Address Locator (Environmental Systems Research Institute 2009c) portion of the ESRI ArcGIS 9.3 platform (Environmental Systems Research Institute 2009b), a commercial-off-the-shelf geocoding system routinely used by researchers, institutions, and government agencies (Rull et al. 2009; Omer et al. 2008; Blondin et al. 2007; Wagner et al. 2009; Macintyre et al. 2007; Ries et al. 2009; Wrobel et al. 2008; Ruiz et al. 2007; Pezzoli et al. 2007); and

(2) the USC geocoder, an experimental geocoding system built by the University of Southern California GIS Research Laboratory as a research platform for implementing and testing novel geocoding techniques and data sources that forms the basis of the online geocoding service provided the USC GIS Research Laboratory.

The implementation details for the ESRI Address Locator can be found in the online documentation accompanying the product (Environmental Systems Research Institute 2009a), but we must note that the ESRI StreetMap North America (Environmental Systems Research Institute 2009d) local roads dataset was chosen as the reference data layer because it is widely used throughout research reports in conjunction with the ESRI Street Map North America Address Locator (Environmental Systems Research Institute 2009d) that is shipped with the data. This combination is commonly used by researchers to investigate spatially-based research questions because of its ease of use and reported levels of high accuracy, e.g. (Wrobel et al. 2008; Ruiz et al. 2007; Gupta et al. 2009). Within the address locator, the default match score settings were used which include 80% spelling sensitivity, 10% maximum candidate score, and 60% minimum match score. The “match candidates if tie” option was turned off (turned on in the default settings) because we did not want to accept ambiguous matches as valid outputs.

Being a closed source solution, it is not easily possible to extend the ESRI Address Locator to support the dynamic feature composition or alternative candidate selection methods described herein. As such, our evaluation of any spatial error and uncertainty improvements resulting from these approaches will focus only on those achieved by the USC geocoder. The version used for this research (2.94) is available online at https://webgis.usc.edu, for which the implementation details and a performance evaluation against several commercial geocoding systems including ESRI, Google, Microsoft Bing, and Yahoo! (Environmental Systems Research Institute 2009b; Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) can be found in our prior work (Goldberg et al. 2009; Swift et al. 2008) and **Chapter 3**. For our experiments in this chapter, the reference data used include the freely available 2008 versions of the U.S. Census Bureau TIGER/Lines (U.S. Census Bureau 2009b), Places, and ZCTA5 files (U.S. Census Bureau 2009a). We utilize the OCG STConvexHull() function provided by the 2008 Microsoft SQL Server spatial data types to create a bounding region containing all ambiguous reference features. The centroid of this region is then returned as the final output geocode using the OGC STCentroid() function as implemented in the 2008 Microsoft SQL Server spatial data types. The same settings as listed in Chapter 3 were used for these experiments along with a diversity threshold of 10% in the bounding box expansion algorithm.

To calculate the frequency of the occurrence ties in street-level matches, each of the reference datasets was processed twice, first with the USC geocoder using the U.S. Census TIGER/Line files and second with the ESRI address locator using the ESRI Street Map North America reference data layer to validate that the rates obtained by the USC geocoder are what other commercial platforms would also produce. To calculate the levels of spatial error and uncertainty improvements, the records that resulted in ties within the USC geocoder were then processed four times, once using each of the approaches to handling ties outlined above: (1) reverting to a lower level of geography following the reference dataset ordering of the NAACCR Coordinate Quality Codes (Hofferkamp et al. 2008); (2) flipping-a-coin to pick a reference feature at random; (3) using the centroid of the dynamically composed reference feature created from all ambiguous candidates; and (4) using the tie-breaking strategy that looks within the region around each candidate feature to pick the most like candidate based on the directional character of the region.

5.7.2 Results

The street-level match and ambiguity rates resulting from processing each reference data file with the respective street-level accuracy reference data layers displayed Table 5.3 reveal that overall, a high proportion of the total input data were able to be successfully geocoded to street-level accuracy, averaging 81% and 77% across all datasets for the USC and ESRI geocoders, respectively.

Table 5.3: Counts of successful and ambiguous street-level matched records per input dataset for the USC and ESRI geocoders

Dataset	Record count	USC street-level match ESRI street-level match	USC ambiguous ESRI ambiguous
USC WebGIS	6,354,666	4752122 (74.8%) 4,510,710 (71%)	45,941 (0.72%) 158,292 (2.49%)
NPI	1,086,196	946971 (87.2%) 922,955 (85%)	7,628 (0.7%) 19,748 (1.82%)
LA County	2,890,639	2649239 (91.6%) 2,564,852 (88.7%)	6,345 (0.22%) 15,785 (0.55%)
Best Western hotels	2,074	1772 (85.4%) 1666 (78.6%)	17 (0.82%) 88 (4.24%)
Target stores	1,648	1374 (83.4%) 1295 (78.6%)	10 (0.61%) 34 (2.1%)
Total	10,335,223	8351478 (80.8%) 8001478 (77.4%)	59941 (0.58%) 193947 (1.88%)

In all instances, the USC and ESRI geocoders obtained match and ambiguity rates within 4% of each other, indicating a high degree of correlation between the tie rates of the two geocoding systems, similar to those found in **Chapter 3**.

The sources of ambiguity for 88% of all ambiguous cases are shown in Table 5.4, broken down by address attribute that caused the ambiguity and what the problem was with the attribute. The other 12% of ambiguity cases resulted from combinations of two or more attributes shown in Table 5.4, combining for 328 distinct combinations of attribute/cause sources of ambiguity. Not shown in this table is the fact that 656 records (1% of all ambiguous cases) resulted in an ambiguous match because two reference features in the reference data layer had the exact same attribute values, indicating either a duplication of a single reference feature or two distinct reference features with the same attributes. The cause of the number attribute resulting in error could be considered an incompleteness of the input data, an incorrectness of the input data, or both.

Table 5.4: Counts of ambiguity sources by address attribute and discrepancy

Count	% of total ambiguous records	Attribute	Cause
26284	43.85	Number	Incomplete/Incorrect
11407	19.03	Pre-directional	Incomplete
3874	6.46	Post-directional	Incomplete
3695	6.16	Pre-directional	Incorrect
2179	3.64	Suffix	Incomplete
1591	2.65	City	Incorrect
1334	2.23	Name	Incorrect
732	1.22	USPS ZIP code	Incomplete
713	1.19	Post-directional	Incorrect
230	0.38	USPS ZIP code	Incorrect
116	0.19	Pre-type	Incorrect
80	0.13	City	Incomplete
22	0.04	USPS ZIP code	Incorrect
8	0.01	Post-qualifier	Incomplete
6	0.01	Pre-type	Incomplete
1	0.01	Pre-qualifier	Incorrect

The counts of ambiguous reference feature address range relationships are show in Table 5.5. The relationships in this table are: (1) contain – one address range is completely contained within another; (2) next to – the ambiguous address ranges are within an address distance of six or less from each other; (3) overlap – the address ranges have some portion which overlaps; (4) disjoint – the address ranges are greater than an address distance of six from each other; and (5) equivalent reversed – the address ranges are equivalent to each other with the exception that the numeric ranges are in a reverse order. Note that the disjoint address range relationships occur in this list because of the use of the nearby feature matching option within the USC geocoder. These are the cases where an exact address range match could not be made and the closest neighboring street segments were selected and were each equally probable, based on the approach taken in **Chapter 3**. Noticeably absent from this list of relations are intersections, where the end address of one address range is the start of another, and equivalent, where the two address ranges are equal.

