A Geocoding Best Practices Guide

By Daniel W. Goldberg
University of Southern California
GIS Research Laboratory

November 2008

SPONSORING ORGANIZATIONS:
Canadian Association of Provincial Cancer Agencies
Canadian Partnership Against Cancer
Centers for Disease Control and Prevention
College of American Pathologists
National Cancer Institute
National Cancer Registrars Association
Public Health Agency of Canada

SPONSORS WITH DISTINCTION:
American Cancer Society
American College of Surgeons
American Joint Committee on Cancer

NAACCR
North American Association of Central Cancer Registries, Inc.

A GEOCODING BEST PRACTICES GUIDE

SUBMITTED TO

THE NORTH AMERICAN ASSOCIATION OF CENTRAL CANCER REGISTRIES
NOVEMBER 10, 2008

BY
DANIEL W. GOLDBERG
UNIVERSITY OF SOUTHERN CALIFORNIA
GIS RESEARCH LABORATORY

This page is left blank intentionally.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF BEST PRACTICES

LIST OF ACRONYMS

0D	Zero Dimensional
1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
4D	Four Dimensional
CI	Confidence Interval
CBG	U.S. Census Bureau Census Block Group
COTS	Commercial Off The Shelf
CT	U.S. Census Bureau Census Tract
DHHS	U.S. Department of Health and Human Services
DoD	U.S. Department of Defense
DMV	Department of Motor Vehicles
E-911	Emergency 911
EMS	Emergency Medical Services
FCC	Feature Classification Code
FGDC	Federal Geographic Data Committee
FIPS	Federal Information Processing Standards
FTE	Full Time Equivalent
GIS	Geographic Information System
G-NAF	Geocoded National Address File
GPS	Global Positioning System
IR	Information Retrieval
LA	Los Angeles
MBR	Minimum Bounding Rectangle
MCD	Minor Civil Division
MTFCC	MAF/TIGER Feature Class Code
NAACCR	North American Association of Central Cancer Registries
NCI	National Cancer Institute
NIH	United States National Institutes of Health
PO Box	USPS Post Office Box
RR	Rural Route
SES	Socio-Economic Status
SQL	Structured Query Language
SVM	Support Vector Machine
TIGER	Topographically Integrated Geographic Encoding and Referencing
TTL	Time to Live
URISA	Urban and Regional Information Systems Association
U.S.	United States
USC	University of Southern California
USPS	United States Postal Service
ZCTA	ZIP Code Tabulation Area

FOREWORD

The advent of geographic information science and the accompanying technologies (geographic information systems [GIS], global positioning systems [GPS], remote sensing [RS], and more recently location-based services [LBS]) have forever changed the ways in which people conceive of and navigate planet Earth. Geocoding is a key bridge linking the old and the new—a world in which streets and street addresses served as the primary location identifiers and the modern world in which more precise representations are possible and needed to explore, analyze, and visualize geographic patterns, their drivers, and their consequences. Geocoding, viewed from this perspective, brings together the knowledge and work of the geographer and the computer scientist. The author, Daniel Goldberg, has done an excellent job in laying out the fundamentals of geocoding as a process using the best contributions from both of these once-disparate fields.

This book will serve as a rich reference manual for those who want to inject more science and less art (uncertainty) into their geocoding tasks. This is particularly important for medical geography and epidemiology applications, as recent research findings point to environmental conditions that may contribute to and/or exacerbate health problems that vary over distances of hundreds and even tens of meters (i.e., as happens with proximity to freeways). These findings call for much better and more deliberate geocoding practices than many practitioners have used to date and bring the contents of this best practices manual to the fore. This book provides a long overdue summary of the state-of-the-art of geocoding and will be essential reading for those that wish and/or need to generate detailed and accurate geographic positions from street addresses and the like.

John Wilson
June 6, 2008

PREFACE

In one sense, writing this manuscript has been a natural continuation of the balancing act that has been, and continues to be, my graduate student career. I am fortunate to be a Computer Science (CS) Ph.D. student at the University of Southern California (USC), working in the Department of Geography, advised by a Professor in the Department of Preventive Medicine, who at the time of this writing was supported by the Department of Defense. While at times unbearably frustrating and/or strenuous, learning to tread the fine lines between these separate yet highly related fields (as well as blur them when necessary) has taught me some important lessons and given me a unique perspective from which I have written this manuscript and will take with me throughout my career. This combination of factors has led to my involvement in many extremely interesting and varied projects in diverse capacities, and to interact with academics and professionals with whom I would most likely not have otherwise met or had any contact.

Case in point is this manuscript. In November of 2006, Dr. John P. Wilson, my always industrious and (at-the-time) Geography advisor (now jointly appointed in CS) was hard at work securing funding for his graduate students (as all good faculty members should spend the majority of their time). He identified an opportunity for a student to develop a GIS-based traffic pollution exposure assessment tool for his colleague in the USC Department of Preventive Medicine, (my soon-to-be advisor) Dr. Myles G. Cockburn, which was right in line with my programming skills. What started off as a simple question regarding the supposed accuracy of the geocodes being used for the exposure model quickly turned into a day-long discussion about the geocoder I had built during the previous summer as a Research Assistant for my CS advisor, Dr. Craig A. Knoblock. This discussion eventually spawned several grant proposals, including one entitled *Geocoding Best Practices Document Phase I: Consultant for NAACCR GIS Committee Meeting & Development of Annotated Outline*, submitted to the North American Association of Central Cancer Registries (NAACCR) on April 21, 2006.

To my great surprise, I was awarded the grant and immediately set to work creating the outline for the meeting and the Annotated Geocoding Reading List I had promised in my proposal. Ambitiously, I started reading and taking notes on the 150 latest geocoding works, at which point the NAACCR GIS Committee, chaired at that time by David O'Brien of the Alaska Cancer Registry, should have run for cover. The first draft I produced after the in-person meeting during the June NAACCR 2006 Annual Meeting in Regina, Saskatchewan, Canada was far too detailed, too CS oriented, and too dense for anyone to make sense of. However, guided by the thoughtful but sometime ruthless suggestions of Drs. Wilson and Cockburn, I was able to transform that draft into an earlier version of this document for final submission to the NAACCR GIS Committee, which then sent it to the NAACCR Executive Board for approval in October 2006. It was approved, and I was subsequently selected to write the full version of the current work, *A Geocoding Best Practices Guide.*

I dare say that this exercise would prove longer and more in-depth than anyone could have anticipated. Looking back 2 years, I do not think I could have imagined what this project would have eventually turned into; 200 plus pages of text, 200 plus references, an annotated reading list the size of a small phone book, example research assurance documents, and a full glossary.