Table 5.5: Counts of ambiguous address range relationship types

Count	% of total	Relation
10,114	38.95	Contains
8,992	34.63	Next to
4,059	15.63	Overlap
2,379	9.16	Disjoint
420	1.62	Equivalent reversed

From the set of 10,005 ambiguous records, 4,357 were from the Best Western, LA County, or Target reference datasets for which either a GPS ground truth or a parcel centroid location were available. Of these, 189 were ambiguous because of a pre-directional or post-directional conflict, and 2,968 were ambiguous because of an address range conflict. If we consider a correctly chosen candidate feature to be the candidate reference feature with the least spatial error with respect to the ground truth and/or parcel centroid point, our neighborhood approach made the correct decision across all ambiguity types 82% of the time. For directional ambiguities, the system chose correctly 185 times (98% of all cases). The four that did not succeed resulted from cases in areas where there were equal numbers of the directional in question present in each of the regions around all of the candidate reference features. In these cases, our approach resulted in assigning all candidates an equal probability, resulting in incorrect assignments in each instance. The address range relationships and percentage correctly chosen by our algorithm versus those chosen across five independent runs of a coin-flipping strategy are displayed in Table 5.6.

Table 5.6: Counts of ambiguous address range relationship types correctly selected

Relationship	Total	Neighborhood Correct (%)	Average of 5 coin flips Correct (%)
Contains	1,727	1,704 (98.67%)	858.6 (49.72%)
Next To	1,174	857 (73%)	433 (36.88%)
Overlap	1,009	606 (60.06%)	292 (28.94%)
Disjoint	126	99 (78.57%)	49.8 (39.52%)

To measure the statistical significance of improvements in spatial error, a log transformation was performed to normalize the distribution of spatial errors inherent in our highly skewed data before attempting a Student’s one-tailed t-test with $\alpha = 0.05$ (Zimmerman et al. 2010; Zandbergen 2008b). In terms of distance to ground truth or parcel centroid, the average spatial error for the ambiguous records was reduced from 0.551 to 0.447 km and 0.171 to 0.14 km for the directional and address range ambiguity classes using our approach, respectively, when compared to the average error of the five flip-a-coin runs. Both were statistically significant $(p < .001)$. In the address range ambiguity cases, spatial error was reduced from 0.162 to 0.141 km for the dynamic feature composition and our neighborhood approach, respectively, again both statistically significant $(p < .001)$. However, in the case of directional ambiguities our approach fared worse, averaging 0.447 km as compared to the dynamic feature composition method which averaged 0.408 km, which was statistically significant $(p < .001)$. When considering the sample dataset as a whole, our region-based approach improved spatial accuracy by 34 and 18 m with statistical significance of $(p < .001)$ and $(p < .011)$ over the coin-flipping and dynamic feature composition methods, respectively.

To perform a comparison of the reduction of spatial uncertainty associated with the tie-breaking approaches versus reverting to a hierarchy, the maximum average spatial uncertainty for any of the tie-breaking approaches (flip-a-coin or neighborhood) can be approximated by the area of the convex hull of the region that contains all candidate reference features. This equates to the same region used for interpolation in the dynamic feature composition approach. Across all ambiguous records, the average spatial uncertainty was 14,397 m^2 when using any of the tie-breaking approaches or our dynamic feature composition centroid. In contrast, when using a hierarchy-based approach that reverts to a lower level of geography instead of attempting to break a tie, this average spatial uncertainty increases to 29.507 km^2.

5.7.3 Discussion

The results of our experiments shed light on several interesting factors. First, it is clear from Table 5.3 that ambiguous geocoding results do in fact occur with some regularity among many different classes of input datasets that geocoding systems would need to be able to process. The presence of a significant number of ambiguous cases in the Los Angeles County address dataset shows that these cases can occur in input data which have been cleaned to the highest possible quality. These cases occur even when using the ESRI StreetMap North America (Environmental Systems Research Institute 2009d) local roads reference dataset, considered by many to be a far superior reference data layer than the U.S. Census Bureau TIGER/Lines (U.S. Census Bureau 2009b). What is more striking is the far higher magnitude of ambiguous cases resulting from the ESRI Address Locator over the USC geocoder. The work presented in **Chapter 3** has shown that the feature matching approaches taken in our geocoding system are successful at increasing the number of address-level matches, thereby lowering the occurrence of ambiguous cases but clearly not able to eliminate them. So, even in the best of all cases (excellent reference data and sophisticated matching algorithms), ambiguous geocode results represent a challenge that geocoding systems will need to deal with.

Second, Table 5.4, shows us that the address ranges associated with reference data features and/or the address number component of an input address are by far the most common source of geocoding ambiguity, representing 44% of all ambiguous cases. Although, as noted earlier, the U.S. Census Bureau TIGER/Line MAF improvement project (Broome et al. 2003) is specifically geared toward addressing this challenge as one of its goals, it becomes clear from our analysis that much more work is needed to shrink the gap between the address ranges associated with street segments in national level geospatial products such as the U.S. Census Bureau TIGER/Line files (U.S. Census

Bureau 2009b) and what actually exists on the ground. Our results show that the rule-based approach that we have outlined herein can make significant strides toward intelligent tie-breaking given the characteristics of the candidate features, but a better course of action would be to eliminate the problem by correcting the address ranges associated with streets in the reference data files.

Third, our approach to tie-breaking when a directional attribute caused the ambiguity picked the correct reference feature in all cases except when the street pattern was irregular, specifically, diagonally-based as opposed to a typical rectangular grid where the streets run horizontal and vertical. It may be the case that there are other street patterns that could cause similar problems with our approach, and it may also be the case that our approach would not fail in every area with diagonal-based streets. In particular, the specific cases that failed here were also quite close to the axis of the street pattern (near the 100 N and 100 S blocks) where there is the most variability in the street pattern in terms of the directional values present on streets in the area. Because of this, there was not enough information available in the region to discriminate one from another based on a particular directional characteristic. Thus our approach assigned equivalent probabilities to all candidates and essentially reverted to a simple coin-flipping strategy. These results add to the existing body of work geared toward making strides in this area (Michalowski et al. 2007).

Fourth, from Table 5.6 it is clear that our strategy for tie-breaking an address range ambiguity when one address contains another provides the most consistent benefit of all possible relationships resulting in the correct choice in 99% of the cases. Likewise, our strategy of picking the candidate feature that is on the same numeric block range as the input address (i.e., 2652 = 2600 block) appears to outperform coin-flipping when the two address ranges are exactly next to each other (next to relationship) as well as when they are farther apart (disjoint relationship). It is interesting to note that in an actual repeated random simulation, the flip-a-coin strategy actually performs far worse than a 50-50 split as anticipated because of the number of scenarios where there were more than two options, 4% of the time in our analysis. That said, our strategy does not perform as well when the two address ranges overlap as in the other relationship types – only marginally better than if one were to simply flip an ideal coin with a 50% probability of picking the correct candidate feature. Upon investigating 20 randomly chosen of these cases, no systematic pattern was revealed as to why our strategy failed to produce better results in these cases.

Fifth, the comparison of our neighborhood approach to a dynamic feature composition shows that in many cases, choosing the centroid of the bounding region (convex hull) of all ambiguous features will do a good job of reducing overall spatial error in a geocoded dataset. However, when looking at the larger dataset of all ground truth data points and parcel centroids, it is clear that the benefits of this approach begin to dissipate as either the number of or spatial distance between candidate reference feature increases.

Sixth, our analysis of both the address number and directional ambiguities shows that the spatial error reduction achieved using our approach is significant. Although these improvements are on an order of tens to hundreds of meters, there are many situations that require the highest levels of geocode accuracy where these improvements could make all the difference such as in an environmental exposure analysis where exposure misclassification can occur within just a few meters.

Finally, when considering the total region encompassed by the reference feature used for interpolation as a measure of the spatial uncertainty associated with a geocoded record as described in **Chapter 4**, our results indicate that as expected, utilizing any approach to dealing with ambiguities (flip-a-coin, dynamic feature composition, or neighborhood-based) other than reverting to the next level accuracy in a NAACCR-type hierarchy (Hofferkamp et al. 2008) greatly reduces the amount of spatial uncertainty. These results are consistent with those found in our prior work that compared using spatially varying block metrics to identify nearby candidate reference features instead of reverting to a lower level of geography (**Chapter 3**) as well as that which compared using one hierarchy-based approach over another (Chapter 4). While intuitive, our results in the present work serve to provide more evidence that such hierarchy-based approaches should be avoided for research endeavors requiring all but the lowest spatial resolution.

5.8 Chapter 5 Conclusions and Future Work

In this study we have investigated a variety of approaches that can be used to deal with ambiguous results in geocoded data with an eye toward reducing the spatial error (distance from true location) and spatial uncertainty (the number of equi-probably locations). In doing so, we have described the potential for spatial error and uncertainty in the currently used approaches of flipping-a-coin and reverting to a lower level of accuracy as the impetus for developing methods to advance over these practices. To improve on the $1/n$ probability of choosing the correct candidate from a set of n equally probable candidate reference features, we have developed a series of rules for dealing with several specific types of address range ambiguity types and have created a region-based approach that used attributes in the region around to determine the most likely candidate.

To evaluate our new rules and region-based methods we have performed a spatial analysis of the accuracy resulting from all available methods using a ground truth set of GPS locations and parcel centroids. This analysis revealed that our approach consistently outperforms all others and the level of spatial improvement is statistically significant. The evaluation of our tie-breaking revealed that our methods choose the correct candidate reference feature 81% of the time in comparison to an ideal 50% chance inherent in flipping-a-coin when only two candidate feature are available. The spatial uncertainty reduction associated with choosing any of the tie-breaking or dynamic feature composition centroid approaches instead of reverting to a lower level of geographic match has been shown to be extremely large, reducing on average from ~30 km^2 to ~14,000 m^2.

The impact of the present study is limited by several factors that we plan to examine in our future work. First, our use of parcel centroids as ground truth values may have affected our evaluation of spatial accuracy on the order of the parcel size within which a particular address falls. Parcel sizes vary across regions so the effect of this on our results is most likely non-stationary across all of Los Angeles County. Second, although our approach chose correctly in 98% of the cases, the number of incidences of directional ambiguity was relatively small $(n = 189)$, which may mean that this aspect of our approach will not work as effectively across other datasets. To address the first of these limitations we plan to perform a large-scale ground truth exercise to collect both local and regional geographic characteristics about portions of Los Angeles County which could then be used to further validate our results. To address the second, we plan to perform a simulation study using the frequencies of occurrences of ambiguity types to evaluate the effectiveness of our approach across a large number of directional ambiguity instances.

CHAPTER 6: CONCLUDING REMARKS

This dissertation has investigated several methods for using spatially-based approaches to increase the level of spatial accuracy and decrease the level of spatial uncertainty in geocoded data. Specifically it has accomplished the following tasks: (1) described the source, scope, and magnitude of spatial error and uncertainty introduction in the geocoding process (**Chapters 2 – 5**); (2) developed a method which uses spatially-varying block metrics to dynamically score nearly candidate reference features (**Chapter 3**); (3) developed the notion of a quantitative best-match criterion for evaluating candidate reference features (**Chapter 4**); (4) developed three new methods for evaluating and combining multiple reference features – the uncertainty-, gravitationally- and topologically-based best-match criteria (**Chapter 4**); (5) developed a method for deriving new candidate reference features for use in feature interpolation using the set of ambiguous candidate reference features – dynamic feature composition (**Chapter 5**); (6) developed a set of rules to improve ambiguous reference feature tie-breaking using the relationships between ambiguous candidate feature address ranges (**Chapter 5**); and (7) developed a neighborhood-based approach to improve ambiguous reference feature tie-breaking using the predominant characteristics of the region immediately surrounding the ambiguous candidate features (**Chapter 5**).

6.1 Sources of Error and Uncertainty in the Geocoding Process

The present study has thoroughly examined the points during the geocoding process at which spatial error and uncertainty can be introduced as well as new methods for describing and measuring them both. This analysis has been cast as both backward-looking from a historic perspective (**Chapter 2**) as well forward-looking in the development of new geocoding techniques (**Chapters 3 – 5**). This investigation has considered each level of the geocoding process; from the input data submitted by the user, to the address parsing and normalization algorithms used to transform the data into a known set of attributes and consistent format, to the various types of geographic data that can be used as reference data layers, to the feature matching algorithms that select candidate reference features from the reference datasets, to the interpolation algorithms that determine where within or along a reference feature the geocode output should be located.

In focusing the technical improvements to predominantly the feature matching and interpolation components of this process (**Chapters 3 – 5**), the present work serves to provide guidance and insight to geocode consumers, practitioners, and developers on the potential levels of error and uncertainty that may be present in their geocoded datasets given the particular strategies that were used to create it. This contribution is important because in many cases, the sources of and propagation of error are ignored during the development or geocoding techniques and the utilization of geocoded data in scientific studies. The evaluation frameworks for assessing geocode quality based on spatial error and uncertainty presented herein can be used as a starting point for developing and/or refining classical and newly emerging geocoding techniques as computing power increases and reference data layers become more ubiquitous.

6.2 Nearby Candidate Scoring

Our method for scoring nearby candidate reference features using a spatially-varying block distance measure has been shown to greatly reduce the spatial error and uncertainty over the alternative approach of simply reverting to a lower level of geographic accuracy as is the common practice in commercially available geocoding systems (**Chapter 3**). The comparison of the USC geocoder to the industry standard ESRI Address Locator (Environmental Systems Research Institute 2009b) indicates that the base line level of accuracy of the USC geocoder performs on-par with existing state-of-the art-geocoding systems and validates its use in the development and testing of our novel nearby matching method. Our analysis of the results of this method in comparison to other online geocoding platforms with access to substantially better reference data layers Google, Microsoft Bing, and Yahoo! (Google Inc. 2009; Microsoft Corporation 2009a; Yahoo! Inc. 2009b) reveal that our 1.1% improvement in the match rates (the number of records able to be successfully geocoded) are not false positives. Instead, these improvements represent real spatial accuracy improvements have been shown to be useful in addressing the problem inherent in utilizing inaccurate and/or incomplete reference data files in geocoding systems such as the widely popular U.S. Census Bureau's TIGER/Line files (U.S. Census Bureau 2009b).

6.3 Geocode Candidate Best-Match Criteria

The development of a theoretical framework to support the quantitative evaluation of a set of candidate geocodes (**Chapter 4**) is an important contribution because it allows researchers to develop and test alternative strategies for choosing between potential geocoding outcomes. The three alternative best-match methods developed using this framework (**Chapter 4**), the uncertainty-, gravitationally, and topologically-based approaches, have been shown to vastly outperform the current state-of-the-art practice of reverting to a lower level of geographic reference features using a static hierarchy of feature classes when a failure occurs. Our experiments found statically significant spatial improvements when using both high quality (street-, USPS ZIP code-, and city-level) and low quality reference data layers (only USPS ZIP code- and city-level) showing that this approach works well at both ends of the geocoding reference data layer spectrum.

Our analysis of the hierarchy-, hierarchy reversed-, uncertainty-, gravitationally, and topologically-based best match methods revealed that although any of our new three methods (uncertainty-, gravitationally, or topologically-based) improve spatial accuracy, the simple uncertainty-based approach provides for most of the spatial error reduction. A further important finding of this work is that simply switching the order of layers in a hierarchy-based approach will not consistently improve the spatial accuracy across the board, although it will improve the results in areas with specific characteristics.

In addition to providing improvements to the spatial accuracy of the output geocodes, these methods provide the quantitative basis for describing the spatial uncertainty of an output geocode based on the overall spatial uncertainty inherent in the geographic reference feature used to create it, the alternatives that were not selected or were used in concert, and the spatial, topological, and relative uncertainty relationships between them. In this fashion, the present work succeeds in extending the geocode quality metadata reported along with geocoded results such that data consumers can begin to have a sense of the relative levels of uncertainty present in the data underlying their scientific studies.