At more than one-half million characters and spanning more than 250 pages, it may at first seem a daunting task for one to read and digest this whole document. However, this fear should be laid to rest. More than one-third of this length is comprised of the front matter (e.g., Table of Contents, indices, Foreward, Preface, etc.) and the back matter (–e.g., Glossary, References, and Appendices). Most of this material is intended as reference, and it is expected that only the most motivated and inquisitive of readers will explore it all. The main content of the document, Sections 1-26, are organized such that an interested reader can quickly and easily turn to their topic(s) of interest, at the desired level of detail, at a moment's notice though the use of the Table of Contents and lists of figures and tables found in the front matter.

In addition to this concern, there were three major hurdles that had to be overcome during the writing of this document. The first was a question as to what the focus and tone should be. From the earliest conception, it was clear that this document should be a "Best Practices Guide," which implicitly meant that it should "tell someone what to do when in a particular situation." The question, however, was "*who* was the person who was to be informed?" Was it the technical person performing the geocoding who might run into a sticky situation and need direction as to which of two options they should choose? Was it the manager who needed to know the differences between reference datasets so they could make the correct investment for their registry? Or, was it the researcher who would be utilizing the geocoded data and needed to know what the accuracy measure meant and where it came from? After lengthy discussion, it was determined that the first two—the person performing the geocoding and the person deciding on the geocoding strategy—would be the target audience, because they are the registry personnel for whom this document was being created. Therefore, this document goes into great detail about the technical aspects of the geocoding process such that the best practices developed throughout the text can and should actually be applied during the process of geocoding. Likewise, the theoretical underpinnings are spelled out completely such that the person deciding on which geocoding process to apply can make the most informed decision possible.

The second hurdle that had to be cleared was political in nature. During the process of determining the set of theoretical best practices presented in this document, it came to light that in some cases, the current NAACCR standards and/or practices were insufficient, inappropriate, and/or precluded what I would consider the actual true best practice. Following lengthy discussion, it was decided that the set of best practices developed for this document should remain true to what "should be done," not simply what the current standards allow. Therefore, in several places throughout the manuscript, it is explicitly stated that the best practices recommended are in the ideal case, and may not be currently supported with other existing NAACCR standards. In these cases, I have attempted to provide justification and support for why these would be the correct best practice in the hopes that they can be taken into consideration as the existing NAACCR standards are reviewed and modified over time.

The final challenge to overcome in creating this manuscript was the sheer diversity of the NAACCR member registries in terms of their geocoding knowledge, resources, practices, and standards that needed to be addressed. The members of the NAACCR GIS Committee who contributed to the production of this document came from every corner of the United States, various levels of government, and represented the full geocoding spectrum from highly advanced and extremely knowledgeable experts to individuals just starting out with more questions than answers. Although input from all of these varied user types undoubtedly led to a more accessible finished product, it was quite a task to produce a document that would be equally useful to all of them. I feel that their input helped produce a much stronger

text that should be appropriate to readers of all levels, from those just getting started to those with decades of experience who may be developing their own geocoders.

The content of this manuscript represents countless hours of work by many dedicated people. The individuals listed in the Acknowledgments Section each spent a significant amount of time reviewing and commenting on every sentence of this document. Most participated in weekly Editorial Review Committee calls from March 2007 to March 2008, and all contributed to making this document what it is. In particular, I would like to thank Frank Boscoe for his steady leadership as NAACCR GIS Committee Chair during the period covering most of the production of this book. I take full responsibility for all grammatical errors and run-on sentences, and believe me when I tell you that this book would be in far worse shape had John Wilson not volunteered to copyedit every single word. I would not be writing this if it was not for Myles Cockburn, so for better or worse, all blame should be directed toward him. The other members of the weekly Editorial Review Committee, namely Stephanie Foster, Kevin Henry, Christian Klaus, Mary Mroszczyk, Recinda Sherman and David Stinchcomb, all volunteered substantial time and effort and contributed valuable expert opinions, questions, corrections, edits, and content, undoubtedly improving the quality of the final manuscript. These detailed and often heated discussions served to focus the content, tone, and direction of the finished product in a manner that I would have been incapable of on my own. I would not currently be a Ph.D. student, much less know what a geocoder was, if it were not for the support of Craig Knoblock. Last but in no way least, Mona Seymour graciously volunteered her time to review portions of this manuscript, resulting in a far more readable text.

Sadly, everyone who reads this document will most likely have already been affected by the dreadful toll that cancer can take on a family member, friend, or other loved one. I whole-heartedly support the goal of NAACCR to work toward reducing the burden of cancer in North America, and I am honored to have been granted the opportunity to give in this small way to the world of cancer-related research. What follows in this document is my attempt to contribute through the production of a *Geocoding Best Practices Guide* for use in standardizing the way that geocoding is discussed, performed, and used in scientific research and analysis.

Daniel W. Goldberg
June 6, 2008

ACKNOWLEDGEMENTS

Much of the material in this handbook was generated at the North American Association of Central Cancer Registries (NAACCR) GIS Workgroup meeting held in Regina, Saskatchewan, Canada on June 16, 2006. The following individuals contributed to the development of this document:

Meeting Participants:
Francis P. Boscoe, Ph.D., New York State Cancer Registry
Myles G. Cockburn, Ph.D., University of Southern California
Stephanie Foster, Centers for Disease Control and Prevention
Daniel W. Goldberg, Ph.D. Candidate, Facilitator and Handbook Author, University of Southern California
Kevin Henry, Ph.D., New Jersey State Department of Health
Christian Klaus, North Carolina State Center for Health Statistics
Mary Mroszczyk, C.T.R., Massachusetts Cancer Registry
David O'Brien, Ph.D., Alaska Cancer Registry
David Stinchcomb, Ph.D., National Cancer Institute