6.4 Tie-Handling

Our detailed analysis of the sources and types of ambiguous geocode results is the first of its kind to investigate the specific causes of ambiguity in terms of the address attributes at the root of the problem as well as the specific relationships between the attributes of candidate features (**Chapter 5**). This investigation allowed us to develop a novel approach to refining the spatial uncertainty of the output geocode based on the convex hull of all available ambiguous candidate reference features – dynamic feature composition.

In addition, this analysis led us to develop an approach to intelligently break ties using a set of attribute relationship rules and a spatial neighborhood-based approach (**Chapter 5**). These methods use spatial reasoning techniques that look for relationships between ambiguous candidate features and trends in the local neighborhood around the ambiguous candidate features to determine which is the most likely location given what is known about the attributes of the input address data.

Our analysis of these novel methods showed that they consistently choose the correct ambiguous candidate feature 82% of the time, and nearly 98% of the time in specific cases effectively reducing the spatial error on the order of tens to hundreds of meters which was shown to be statistically significant. Perhaps even more importantly, using either the dynamic feature composition centroid or the neighborhood-based spatial reasoning approach dramatically reduces the overall spatial uncertainty associated with the geocode output.

In conjunction with our analysis of the occurrence of ambiguous cases which showed that these cases occur quite frequently even in supposedly super-high-quality input data such as the Los Angeles County address file (Los Angeles County Chief Information Office 2009) and when geocoding with supposedly super-high-quality reference data layers such as the ESRI StreetMap North America (Environmental Systems Research Institute 2009d), our approaches to tie-handling can be seen as an important contribution. These approaches make great strides toward addressing two fundamental issues: (1) dramatically reducing the spatial uncertainty associated with geocodes that would have otherwise reverted to lower levels of accuracy following a standard NAACCR-type hierarchy approach; and (2) reducing spatial error by picking the correct ambiguous candidate feature with greater probability than simply flipping-a-coin.

REFERENCES

Apparicio, P., M. Abdelmajid, M. Riva, and R. Shearmur. 2008. Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. *International Journal of Health Geographics* 7 (7).

Arbia, G., D. Griffith, and R. P. Haining. 1998. Error propagation modeling in raster GIS: overlay operations. *International Journal of Geographical Information Science* 12 (2):145-167.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18 (5):497-525.

Bakshi, R., C. A. Knoblock, and S. Thakkar. 2004. Exploiting online sources to accurately geocode addresses. In *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, 194-203. Washington, DC: ACM Press.

Beal, J. R. 2003. Contextual geolocation, a specialized application for improving indoor location awareness in wireless local area networks. Paper read at MICS2003: The 36th Annual Midwest Instruction and Computing Symposium, at Duluth, Minnesota.

Bell, B. S., R. E. Hoskins, L. W. Pickle, and D. Wartenberg. 2006. Current practices in spatial analysis of cancer data: Mapping health statistics to inform policymakers and the public. *International Journal of Health Geographics* 5 (49).

Beyer, K. M. M., A. F. Schultz, and G. Rushton. 2008. Using ZIP codes as geocodes in cancer research. In *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice,* eds. G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West and D. L. Zimmerman, 37-68. Boca Raton, FL: CRC Press.

Block, R. ed. 1995. *Geocoding of crime incidents using the 1990 TIGER file: The Chicago example*. Washington, DC: Police Executive Research Forum.

Blondin, N., D. J. Baumgardner, G. E. Moore, and L. T. Glickman. 2007. Blastomycosis in indoor cats: Suburban Chicago, Illinois, USA. *Mycopathologia* 163 (2):59-66.

Bonner, M. R., D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim. 2003. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 14 (4):408-411.

Boscoe, F. P. 2008. The science and art of geocoding. In *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, eds. G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West and D. L. Zimmerman, 95-109. Boca Raton, FL: CRC Press.

Boscoe, F. P., C. L. Kielb, M. J. Schymura, and T. M. Bolani. 2002. Assessing and improving census track completeness. *Journal of Registry Management* 29 (4):117-120.

Boscoe, F. P., M. H. Ward, and P. Reynolds. 2004. Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. *International Journal of Health Geographics* 3 (28).

Boulos, M. N. K. 2004. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. *International Journal of Health Geographics* 3 (1).

Brody, J. G., D. J. Vorhees, S. J. Melly, S. R. Swedis, P. J. Drivas, and R. A. Rudel. 2002. Using GIS and historical records to reconstruct residential exposure to large-scale pesticide application. *Journal of Exposure Analysis and Environmental Epidemiology* 12 (1):64-80.

Broome, F. R., and L. S. Godwin. 2003. Partnering for the People: Improving the US Census Bureau's MAF/TIGER Database. *Photogrammetric engineering and remote sensing* 69 (10):1119-1126.

Cantador, I., M. Szomszor, H. Alani, M. Fernandez, and P. Castells. 2008. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proceedings of the Workshop on Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb 2008)*. Tenerife, Spain.

Cayo, M. R., and T. O. Talbot. 2003. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2 (10).

Chainey, S., and J. H. Ratcliffe. 2005. Mapping crime with local and community data. In *GIS and Crime Mapping*, 205-212. Chichester, UK: Wiley.

Chao, A., and T. J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* 10 (4):429-443.

Chen, C.-C., C. A. Knoblock, C. Shahabi, and S. Thakkar. 2003. Building finder: a system to automatically annotate buildings in satellite imagery. In *NG2I '03: Proceedings of the International Workshop on Next Generation Geospatial Information*, ed. P. Agouris. Cambridge, MA.

Chen, C.-C., C. A. Knoblock, C. Shahabi, S. Thakkar, and Y.-Y. Chiang. 2004. Automatically and accurately conflating orthoimagery and street maps. In *ACMGIS '04: Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems*, eds. D. Pfoser, I. F. Cruz and M. Ronthaler, 47-56. Washington D.C.

Christen, P., and T. Churches. 2005. A probabilistic deduplication, record linkage and geocoding system. In *Proceedings of the Australian Research Council Health Data Mining Workshop*. Canberra, AU.

Christen, P., T. Churches, and A. Willmore. 2004. A probabilistic geocoding system based on a national address file. In *Proceedings of the Australasian Data Mining Conference*. Cairns, AU.

Chung, K., D. H. Yang, and R. Bell. 2004. Health and GIS: Toward spatial statistical analyses. *Journal of Medical Systems* 28 (4):349-360.

Church, K. W., W. A. Gale, P. Hanks, and D. M. Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, ed. U. Zernik, 115-164. Hillsdale, NJ: Lawrence Erlbaum Associates.

Churches, T., P. Christen, K. Lim, and J. X. Zhu. 2002. Preparation of name and address data for record linkage using Hidden Markov Models. *Medical Informatics and Decision Making* 2 (9).

Cirasella, J. 2007. You and Me and Google Makes Three: Welcoming Google into the Reference Interview. *Library Philosophy and Practice* 9 (2):1-8.

Collins, S. E., R. P. Haining, I. R. Bowns, D. J. Crofts, T. S. Williams, A. S. Rigby, and D. M. Hall. 1998. Errors in postcode to enumeration district mapping and their effect on small area analyses of health data. *Journal of Public Health Medicine* 20 (3):325-330.

Costello, S., M. Cockburn, J. Bronstein, X. Zhang, and B. Ritz. 2009. Parkinson disease and residential exposure to maneb and paraquat from agricultural applications in the Central Valley of California. *American Journal of Epidemiology* 169 (8):919–926.

Cressie, N., and J. Kornak. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* 18 (4):436-456.

Croner, C. M., J. Sperling, and F. R. Broome. 1996. Geographic information systems (GIS): new perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15 (18):1961-1977.