Comments, Reviewers, and Editors:
Robert Borchers, Wisconsin Department of Health and Human Services
Francis P. Boscoe, Ph.D., New York State Cancer Registry
Myles G. Cockburn, Ph.D., University of Southern California
Stephanie Foster, Centers for Disease Control and Prevention
Kevin Henry, Ph.D., New Jersey State Department of Health
Christian Klaus, North Carolina State Center for Health Statistics
Mary Mroszczyk, C.T.R., Massachusetts Cancer Registry
David O'Brien, Ph.D., Alaska Cancer Registry
Mona N. Seymour, Ph.D. Candidate, University of Southern California
Recinda L. Sherman, M.P.H., C.T.R., Florida Cancer Data System
David Stinchcomb, Ph.D., National Cancer Institute
John P. Wilson, Ph.D., University of Southern California

This project has been funded in part with federal funds from the National Cancer Institute (NCI), National Institutes of Health (NIH), Department of Health and Human Services (DHHS) under Contract No. HHSN26120044401C and ADB Contract No. N02-PC-44401, and from the Centers for Disease Control and Prevention (CDC) under Grant/Cooperative Agreement No. U75/CCU523346. Daniel Goldberg was supported by a U.S. Department of Defense (DoD) Science, Mathematics, and Research for Transformation (SMART) Defense Scholarship for Service Program fellowship and National Science Foundation (NSF) Award No. IIS-0324955 during portions of the production of this document. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

DEDICATION

This book is dedicated to the life and memory of Michael Owen Wright-Goldberg, beloved husband, son, and older brother (1975-2006).

ABOUT THIS DOCUMENT

BEST PRACTICES GUIDE

The main purpose of this document is to act as a best practices guide for the cancer registry community, including hospitals as well as state and provincial registries. Accordingly, it will advise those who want to know specific best practices that they should follow to ensure the highest level of confidence, reliability, standardization, and accuracy in their geocoding endeavors. These best practices will be framed as both policy and technical decisions that must be made by a registry as a whole and by the individual person performing the geocoding or using the results. Best practices are listed throughout the text, placed as close to the section of text that describes them as possible.

STANDARDIZATION

Due to a fundamental lack of standardization in the way that geocoding is defined and implemented across cancer registries, it is difficult to compare or integrate data created at different sources. This document will propose numerous definitions germane to the geocoding process, thus developing a consistent vocabulary for use as a first step toward a larger standardization process. Throughout the document, specific terms will be written in bold with definitions closely following. A **geocoding best practice** is a policy or technical decision related to geocoding recommended (but not required) by NAACCR for use in a cancer registry's geocoding process. The **geocoding best practices** are a set of suggested best practices developed throughout this document. In addition, the document attempts to detail software implementation preferences, current limitations, and avenues for improvement that geocoding vendors should be aware are desired by the cancer research communities.

Note that the **best practices** developed in this document are not as-of-yet official **NAACCR data standards**, meaning that they will not be found in the current version of *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008), and thus their use is not officially required by any means. More specifically, although the content of Hofferkamp and Havener (2008) represent the current mandatory **NAACCR data standards** that registries are required to follow, the **best practices** found herein are recommended for adoption by researchers, registries, and/or software developers that seek to begin conducting their geocoding practices in a consistent, standardized, and more accurate manner. It is the hope of the author that the contents of this document will assist in the eventual official standardization of the geocoding process, fully accepted and recognized by the NAACCR Executive Board. As such, software developers are encouraged to adopt and incorporate the recommendations included in this document to: (1) be ahead of the curve if/when the recommendations contained herein (or their derivates and/or replacements) are accepted as true NAACCR data standards; and (2) improve the quality, transparency, usability, and legitimacy of their geocoding products.

LEARNING TOOL

To make informed decisions about geocoding choices, an understanding of both the theoretical and practical aspects of the geocoding process is necessary. Accordingly, this document provides a high level of detail about each aspect of the geocoding process such that a reader can obtain a complete understanding of the best practice recommended, other possible options, and the rationale behind the recommended practice. It serves to centralize

much of the available research and practice scholarship on these topics to provide a single, comprehensive perspective on all aspects of the geocoding process.

The document has been specifically divided into six parts. Each part attempts to address the topics contained in it at a consistent level of detail. The decision was made to organize the document in this format so that it would be easy for a reader interested in certain topics to find the information he or she is looking for (e.g., to learn about components of geocoding or find solutions to an exact problem) without being bogged down in either too much or too little detail.

REFERENCE TOOL

Appendix A includes example research assurance documents that can be tailored to an individual registry for ensuring that researchers understand the acceptable manner in which registry data may be obtained and used. Every attempt was made to back up all claims made in the document using published scientific literature so that it can be used as a reference tool. The Annotated Bibliography included as Appendix B includes more than 250 of the most recently published geocoding works classified by the topic(s) they cover, and should prove a useful resource for those interested in further reading.

TYPES OF READERS

In this document, four distinct types of readers will be identified based on their specific roles in, or uses of, the geocoding process. These are: (1) the practitioner, (2) general interest, (3) process designer, and (4) data consumer groups. The roles of these groups are described in Table 1, as are their main concerns regarding the geocoding process and the sections in this document that address them.

SUGGESTED CITATION

Goldberg DW: A Geocoding Best Practices Guide. Springfield, IL: North American Association of Central Cancer Registries; 2008.

Table 1 – Types of readers, concerns, and sections of interest

Group	Role	Concerns	Sections of Interest
Practitioner	Registry staff performing the geocoding task using some pre-defined method with existing tools, ultimately responsible for the actual production of the geospatial data from the raw aspatial address data	Practical aspects of the geocoding process Handling instances in which data do not geocode	1, 4, 5, 14, 15, 16, 17, 18, 19, 20, 24
General Interest	Registry staff interested in geocoding but not formally involved in the process as part of their duties, akin to the general public	Why is geocoding important? How does geocoding fit into the larger operations of the registry?	1, 2.1, 2.4, 3, 4, 11, 12.5, 14, 15, 18, 26
Process Designers	Registry staff overseeing and designing the geocoding process used at a registry, ultimately responsible for the overall outcome of the geocoding performed at a registry	All design and policy decisions that affect the outcome of geocoding Data definition, representation, and validation; components and algorithms involved in the geocoding process; forms and formats of reference data sources; and accuracy metrics and reporting	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26
Data Consumers	Cancer researchers consuming the geocoded data Others responsible for monitoring annually reported aggregate statistics to discover important trends	Accuracy of the geocoded output in terms of its lineage, confidence/reliability, accountability, and any assumptions that were used	1, 3, 5.3, 6.1, 7, 8, 11, 12, 13, 14, 15, 18, 19, 20, 24, 25, 26

EXECUTIVE SUMMARY

PURPOSE

As a rule, health research and practice (and cancer-related research in particular) takes place across a multitude of administrative units and geographic extents (country-wide, state-wide, etc.). The data used to develop and test cancer-related research questions are created, obtained, and processed by disparate organizations at each of these different levels. Studies requiring the aggregation of data from multiple administrative units typically must integrate these disparate data, which occur in incompatible formats with unknown lineage or accuracy. The inconsistencies and unknowns amongst these data can lead to uncertainty in the results that are generated if the data are not properly integrated. This problem of data integration represents a fundamental hurdle to cancer-related research.