Davis Jr., C. A. 1993. Address base creation using raster/vector integration. In *Proceedings of the URISA 1993 Annual Conference*. Atlanta, GA.

Davis Jr., C. A., and F. T. Fonseca. 2007. Assessing the certainty of locations produced by an address geocoding system. *GeoInformatica* 11 (1):103-129.

Davis Jr., C. A., F. T. Fonseca, and K. A. V. Borges. 2003. A flexible addressing system for approximate geocoding. In *Proceedings of the 5th Brazilian Symposium on Geoinformatics*.

Dearwent, S. M., R. R. Jacobs, and J. B. Halbert. 2001. Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis Environmental Epidemiology* 11 (4):329-334.

Diez-Roux, A. V., S. S. Merkin, D. Arnett, L. Chambless, M. Massing, F. J. Nieto, P. Sorlie, M. Szklo, H. A. Tyroler, and R. L. Watson. 2001. Neighborhood of residence and incidence of coronary heart disease. *New England Journal of Medicine* 345 (2):99-106.

Drummond, W. J. 1995. Address matching: GIS technology for mapping human activity patterns. *Journal of the American Planning Association* 61 (2):240-251.

Dueker, K. J. 1974. Urban geocoding. *Annals of the Association of American Geographers* 64 (2):318-325.

Durr, P. A., and A. E. A. Froggatt. 2002. How best to georeference farms? A case study from Cornwall, England. *Preventive Veterinary Medicine* 56:51-62.

Egenhofer, M. 1991. Reasoning about binary topological relations. *Lecture Notes in Computer Science* 523:143–160.

Eichelberger, P. 1993. The importance of addresses: The locus of GIS. In *Proceedings of the URISA 1993 Annual Conference*, 212-213. Atlanta, GA.

Eldridge, D., and J. P. Jones. 1991. Warped Space: A Geography of Distance Decay. *Professional Geographer* 43 (4):500-511.

Environmental Systems Research Institute. 2009a. *ArcGIS 9 Geocoding Rule Base Developer Guide*. Redlands, CA: Environmental Systems Research Institute.

———. *ArcGIS: A Complete Integrated System*. Environmental Systems Research Institute 2009b. Available from http://www.esri.com/software/arcgis/.

———. *Building a composite address locator*. ESRI Press 2009c. Available from http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Building_a_composite_address_locator.

———. *ESRI StreetMap Premium*. ESRI Press 2009d. Available from http://www.esri.com/data/streetmap/index.html.

Federal Geographic Data Committee. 2006. *Homeland Security and Geographic Information Systems – How GIS and mapping technology can save lives and protect property in post-September 11th America*. Reston, VA: Federal Geographic Data Committee.

Fonda-Bonardi, P. 1994. House Numbering Systems in Los Angeles. In *Proceedings of the GIS/LIS '94 Annual Conference and Exposition*, 322–331. Phoenix, AZ.

Frew, J., M. Freeston, N. Freitas, L. L. Hill, G. Janee, K. Lovette, R. Nideffer, T. R. Smith, and Q. Zheng. 1998. The Alexandria digital library architecture. In *ECDL '98: Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, eds. C. Nikalaou and C. Stephanidis, 61-73. Heraklion, Crete, Greece: Springer.

Frizzelle, B., K. Evenson, D. Rodriguez, and B. Laraia. 2009. The importance of accurate road data for spatial applications in public health: Customizing a road network. *International Journal of Health Geographics* 8 (24).

Fulcomer, M. C., M. M. Bastardi, H. Raza, M. Duffy, E. Dufficy, and M. M. Sass. 1998. Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996. In *Proceedings of the 1998 Geographic Information Systems in Public Health Conference,* eds. R. C. Williams, M. M. Howie, C. V. Lee and W. D. Henriques, 547-560. San Diego, CA.

Gabrosek, J., and N. Cressie. 2002. The effect on attribute prediction on location uncertainty in spatial data. *Geographical Analysis* 34:262-285.

Gaffney, S. H., F. C. Curriero, P. T. Strickland, G. E. Glass, K. J. Helzlsouer, and P. N. Breysse. 2005. Influence of geographic location in modeling blood pesticide levels in a community surrounding a U.S. Environmental Protection Agency Superfund Site. *Environmental Health Perspectives* 113 (12):1712-1716.

Gatrell, A. C. 1989. On the spatial representation and accuracy of address-based data in the United Kingdom. *International Journal of Geographical Information Science* 3 (4):335-348.

Geolytica Inc. 2010. *Geocoder.ca: Geocoding for North America - USA and Canada* 2010 [cited March 6 2010]. Available from http://geocoder.ca.

Geronimus, A. T., and J. Bound. 1998. Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology* 148 (5):475-486.

———. 1999a. RE: Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology* 150 (8):894-896.

———. 1999b. RE: Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology* 150 (9):997-999.

Geronimus, A. T., J. Bound, and L. J. Neidert. 1995. On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. In *Technical Working Paper*. Cambridge, MA: US National Bureau of Economic Research.

Gilboa, S. M., P. Mendola, A. F. Olshan, C. Harness, D. Loomis, P. H. Langlois, D. A. Savitz, and A. H. Herring. 2006. Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research* 101 (2):256-262.

Goldberg, D. W. 2008. *A Geocoding Best Practices Guide*. Springfield, IL: North American Association of Central Cancer Registries.

———. 2009. *Free geocoding, address validation, shortest path, and data capture*. CA 20092009]. Available from http://webgis.usc.edu.

Goldberg, D. W., and J. P. Wilson. 2009. *The USC WebGIS Geocoding Platform*. University of Southern California 2009 [cited August 4 2009]. Available from https://webgis.usc.edu.

———. 2010. *The USC WebGIS Geocoding Platform*. University of Southern California 2010 [cited March 6 2010]. Available from https://webgis.usc.edu.

Goldberg, D. W., J. P. Wilson, and C. A. Knoblock. 2007. From text to geographic coordinates: The current state of geocoding. *Urisa Journal* 19 (1):33-47.

———. 2008. Exploring the use of gazetteers and geocoders for the analysis and interpretation of a dynamically changing world. In *Understanding Dynamics of Geographic Domains*, eds. K. S. Hornsby and M. Yuan, 51-76. Boca Raton, FL: CRC Press.

Goldberg, D. W., J. P. Wilson, C. A. Knoblock, B. Ritz, and M. G. Cockburn. 2008. An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics* 7 (60).

Google Inc. *Maps API Reference - Google Maps API - Google Code* Google Inc. 2009. Available from http://code.google.com/apis/maps/documentation/reference.html.

Gottesfeld-Brown, L. 1992. A survey of image registration techniques. *ACM computing surveys* 24 (4):325-376.

Gregorio, D. I., E. Cromley, R. Mrozinski, and S. J. Walsh. 1999. Subject loss in spatial analysis of breast cancer. *Health & Place* 5 (2):173-177.

Gregorio, D. I., L. M. S. DeChello, H. , and M. Kulldorff. 2005. Lumping or splitting: seeking the preferred areal unit for health geography studies. *International Journal of Health Geographics* 4 (6).

Griffin, D. H., J. M. Pausche, E. B. Rivers, A. L. Tillman, and T. J. B. 1990. Improving the coverage of addresses in the 1990 census: preliminary results. In *Proceedings of the American Statistical Association Survey Research Methods Section*, 541-546. Anaheim, CA.

Grubesic, T. H. 2008. Zip codes and spatial analysis: Problems and prospects. *Socio-Economic Planning Sciences* 42 (2):129-149.

Gupta, R. S., X. Zhang, L. K. Sharp, J. J. Shannon, and K. B. Weiss. 2009. The protective effect of community factors on childhood asthma. *Journal of Allergy and Clinical Immunology* 123 (6):1297-1304.

Han, D., P. A. Rogerson, M. R. Bonner, J. Nie, J. E. Vena, P. Muti, M. Trevisan, and J. L. Freudenheim. 2005. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. *International Journal of Health Geographics* 4 (9).