To overcome the difficulties associated with disparate data, a specific set of actions must be undertaken. First, key stakeholders must be identified and informed of potential issues that commonly arise and contribute to the problem. Next, a common vocabulary and understanding must be defined and developed such that thoughtful communication is possible. Finally and most importantly, advice must be provided in the form of a set of best practices so that processes can begin to be standardized across the health research communities. Together, these will allow health researchers to have a reasonable level of certainty as to how and where the data they are working with have been derived as well as an awareness of any overarching data gaps and limitations.

Person, place, event, and time form the four fundamental axes of information around which epidemiologic research is conducted. The spatial data representing the subject's location is particularly susceptible to the difficulties that plague multi-source data because much of the spatial data are derived from textual addresses through the process of **geocoding.** These data are vulnerable to inconsistencies and unknown quality because of the wide range of methods by which they are defined, described, collected, processed, and distributed. To contextualize the heterogeneity of current geocoding practices among cancer registries, see the recent work by Abe and Stinchcomb (2008), which highlights the far-ranging approaches used at several cancer registries throughout the United States. This lack of uniformity and/or standardization with regard to geocoding processes represents a current and significant problem that needs to be addressed.

Although there is a substantial amount of available literature on the many topics germane to geocoding, there is no single source of reference material one can turn to that addresses many or all of these topics in the level of detail required to make well-informed decisions. Recent works such as Rushton et al. (2006, 2008a), Goldberg et al. (2007a), and Mechanda and Puderer (2007) provide a great deal of review and detail on geocoding and related topics, but available scholarship as to specific recommendations, their rationale, and alternative considerations is lacking.

To these ends, The North American Association of Central Cancer Registries (NAACCR) has promoted the development of this work, *A Geocoding Best Practices Guide*, the purpose of which is to help inform and standardize the practice of geocoding as performed by the cancer registries and research communities of the United States and Canada. This work primarily focuses on the theoretical and practical aspects of the actual production of geocodes, and will briefly touch upon several important aspects of their subsequent usage in cancer-related research.

SCOPE

This document will cover the fundamental underpinnings of the geocoding process. Separate sections describe the components of the geocoding process, ranging from the input data, to the internal processing performed, to the output data. For each topic, choices that affect the accuracy of the resulting data will be explored and possible options will be listed.

INTENDED AUDIENCE

The primary purpose of this document is to provide a set of best practices that, if followed, will enable the standardization of geocoding throughout the cancer research communities. Thus, the main focus of this document will be to provide enough detailed information on the geocoding process such that informed decisions can be made on each aspect—from selecting data sources, algorithms, and software to be used in the process; to defining the policies with which the geocoding practitioner group perform their task and make decisions; to determining and defining the metadata that are associated with the output.

For those with varying levels of interest in the geocoding process, this document presents detailed information about the components and processes involved in geocoding, as well as sets of best practices designed to guide specific choices that are part of any geocoding strategy. Benefits and drawbacks of potential options also are discussed. The intent is to establish a standardized knowledge base that will enable informed discussions and decisions within local registries and result in the generation of consistent data that can be shared between organizations.

For researchers attempting to use geocoded data in their analyses, this document outlines the sources of error in the geocoding process and provides best practices for describing them. If described properly, accuracy values for each stage of the geocoding process can be combined to derive informative metrics capable of representing the accuracy of the output in terms of the whole process. The data consumer can use these to determine the suitability of the data with respect to the specific needs of their study.

For practitioners, this document presents detailed, specific solutions for common problems that occur during the actual process of geocoding, with the intent of standardizing the way in which problem resolution is performed at all registries. Uniform problem resolution would remove one aspect of uncertainty (arguably the most important level) from the geocoding process and ultimately from the resulting data and analyses performed on them.

Most of the information contained in this document (e.g., examples, data sources, laws, and regulations) will primarily focus on U.S. and Canadian registries and researchers, but the concepts should be easily translated to other countries. Likewise, some of the information and techniques outlined herein may only be applicable to registries that perform their own geocoding instead of using a commercial vendor. Although the number of these registries performing their own geocoding is currently small, this number has been and will continue to increase as access to geocoding software and required data sources continue to improve. Additionally, the information within this document should assist those registries currently using a vendor in becoming more understanding of and involved in the geocoding process, better able to explain what they want a vendor to do under what circumstances, and more cognizant of the repercussions of choices made during the geocoding process.

Part 1: The Concept and Context of Geocoding

As a starting point, it is important to succinctly develop a concrete notion for exactly what geocoding is and identify how it relates to health and cancer-related research. In this part of the document, geocoding will be explicitly defined and its formal place in the cancer research workflow will be identified.

This page is left blank intentionally.

1. INTRODUCTION

This section provides the motivation for standardized geocoding.

1.1 WHAT IS GEOCODING?

Person, place, event, and time are the four key pieces of information from which epidemiologic research in general is conducted. This document will focus primarily on issues arising in the description, definition, and derivation of the place component. In the course of this research, scientists frequently use a variety of spatial analysis methods to determine trends, describe patterns, make predictions, and explain various geographic phenomena.

Although there are many ways to denote place, most people rely almost exclusively on locationally descriptive language to describe a geospatial context. In the world of cancer registries this information typically includes the address, city, and province or state of a patient at the diagnosis of their disease (dxAddress, dxCity, dxProvince, dxState), most commonly in the form of postal street addresses. These vernacular, text-based descriptions are easily understood by people, but they are not directly suitable for use in a computerized environment. Performing any type of geospatial mapping or investigation with the aid of a computer requires discrete, non-ambiguous, geographically valid digital data rather than descriptive textual strings.

Thus, some form of processing is required to convert these text descriptors into valid geospatial data. In the parlance of geographic information science (GIS), this general concept of making implicit spatial information explicit is termed **georeferencing,** or transforming **non-geographic information,** information that has no geographically valid reference that can be used for spatial analyses, into **geographic information,** information that has a valid geographic reference that can be used for spatial analyses (Hill 2006).

Throughout the years, this general concept has been realized in a multitude of actual processes to suit the needs of various research communities. For instance, a **global positioning system** (GPS) device can produce coordinates for the location on the Earth's surface based on a system of satellites, calibrated ground stations, and temporally based calculations. The coordinates produced from these devices are highly accurate, but can be expensive in terms of time and effort required to obtain the data, as they typically require a human to go into the field to obtain them.

Geocoding describes another method of georeferencing (Goldberg et al. 2007a). As seen in scholarship and practice, the term **geocoding** is used throughout almost every discipline of scientific research that includes any form of spatial analysis, with each field usually either redefining it to meet their needs or adopting another field's definition wholesale. As a result, there is a great deal of confusion as to what geocoding—and its derivatives, most notably the terms **geocode** and **geocoder**—actually refer to. What do these words mean, and how should they be used in the cancer registry field?

For example, does **geocoding** refer to a specific computational process of transforming something into something else, or simply the concept of a transformation? Is a **geocode** a real-world object, simply an attribute of something else, or the process itself? Is a **geocoder**

the computer program that performs calculations, a single component of the process, or the human who makes the decisions?

An online search performed in April of 2008 found the various definitions of the term **geocoding** shown in Table 2. These definitions have been chosen for their geographic diversity as well as for displaying a mix of research, academic, and industry usages. It is useful to contrast our proposed definition with these other definitions that are both more constrained and relaxed in their descriptions of the geocoding process to highlight how the proposed definition is more representative of the needs of the health/cancer research communities.

Table 2 – Alternative definitions of "geocoding"

Source	Definition	Possible Problems
Environmental Sciences Research Institute (1999)	The process of matching tabular data that contains location information such as street addresses with real-world coordinates.	Limited to coordinate output only.
Harvard University (2008)	The assignment of a numeric code to a geographical location.	Limited to numeric code output only.
Statistics Canada (2008)	The process of assigning geographic identifiers (codes) to map features and data records.	Limited input range.
U.S. Environmental Protection Agency (2008)	The process of assigning latitude and longitude to a point, based on street addresses, city, state and USPS ZIP Code.	Limited to coordinate output only.

As a further complication, it must be noted that the methods and data sources employed throughout all registries in the United States and Canada are quite diverse and varied so a single definition explicitly defining, requiring, or endorsing a particular technology would not be useful. Each registry may have different restrictions or requirements on what can be geocoded in terms of types of input data (postal addresses, named places, etc.), which algorithms can be used, what geographic format the results must be in, what can be produced as output, or what data sources can be used to produce them. Differing levels of technical skills, budgetary and legal constraints, and varied access to types of geographic data, along with other factors, also may dictate the need for a broad definition of geocoding. As such, the definition offered herein is meant to serve the largest possible audience by specifically not limiting any of these characteristics of the geocoding process, intentionally leaving the door open for different flavors of geocoding to be considered as valid. In the future, as the vast body of knowledge of geocoding constraints, ontologies, and terminologies spreads and is utilized by registry personnel, it is expected that there will be a common desire in the registry community to achieve consensus on standardizing geocoding and geocoding-related processes to achieve economies of scale.

The remainder of this document will explicitly define geocoding as well as its component parts as follows:

Geocoding (verb) is the act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process.

A **geocoder** (noun) is a set of inter-related components in the form of operations, algorithms, and data sources that work together to produce a spatial representation for descriptive locational references.

A **geocode** (noun) is a spatial representation of a descriptive locational reference.

To **geocode** (verb) is to perform the process of geocoding.

In particular, these definitions help to resolve four common points of confusion about geocoding that often are complicated by disparate understandings of the term: (1) the types of data that can be geocoded, (2) the methods that can be employed to geocode data, (3) the forms and formats of the outputs, and (4) the data sources and methods that are germane to the process. These definitions have been specifically designed to be broad enough to meet the diverse needs of both the cancer registry and cancer research communities.

1.1.1 Issue #1: Many data types can be (and are) geocoded

There are many forms of information that registries and researchers need geocoded. Table 3 illustrates the magnitude of the problem in terms of the many different types of addresses that may be encountered to describe the same physical place, along with their best and worst resolutions resulting from geocoding and common usages.

Table 3 – Possible input data types (textual descriptions)

Name	Type	Usage	Best/Worst Case Output Resolution
The University of Southern California	Named place	County counts	Parcel-level/ Non-matchable
The University of Southern California GIS Research Lab	Named place	Cluster screening	Sub parcel-level/ Non-matchable
Kaprielian Hall, Unit 444	Named place	Cluster screening	Sub parcel-level/ Non-matchable
The northeast corner of Vermont Avenue and 36th Place	Relative intersection	Cluster screening	Intersection-level/ Non-matchable
Across the street from Togo's, Los Angeles 90089	Relative direction	Cluster screening	Street-level/ Non-matchable
3620 South Vermont Ave, Los Angeles, CA 90089	Street address	Cluster screening	Building-level/ Street-level
USPS ZIP Code 90089-0255	USPS ZIP Code	County counts	Building-level/ USPS ZIP Code-level
34.022351, -118.291147	Geographic coordinates	Cluster screening	Sub parcel-level/ Non-matchable

It should be clear from this list that a location can be described as a named place, a relative location, a complete postal address (or any portion thereof), or by its actual coordinate representation. All of these phrases except the last (actual coordinates) are commonly occurring representations found throughout health data that need to be translated into spatial coordinates. Obviously, some are more useful than others in that they relay more detailed information (some may not even geocode). Registry data standards are heading toward the enforcement of a single data format for input address data, but utilization of a single representation across all registries is presently not in place due to many factors. In keeping with the stated purpose of this document, the definition provided should be general enough to encompass each of the commonly occurring reporting styles (i.e., forms).