Hanley, N., F. Schläpferb, and J. Spurgeonc. 2003. Aggregating the benefits of environmental improvements: distance-decay functions for use and non-use values. *Journal of Environmental Management* 68 (3):297-304.

Haspel, M., and H. G. Knotts. 2005. Location, location, location: Precinct placement and the costs of voting. *The Journal of Politics* 67 (2):560-573.

Hay, G., and G. Castilla. 2008. Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline. In *Object-based image analysis*, eds. T. Blaschke, S. Lang and G. Hay, 75-89. Berlin: Springer.

Henry, K. A., and F. P. Boscoe. 2008. Estimating the accuracy of geographical imputation. *International Journal of Health Geographics* 7 (3).

Hibbert, J., A. Liese, A. Lawson, D. Porter, R. Puett, D. Standiford, L. Liu, and D. Dabelea. 2009. Evaluating geographic imputation approaches for zip code level data: an application to a study of pediatric diabetes *International Journal of Health Geographics* 8 (54).

Higgs, G., and W. Richards. 2002. The use of geographical information systems in examining variations in sociodemographic profiles of dental practice catchments: A case study of a Swansea practice. *Primary Dental Care* 9 (2):63-69.

Higgs, G., and D. J. Martin. 1995. The address data dilemma part 1: Is the introduction of address-point the key to every door in Britain? *Mapping Awareness* 8:26–28.

Hill, L. L. 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. In *ECDL '00: research and advanced technology for digital libraries. 4th European Conference*, eds. J. L. Borbinha and T. Baker, 280-290. Lisbon, Portugal: Springer.

Hill, L. L., and Q. Zheng. 1999. Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, 57-69. Washington, D.C.

Hill, L. L., and M. F. Goodchild. 2000. *Digital Gazetteer Information Exchange (DGIE) Final Report of Workshop*.

Hofferkamp, J., and L. Havener. 2008. *Standards for Cancer Registries: Data Standards and Data Dictionary*. 12th ed. Springfield, IL: North American Association of Central Cancer Registries.

Hurley, S. E., T. M. Saunders, R. Nivas, A. Hertz, and P. Reynolds. 2003. Post office box addresses: A challenge for geographic information system-based studies. *Epidemiology* 14:386-391.

Hutchinson, M., and B. Veenendall. 2005a. Towards a framework for intelligent geocoding. In *Proceedings of the Spatial Intelligence, Innovation and Praxis: The National Biennial Conference of the Spatial Sciences Institute*. Melbourne, AU.

———. 2005b. Towards using intelligence to move from geocoding to geolocating. In *Proceedings of the 7th Annual URISA GIS in Addressing Conference*.

Jaro, M. 1984. Record linkage research and the calibration of record linkage algorithms. In *Statistical Research Division Report Series* Washington, DC: United States Census Bureau.

———. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 89:414-420.

Johnson, S. D. 1998a. Address matching with commercial spatial data: Part 1. *Business Geographics* March:24-32.

———. 1998b. Address matching with stand-alone geocoding engines: Part 2. *Business Geographics* April (30-36).

Karimi, H. A., M. Durcik, and W. Rasdorf. 2004. Evaluation of uncertainties associated with geocoding techniques. *Journal of Computer-Aided Civil and Infrastructure Engineering* 19 (3):170-185.

Kennedy, T. C., J. G. Brody, and J. N. Gardner. 2003. Modeling historical environmental exposures using GIS: implications for disease surveillance. In *Proceedings of the 2003 ESRI Health GIS Conference*. Arlington, Virginia.

Krieger, N. 1992. Overcoming the absence of socioeconomic data in medical records: Validation and application of a Census-based methodology. *American Journal of Public Health* 82 (5):703-710.

———. 2003. Place, space, and health: GIS and epidemiology. *Epidemiology* 14 (4):384-385.

Krieger, N., J. T. Chen, P. D. Waterman, D. H. Rehkopf, and S. V. Subramanian. 2005. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: The public health disparities geocoding project. *American Journal of Public Health* 95 (2):312-323.

Krieger, N., J. T. Chen, P. D. Waterman, M. J. Soobader, S. V. Subramanian, and R. Carson. 2002. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? *American Journal of Epidemiology* 156 (5):471-482.

Krieger, N., and D. Gordon. 1999. RE: Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology* 150 (8):894-896.

Krieger, N., P. Waterman, K. Lemieux, S. Zierler, and J. W. Hogan. 2001. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American Journal of Public Health* 91 (7):1114-1116.

Krieger, N., P. D. Waterman, J. T. Chen, M. J. Soobader, and S. V. Subramanian. 2003. Monitoring socioeconomic inequalities in sexually transmitted infections, Tuberculosis, and violence: Geocoding and choice of area-based socioeconomic measures. *Public Health Reports* 118 (3):240-260.

Krieger, N., P. D. Waterman, J. T. Chen, M. J. Soobader, S. V. Subramanian, and R. Carson. 2002. ZIP code caveat: Bias due to spatiotemporal mismatches between ZIP codes and US census-defined areas: The public health disparities geocoding project. *American Journal of Public Health* 92 (7):1100-1102.

Kwok, R. K., and B. C. Yankaskas. 2001. The use of census data for determining race and education as SES indicators: a validation study. *Annals of Epidemiology* 11 (3):171-177.

Laender, A. H. F., K. A. V. Borges, J. C. P. Carvalho, C. B. Medeiros, A. S. da Silva, and C. A. Davis Jr. 2005. Integrating Web data and geographic knowledge into spatial databases. In *Spatial databases: Technologies, techniques and trends*, eds. Y. Manalopoulos and A. N. Papadapoulos, 23-47. Hershey, PA: Idea Group Inc.

Lee, B. 2009. Spatial Pattern of Uncertainties: An Accuracy Assessment of the TIGER Files. *Journal of Geography and Geology* 1 (2):2-12.

Lee, J. 2004. 3D GIS for geo-coding human activity in micro-scale urban environments. In *Third International Conference on Geographic Information Science, GIScience 2004*, eds. M. J. Egenhofer, C. Freksa and H. J. Miller, 162-178. College Park, MD.

Levesque, M. 2003. West Virginia state-wide addressing and mapping project. In *Proceedings of the 5th Annual URISA Street Smart and Address Savvy Conference*.

Levine, N., and K. E. Kim. 1998. The spatial location of motor vehicle accidents: A methodology for geocoding intersections. *Computers, Environment and Urban Systems* 22 (6):557-576.

Li, H., R. K. Srihari, C. Niu, and W. Li. 2003. InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, 39-44. Edmonton, Alberta: Association for Computational Linguistics.

Li, Y. 2007. Probabilistic Toponym Resolution and Geographic Indexing and Querying, Department of Computer Science and Software Engineering, The University of Melbourne, Melbourne, AU.

Liang, J. 2010. *JGeocoder - Free Java Geocoder* 2008 [cited March 6 2010]. Available from http://jgeocoder.sourceforge.net/.

Lind, M. 2001. Developing a system of public addresses as a language for location dependent information. In *Proceedings of the 2001 URISA Annual Conference*. Long Beach, CA: URISA.

Lixin, Y. 1996. Development and evaluation of a framework for assessing the efficiency and accuracy of street address geocoding strategies, Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York, Albany, NY.

Locative Technologies. 2009. *Geocoder.us/.NET - Find the latitude & longitude of any US address - for free*. Sierraville CA: Locative Technologies.

Lockyer, B. 2005. Legal Opinion 04-1105: Office of the Attorney General of the State of California.

Lookingbill, A. 2009. *Google LatLong: Your world, your map*. Mountain View, CA: Google Inc.

Los Angeles County Assessor's Office. 2009. *GIS ready map base data*. Los Angeles, CA: Los Angeles County Assessor's Office.