1.1.2 Issue #2: Many methods can be (and are) considered geocoding

Turning to the host of methods researchers have used to geocode their data, it becomes clear that there are still more varieties of geocoding. The process of utilizing a GPS device and physically going to a location to obtain a true geographic position has been commonly cited throughout the scientific literature as one method of geocoding. This is usually stated as the most accurate method, the **gold standard.** Obtaining a geographic position by identifying the geographic location of a structure through the use of georeferenced satellite or aerial imagery also has been defined as geocoding. The direct lookup of named places or other identifiable geographic regions (e.g., a U.S. Census Bureau ZIP Code Tabulation Area [ZCTA]) from lists or **gazetteers** (which are databases with names, types, and footprints of geographic features) also has been referred to as geocoding. Most commonly, geocoding refers to the use of interpolation-based computational techniques to derive estimates of geographic locations from GIS data such as linear street vector files or areal unit parcel vector files.

1.1.3 Issue #3: Many output types are possible

Geocoding output is typically conceived of as a geographic point, a simple geographic coordinate represented as latitude and longitude values. However, the base geographic data used for the derivation of the point geocode (e.g., the polygon boundary of the parcel or the polyline of the street vector) also could be returned as the output of the geocoding process.

1.1.4 Issue #4: Geocoding can be (and usually is) a multi-component process

Finally, the geocoding process is not achieved by one single instrument, software, or geographic data source. The process of geocoding can be conceptualized as a single operation, but there are multiple components such as operations, algorithms, and data sources that work together to produce the final output. Each one of these components is the result of significant research in many different scientific disciplines. Each is equally important to the process. Thus, when one speaks of geocoding, it begs the question: are they speaking of the overall process, or do they mean one or more of the components? The proposed definition therefore must take this into account and make these distinctions.

By design, any of the processes stated earlier in Section 1.1.2 that are known as geocoding are valid (e.g., using a software geocoder, GPS in the field, or imagery would fit into this definition). By using the terms "locationally descriptive text" and "spatial representation," any of the forms of data listed earlier in Sections 1.1.1 and 1.1.3 are valid as input and output, respectively. Finally, instead of explicitly stating what must be a part of a geocoder, it may be best to leave it open-ended such that different combinations of algorithms and data sources can be employed and still adhere to this definition.

Again, the primary purpose of this document is to assist registries in making the appropriate choices given their particular constraints and to explain the repercussions these decisions will have. Because the definition presented here is tailored specifically for a certain community of researchers with unique characteristics, it may not be appropriate for other research disciplines. It should be noted that although this definition allows for any type of geographic output, registries must at least report the results in the formats explicitly defined in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008). Best practices relating to the fundamental geocoding concepts developed in this section are listed in Best Practices 1.

Best Practices 1 – Fundamental geocoding concepts

Topic	Policy Decision	Best Practice
Geocoding concept	What does geocoding refer to in their organization?	The meaning of geocoding within an organization should be consistent with that presented in this document.
Geocoding motivation	When should geocoding be performed?	Geocoding should be performed when descriptive location data need to be transformed into numeric spatial data to support spatial analysis.
Geocoding vocabulary	What are the definitions of geocoding and its related terms?	The definitions of geocoding should be based on those within this document.

This page is left blank intentionally.

2. THE IMPORTANCE OF GEOCODING

This section places geocoding in the larger context of spatial analysis performed as part of cancer-related research.

2.1 GEOCODING'S IMPORTANCE TO HOSPITALS AND CENTRAL REGISTRIES

As the component ultimately responsible for generating the spatial attributes associated with patient/tumor data, geocoding's primary importance to cancer research becomes clear (Rushton et al. 2006, 2008a). Spatial analysis would be extremely difficult, if not impossible, in the absence of geocoding. Note that the patient's address at the time of diagnosis is tumor-level data (i.e., each tumor will have its own record, and each record will have its own address). Being time-dependent, this address may vary with each cancer.

The recent work by Abe and Stinchcomb (2008) relating the results of a North American Association of Central Cancer Registries (NAACCR) GIS Committee survey of 72 NAACCR member registries crystallizes the importance of geocoding in current registry practice. They found that 82 percent of the 47 responding registries performed some type of address geocoding, and that the average number of addresses geocoded was 1.7 times the annual caseload of the registry. For complete details of this survey, refer to NAACCR (2008a, 2008b).

2.2 TYPICAL RESEARCH WORKFLOW

The important role of the geocoder in cancer-related research is easily highlighted through an example of a prototypical research workflow that utilizes geocoding as a component. Generally, a spatially inspired research investigation will have two basic components—data gathering and data analysis. Looking deeper, the data gathering portion involves data collection, consolidation, and processing. The data analysis portion involves hypothesis testing using statistical analysis to assess the phenomena in question given the relative strength of the supporting evidence. Figure 1 displays an outline of a research workflow that should solidify the importance of the geocoder as the link between descriptive locational data and numeric (digital) geographic information.

2.2.1 Data gathering

Each registry will undoubtedly vary in its exact implementation of the data gathering protocol. Generally, when an incidence of cancer is diagnosed, a series of best efforts are made to obtain the most accurate data available about the person and their tumor. Many types of confidential data will be collected as the patient receives care, including identifiable information about the person and their life/medical history, as well as information about the diagnosed cancer. Portions of this identifiable information may be obtained directly from the patient through interviews with clerical or medical staff at a hospital. Here, the patient will be asked for his/her current home address, and possibly a current work address and former addresses.

Time and/or confidentiality constraints may limit the amount of information that can be collected by a hospital and only the patient's current post-diagnosis address may be available when consolidation occurs at the registry. The central registry therefore may be expected to

determine an address at diagnosis after the data have been submitted from a diagnosing facility for consolidation. Searching a patient's medical records, or in some cases linking to Department of Motor Vehicles (DMV) records, may be options to help achieve this (more detail on these practices is provided in Section 19).