Los Angeles County Chief Information Office. 2009. *Los Angeles County Address File*. Los Angeles, CA: Los Angeles Chief Information Office.

Macintyre, S., L. Macdonald, and A. Ellaway. 2007. Lack of agreement between measured and self-reported distance from public green parks in Glasgow, Scotland. *International Journal of Behavioral Nutrition and Physical Activity* 5 (26).

Manifold Net Ltd. 2009. Manifold System Release 8 User Manual. Carson City, NV: Manifold Net Ltd.

Martin, D. J. 1998. Optimizing census geography: The separation of collection and output geographies. *International Journal of Geographical Information Science* 12 (7):673-685.

Martin, D. J., and G. Higgs. 1996. Georeferencing people and places: A comparison of detailed datasets. In *Innovations in GIS 3: Selected papers from the Third National Conference on GIS Research UK (Gisruk)*, ed. D. Parker, 37-47. Canterbury, UK: Taylor and Francis.

———. 1999. Georeferencing people and places: A comparison of detailed datasets. In *Proceedings of the 3rd National Conference on GIS Research UK (Innovations in GIS 3),* ed. D. Parker, 37-47. Canterbury, UK: Taylor & Francis.

Mazumdar, S., G. Rushton, B. J. Smith, D. L. Zimmerman, and K. J. Donham. 2008. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics* 7 (13).

McElroy, J. A., P. L. Remington, A. Trentham-Dietz, S. A. Roberts, and P. A. Newcomber. 2003. Geocoding addresses from a large population based study: Lessons learned. *Epidemiology* 14 (4):399-407.

Mckercher, B., and A. A. Lew. 2003. Distance decay and the impact of effective tourism exclusion zones on international travel flows. *Journal of Travel Research* 42 (2):159.

Meliker, J., and G. Jacquez. 2007. Space–time clustering of case–control data with residential histories: insights into empirical induction periods, age-specific susceptibility, and calendar year-specific effects *Stochastic Environmental Research and Risk Assessment* 21 (5):625-634.

Michalowski, M., C. Knoblock , K. Bayer, and B. Choueiry. 2007. Exploiting Automatically Inferred Constraint Models for Building Identification in Satellite Imagery. Paper read at Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems (ACMGIS 07).

Microsoft Corporation. *Bing Maps Web Services SDK, Version 1.0*. Microsoft Corporation 2009a. Available from http://msdn.microsoft.com/en-us/library/cc980922.aspx.

———. *OGC Methods on Geography Instances*. Microsoft Corporation 2009b. Available from http://msdn.microsoft.com/en-us/library/bb933917.aspx.

Murphy, J., and R. Armitage. 2005. Merging the modeled and working address database: A question of dynamics and data quality. In *Proceedings of GIS Ireland 2005*. Dublin, IE.

Navarro, G. 2001. A guided tour to approximate string matching. *ACM computing surveys* 33 (1):31-88.

Nicoara, G. 2005. Exploring the geocoding process: A municipal case study using crime data, University of Texas at Dallas, Dallas, TX.

North American Association of Central Cancer Registries. 2009. *Geocoded National Provider Identification file*. Springfield, IL: North American Association of Central Cancer Registries.

Nuckols, J. R., M. H. Ward, and L. Jarup. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* 112 (9):1007-1015.

O’Reagan, R. T., and A. Saalfeld. 1987. Geocoding theory and practice at the Bureau of the Census. In *Statistical Research Report*. Washington, DC: United States Bureau of Census.

Oliver, M. N., K. A. Matthews, M. Siadaty, F. R. Hauck, and L. W. Pickle. 2005. Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics* 4 (29).

———. 2009. Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics* 4 (29).

Olligschlaeger, A. M. 1998. Artificial neural networks and crime mapping. In *Crime mapping and crime prevention*, eds. D. Weisburd and T. McEwen, 313-347. Monsey, NY: Criminal Justice Press.

Omer, S. B., K. S. Enger, L. H. Moulton, N. A. Halsey, S. Stokley, and D. A. Salmon. 2008. Geographic Clustering of Nonmedical Exemptions to School Immunization Requirements and Associations With Geographic Clustering of Pertussis. *American Journal of Epidemiology* 168 (12):1389-1396.

Openshaw, S. 1989. Learning to live with errors in spatial databases. In *Accuracy of spatial databases*, eds. M. F. Goodchild and S. Gopal, 263-276. Bristol, PA: Taylor and Francis.

Oppong, J. R. 1999. Data problems in GIS and health. In *Proceedings of Health and Environment Workshop 4: Health Research Methods and Data*. Turku, Finland.

Ordnance Survey. 2010. *ADDRESS-POINT: Ordnance Survey's map dataset of all postal addresses in Great Britain* 2010 [cited March 6 2010]. Available from http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint.

Paull, D. 2003. A geocoded national address file for Australia: The G-NAF what, why, who and when?

Pendleton, C. 2008. *New Feature Release of Live Search Maps!* Redmond, Wa: Microsoft Corporation.

Pezzoli, K., R. Tukey, H. Sarabia, I. Zaslavsky, M. L. Miranda, W. A. Suk, A. Lin, and M. Ellisman. 2007. The NIEHS Environmental Health Sciences Data Resource Portal: Placing Advanced Technologies in Service to Vulnerable Communities. *Environmental Health Perspectives* 115 (4):564 - 571.

Pohl, C., and J. L. Van Genderen. 1998. Review article: Multisensor image fusion. In Remote sensing: concepts, methods and applications. *International Journal of Remote Sensing* 19 (5):823-854.

POIfriend Inc. *POIfriend.com: GPS POI Group: Best Western Hotels* 2009. Available from http://poifriend.com/poigroup.php?poigroup_id=8870.

Poulin, R. 2003. The decay of similarity with geographical distance in parasite communities of vertebrate hosts. *Journal of Biogeography* 30 (10):1609-1615.

Ratcliffe, J. H. 2001. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science* 15 (5):473-485.

———. 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science* 18 (1):61-72.

Reimers, J. 2009. *ESRI Shapefile to KML Converter*.

Ries, A. V., C. C. Voorhees, K. M. Roche, J. Gittelsohn, A. F. Yan, and N. M. Astone. 2009. A Quantitative Examination of Park Characteristics Related to Park Use and Physical Activity Among Urban Youth. *Journal of Adolescent Health* 45 (3):S64-S70.

Ritz, B., A. Manthripragada, S. Costello, S. Lincoln, M. Farrer, M. Cockburn, and J. Bronstein. 2009. Dopamine transporter genetic variants and pesticides in Parkinson's disease. *Environmental Health Perspectives* 117 (6):964-969.

Rose, K. M., J. L. Wood, S. Knowles, R. A. Pollitt, E. A. Whitsel, A. V. Diez Roux, D. Yoon, and G. Heiss. 2004. Historical measures of social context in life course studies: retrospective linkage of addresses to decennial censuses. *International Journal of Health Geographics* 3 (27).

Ruiz, M., E. Walker, E. Foster, L. Haramis, and U. Kitron. 2007. Association of West Nile virus illness and urban landscapes in Chicago and Detroit. *International Journal of Health Geographics* 6 (1).

Rull, R. P., R. Gunier, J. V. Behren, A. Hertz, V. Crouse, P. A. Buffler, and P. Reynolds. 2009. Residential proximity to agricultural pesticide applications and childhood acute lymphoblastic leukemia. *Environmental Research* 109 (7):891-899.

Rushton, G., M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West, and D. L. Zimmerman. 2006. Geocoding in cancer research: A review. *American Journal of Preventive Medicine* 30 (2):S16-S24.

Schootman, M., D. Jeffe, E. Kinman, G. Higgs, and J. Jackson-Thompson. 2004. Evaluating the utility and accuracy of a reverse telephone directory to identify the location of survey respondents. *Annals of Epidemiology* 15 (2):160-166.

Schootman, M., D. A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo, and G. Higgs. 2007. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology* 17 (6):379-387.

Sheehan, T. J., S. T. Gershman, L. MacDougal, R. A. Danley, M. Mroszczyk, A. M. Sorensen, and M. Kulldorff. 2000. Geographic surveillance of breast cancer screening by tracts, towns and zip codes. *Journal of Public Health Management Practices* 6:48-57.

Shi, X. 2007. Evaluating the uncertainty caused by P.O. box addresses in environmental health studies: A restricted Monte Carlo approach. *International Journal of Geographical Information Science* 21 (3):325-340.

Smith, G. D., Y. Ben-Shlomo, and C. Hart. 1999. Re: Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology* 150 (9):996-997.

Soobader, M., F. B. LeClere, W. Hadden, and B. Maury. 2001. Using aggregate geographic data to proxy individual socioeconomic status: does size matter? *American Journal of Public Health* 91 (4):632-636.

Stage, D., and N. von Meyer. 2005. An assessment of parcel data in the United States 2005 Survey results: Federal Geographic Data Committee Subcommittee on Cadastral Data.

Stevenson, M. A., J. Wilesmith, J. Ryan, R. Morris, A. Lawson, D. Pfeiffer, and D. Lin. 2000. Descriptive spatial analysis of the epidemic of bovine spongiform encephalopathy in Great Britain to June 1997. *The Veterinary Record* 147 (14):379-384.

Stewart, J. 1995. *Calculus - Early Transcendentals*. 3rd ed. Pacific Grove, CA: Brooks/Cole.

Strickland, M. J., C. Siffel, B. R. Gardner, A. K. Berzen, and A. Correa. 2007. Quantifying geocode location error using GIS methods. *Environmental Health* 6 (10).

Suarez, L., J. D. Brender, P. H. Langlois, F. B. Zhan, and K. Moody. 2007. Maternal exposures to hazardous waste sites and industrial facilities and risk of neural tube defects in offspring. *Annals of Epidemiology* 17 (10):772-777.

Sui, D. Z. 2007. Geographic Information Systems and Medical Geography: Toward a New Synergy. *Geography Compass* 1 (3):556–582.

Swift, J. N., D. W. Goldberg, and J. P. Wilson. 2008. Geocoding Best Practices: Review of Eight Commonly Used Geocoding Systems. Los Angeles, CA: University of Southern California GIS Research Laboratory.

Tobler, W. 1972. Geocoding theory. In *Proceedings of the National Geocoding Conference*. Washington DC: U.S. Department of Transportation.

Toutin, T. 2004. Review article: Geometric processing of remote sensing images: models, algorithms and methods. *International Journal of Remote Sensing* 25 (10):1893-1924.

U.S. Census Bureau. *Cartographic boundary files*. U.S. Census Bureau 2009a. Available from http://www.census.gov/geo/www/cob/.

———. *U.S. Census Bureau TIGER/Line*. U.S. Census Bureau 2009b. Available from http://www.census.gov/geo/www/tiger.

U.S. Department of Health and Human Services. 2000. *Healthy people 2010: Understanding and improving health*. 2nd ed. Washington, D.C.: U.S. Government Printing Office.

———. 2004. *HIPAA Administrative Simplification: Standard Unique Health Identifier for Health Care Providers; Final Rule*. Washington, D.C.: U.S. Department of Health and Human Services.

U.S. Postal Service. 2009a. *CASS Technical Guide*. Washington, DC: U.S. Postal Service.

———. 2009b. *Publication 28 - Postal Addressing Standards*. Washington, DC: U.S. Postal Service.

———. 2009c. *Topological Integrated Geographic Encoding and Reference/ZIP + 4 File*. Washington, DC: U.S. Postal Service.

United Nations Economic Commission. 2010. *A functional addressing system for Africa: A discussion paper* 2005 [cited March 6 2010]. Available from http://geoinfo.uneca.org/Docs/Situs Addressing background paper-Draft.pdf.

Urban and Regional Information Systems Association. 2009. *Draft Street Address Data Standard*. Park Ridge, IL: Urban and Regional Information Systems Association.

Veregin, H. 1999. Data quality parameters. In *Geographical information systems*, ed. M. F. G. In P. A. Longley, D. J. Maguire, and D. W. Rhind, 177-189. New York, NY Wiley.

Vine, M. F., D. Degnan, and H. C. 1998. Geographic information systems: Their use in environmental epidemiologic research. *Journal of Environmental Health* 61:7-16.

Wagner, S. E., J. B. Burch, J. Hussey, T. Temples, S. Bolick-Aldrich, C. Mosley-Broughton, Y. Liu, and J. R. Hebert. 2009. Soil zinc content, groundwater usage, and prostate cancer incidence in South Carolina. *Cancer Causes and Control* 20 (3):345.

Walls, M. D. 2003. Is consistency in address assignment still needed? . In *Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference*. Providence, RI.

Wang, J., and N. Ge. 2006. Automatic feature thesaurus enrichment: extracting generic terms from digital gazetteer. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 326-333. Chapel Hill, NC: ACM.

Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, J. S. Colt, and P. Hartge. 2005. Positional accuracy of two methods of geocoding. *Epidemiology* 16 (4):542-547.

Warden, C. R. 2008. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. *International Journal of Health Geographics* 7 (51).

Werner, P. A. 1974. National geocoding. *Annals of the Association of American Geographers* 64 (2):310-317.

Whitsel, E. A., P. M. Quibrera, R. L. Smith, D. J. Catellier, D. Liao, A. C. Henley, and G. Heiss. 2006. Accuracy of commercial geocoding: Assessment and implications. *Epidemiologic Perspectives & Innovations* 3 (8).

Wieczorek, J., Q. Guo, and R. J. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18 (8):745-767.

Wrobel, L., J. K. Whittington, C. Pujol, S.-H. Oh, M. O. Ruiz, M. A. Pfaller, D. J. Diekema, D. R. Soll, and L. L. Hoyer. 2008. Molecular Phylogenetic Analysis of a Geographically and Temporally Matched Collection of Candida albicans Isolates from Humans and Non-Migratory Wildlife in Central Illinois. *Eukaryotic Cell*.

Wu, J., T. H. Funk, F. W. Lurmann, and A. M. Winer. 2005. Improving spatial accuracy of roadway networks and geocoded addresses. *Transactions in GIS* 9 (4):585-601.

Yahoo! Inc. *Yahoo! Maps Terms Of Use*. Yahoo, Inc. 2009a. Available from http://info.yahoo.com/legal/us/yahoo/maps/mapstou/mapstou-278.html.

———. *Yahoo! Maps Web Services - Geocoding API*. Yahoo, Inc. 2009b. Available from http://developer.yahoo.com/maps/rest/V1/geocode.html.

Yang, D.-H., L. M. Bilaver, O. Hayes, and R. Goerge. 2004. Improving Geocoding Practices: Evaluation of Geocoding Tools. *Journal of Medical Systems* 28 (4):361-370.

Zandbergen, P. A. 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health* 7 (37).

———. 2008a. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32:214-232.

———. 2008b. Positional accuracy of spatial data: Non-normal distributions and a critique of the National Standard for Data Accuracy. *Transactions in GIS* 12 (1):103-130.

Zandbergen, P. A., and J. W. Green. 2007. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives* 115 (9):1363-1370.

Zhan, F. B., J. D. Brender, I. De Lima, L. Suarez, and P. H. Langlois. 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology* 16 (11):842-849.

Zimmerman, D., and J. Li. 2010. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies *International Journal of Health Geographics* 9 (10).

Zimmerman, D. L. 2008. Statistical methods for incompletely and incorrectly geocoded cancer data. In *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, eds. G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West and D. L. Zimmerman, 165-180. Boca Raton, FL: CRC Press.

Zimmerman, D. L., X. Fang, S. Mazumdar, and G. Rushton. 2007. Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics* 6 (1).