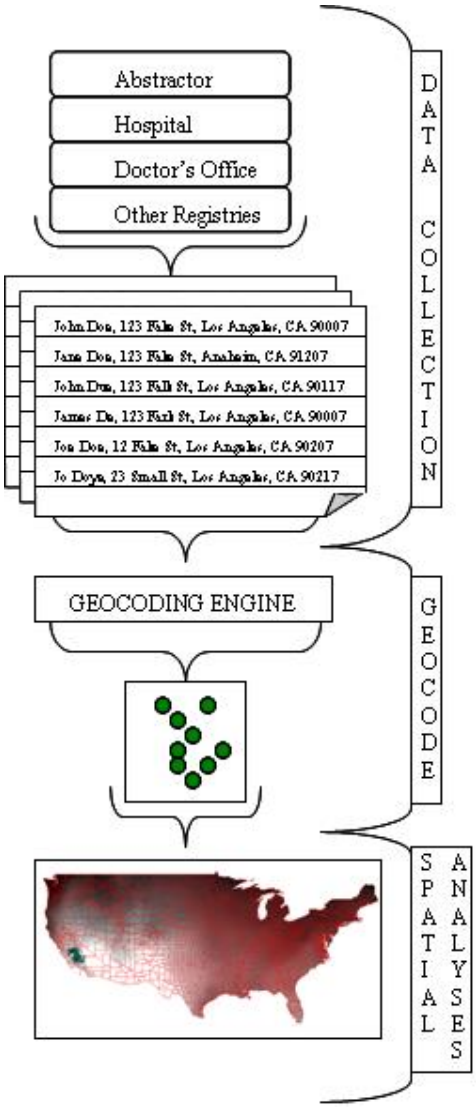


Figure 1 – Typical research workflow

In any case, the address data obtained represent the descriptive text describing the person's location in at least one point in time. At this stage a record has been created, but it is as of yet completely **aspatial** (or **non-spatial**), meaning that it does not include any spatial information. Although there are multiple methods to collect data and perform address consolidation, there are currently no standards in place. Best Practices 2 lists data gathering and metadata guides for the hospital and registry, in the ideal cases. Best Practices 3 lists the type of residential history data that would be collected, also in the ideal case, but one that has a clear backing in the registry (e.g., Abe and Stinchcomb 2008, pp. 123) and research communities (e.g., Han et al. 2005). Best Practices 4 briefly lists guides for obtaining secondary information about address data.

Best Practices 2 – Address data gathering

Policy Decision	Best Practice
When, where, how and what type of address data should be gathered?	Residential history information should be collected. Collect information as early as possible: • As much data as possible at the diagnosing facility • As much data as possible at the registry. Collect information from the most accurate source: • Patient • Relative • Patient/tumor record. Metadata should describe when it was collected: • Upon intake • After treatment • At registry upon arrival • At registry upon consolidation. Metadata should describe where it was collected: • At diagnosing facility • At registry. Metadata should describe how it was collected: • Interview in-person/telephone • Patient re-contacting • From patient/tumor record • Researched online. Metadata should describe the source of the data: • Patient • Relative • Patient/tumor record.

Best Practices 3 – Residential history address data

Policy Decision	Best Practice
What type of residential history information should be collected?	Ideally, complete historical address information should be collected: • Residential history of addresses • How long at each address • Type of address (home, work, seasonal).

Best Practices 4 – Secondary address data gathering

Policy Decision	Best Practice
Should secondary research methods be attempted at the registry? If so, when, which ones, and how often?	Secondary research should be attempted at the registry if address data from the diagnosing facility is inaccurate, incomplete, or not timely. All available and applicable sources should be utilized until they are exhausted or enough information is obtained: • Online searches • Agency contacts • Patient re-contacting. Metadata should describe the data sources consulted: • Websites • Agencies • Individuals. Metadata should describe the queries performed. Metadata should describe the results achieved, even if unsuccessful. Metadata will need to be stored in locally defined fields.

2.2.2 Conversion to numeric (i.e., digital) data

How long the patient record remains unaltered depends on several factors, including:

- The frequency with which data are reported to the central registry
- Whether **per-record geocoding** (geocoding a single record at a time) or **batch-record geocoding** (geocoding multiple records at once) is performed
- Whether the registry geocodes the data or they already are geocoded when the data reaches them
- Whether geocoding is performed in-house or outsourced.

When the data finally are geocoded, addresses will be converted into numeric spatial representations that then will be appended to the record. Best practices related to the

conversion from descriptive to numeric spatial data (assuming that geocoding is performed at the registry) are listed in Best Practices 5. Note that the recommendation to immediately geocode every batch of data received may not be a feasible option for all registries under all circumstances because of budgetary, logistical, and/or practical considerations involved with processing numerous data files.

Best Practices 5 – Conversion to numeric spatial data

Policy Decision	Best Practice
How long can a record remain non-matched (e.g., must it be transformed immediately, or can it wait indefinitely?)	If records are obtained one at a time, they should be geocoded when a sufficient number have arrived to offset the cost of geocoding (i.e., to achieve economies of scale). If records are obtained in batches, they should be geocoded as soon as possible.
Should a record ever be re-geocoded? If so, when and under what circumstances?	The record should retain the same geocode until it is deemed to be unacceptably inaccurate. If new reference data or an improved geocoder are obtained and will provably improve a record's geocode, it should be re-geocoded. If new or updated address data are obtained for a record, it should be re-geocoded. Metadata should describe the reason for the inaccuracy determination. Metadata should retain all historical geocodes.

2.2.3 Spatial association

Once the data are in this numeric form the quality of the geocoding process can be assessed. If the quality is found to be sufficient, other desired attributes can be associated and spatial analysis can be performed within a GIS to investigate any number of scientific outcomes. For example, records can be visually represented on a map, or values from other datasets can be associated with individual records though the spatial intersection of the geocode and other spatial data. Common data associations include spatial intersection with U.S. Census Bureau data to associate socioeconomic status (SES) with a record or intersection with environmental exposure estimates. See Rushton et al. (2008b) for one example of how spatially continuous cancer maps can be produced from geocoded data and/or Waller (2008) for concise introductory material on the type of spatial statistical analysis typically performed on point and areal unit data in public health research. Best practices related to the conversion from descriptive to numeric to spatial data are listed in Best Practices 6.

Best Practices 6 – Spatial association

Policy Decision	Best Practice
Should data be allowed to be spatially associated with geocoded records? If so, when and where?	Spatial associations should be allowed if the data to be associated meet acceptable levels of accuracy and integrity. Spatial association should be examined every time an analysis is to be run (because the viability of statistical analysis techniques will vary with the presence or absence of these associations), and can be performed by data consumers or registry staff. Metadata should include the complete metadata of the spatially associated data.
What requirements must data meet to be spatially associated?	Data should be considered valid for association if: • Its provenance, integrity, temporal footprint, spatial accuracy, and spatial resolution are known and can be proven. • Its temporal footprint is within a decade of the time the record was created. • Its spatial resolution is equal to or less than that of the geocode (i.e., only associate data of lower resolution).

2.3 WHEN TO GEOCODE

When the actual geocoding should be performed during the data gathering process is a critical component that affects the overall accuracy of the output. There are several options that need to be considered carefully, as each has particular benefits and drawbacks that may influence the decision about when to geocode.

2.3.1 Geocoding at the Central Registry

The first option, **geocoding at the central registry,** is the process of geocoding patient/tumor records at the registry once they have been received from facilities and abstractors. Geocoding traditionally has been the role of the central registry. This can be accomplished when a record arrives or after consolidation. The first approach processes one case at a time, while the second processes batches of records at a time. One obvious benefit in the second case results from economies of scale. It will always be cheaper and more efficient to perform automatic geocoding for a set of addresses, or in batch mode, rather than on a single address at a time. Although the actual cost per produced geocode may be the same (e.g., one-tenth of a cent or less), the time required by staff members in charge of the process will be greatly reduced, resulting in definite cost savings.

Common and practical as this method may be, it also suffers from setbacks. First and foremost, if incorrect, missing, or ambiguous data that prevent a geocode from successfully being produced have been reported to a registry, it will be more difficult and time consuming to correct at this stage. In fact, most geocoders will not even attempt such corrections; instead, they will simply either output a less accurate geocode (e.g., one representing a successful geocode at a lower geographic resolution), or not output a geocode at all (Section 18 provides more discussion of these options).

Some of these problems can be rectified by performing **interactive geocoding,** whereby the responsible staff member is notified when problematic addresses are encountered and intervenes to choose between two equally likely options in the case of an ambiguous address or to correct an obvious and easily solvable error that occurred during data entry. Interactive geocoding, however, cannot solve the problems that occur when not enough information has been recorded to make an intelligent decision, and the patient cannot or should not be contacted to obtain further details. Further, interactive geocoding may be too time consuming to be practical.

2.3.2 Geocoding at the diagnosing facility

The second, less-likely option, **geocoding at the diagnosing facility,** is the process of geocoding a record at the intake facility while the person performing the intake is conducting the ingest or performing the interview and creating the electronic record or abstract. This option performs the geocoding as the abstractor, clerical, intake, or registration personnel at the hospital (i.e., whomever on the staff is in contact with the patient) is performing the data ingest, or when he or she is performing the patient interview and creating their electronic record or abstract. Geocoding at this point will result in the highest percentage of valid geocodes because the geocoding system itself can be used as a validation tool. In addition, staff can ask the patient follow-up questions regarding the address if the system returns it as an address that is non-matchable, a sentiment clearly echoed by the emerging trends and attitudes in the cancer registry community (e.g., Abe and Stinchcomb 2008, pp 123). Street-level geocoding at the hospital is ideal, but has yet to be realized at most facilities.

This is an example of one definition for the term **real-time geocoding,** the process of geocoding a record while the patient or the patient's representative is available to provide more detailed or correct information using an iterative refinement approach. Data entry errors resulting from staff input error can be reduced if certain aspects of the address can be filled in automatically as the staff member enters them in a particular order, from lowest resolution to highest. For instance, the staff can start with the state attribute, followed by the United States Postal Service (USPS) ZIP Code. Upon entering this, in some cases both the county and city can be automatically filled in by the geocoding system, which the patient then can verify. However, if the USPS ZIP Code has other USPS-acceptable postal names or represents mail delivery to multiple counties, these defaults may not be appropriate. This process also may be performed as **interactive** or **non-interactive geocoding.**

Looking past this and assuming the case of a USPS ZIP Code that can assign city and county attributes correctly, a further step can be taken. The street, as defined by its attributes (e.g., name, type, directional) can be validated by the geocoding system as actually existing within the already-entered county and USPS ZIP Code, and the building number can be tested as being within a valid range on that street. At any point, if invalid or ambiguous data are discovered by the geocoding system (or the address validation component or stand-alone system) as it is being entered, the staff can be instructed to ask follow-up questions to resolve the conflicts. Depending on the polices of the central registry, all that may be required of a hospital is to ensure that all of the steps that could have provided the patient with the opportunity to resolve the conflict were taken and their outcomes documented, even if the outcome was a refusal to clarify. If a correct address can be determined, entered, and verified, the geocoding system then can associate any attributes that were not entered (e.g., the directional prefix or suffix of the street name), which can be approved and accepted by the staff member if correct, thereby increasing the completeness of the input address data.

In the most accurate and highly advanced scenario possible, the generated point can be displayed on imagery and shown to the patient who then can instruct the staff member in placing the point exactly in the center of their roofline, instead of at its original estimated location. Of course, this may not be desired or appropriate in all cases, such as when people are afraid to disclose their address for one reason or another. Further, if this strategy were employed without properly ensuring the patient's confidentiality (e.g., performing all geocoding and mapping behind a firewall), explaining the technology that enables it (e.g., any of the variety of ArcGIS products and data layers), and what the goal of using it was (i.e., to improve cancer data with the hopes of eventually providing better prevention, diagnosis, and care), it would be understandable for patients to be uncomfortable and think that the diagnosing facility was infringing on their right to privacy.

As described, this process may be impossible for several reasons. It may take a substantially greater amount of time to perform than what is available. It assumes that an on-demand, case-by-case geocoder is available to the staff member, which may not be a reality for registries geocoding with vendors. If available, its use may be precluded by privacy or confidentiality concerns or constraints, or the reference datasets used may not be the correct type or of sufficient quality to achieve the desired level of accuracy. This scenario assumes a great deal of technical competence and an in-depth understanding of the geocoding process that the staff member may not possess. If this approach were to become widely adopted, further questions would be raised as to if and/or when residential history information also should be processed.

2.3.3 Considerations

These two scenarios illustrate that there are indeed benefits and drawbacks associated with performing the geocoding process at different stages of data gathering. When deciding which option to choose, an organization also should take note of any trends in past performance that can be leveraged or used to indicate future performance. For example, if less than 1 percent of address data fails batch-mode processing and requires very little of a staff member's time to manually correct, it may be worth doing. However, if 50 percent of the data fail and the staff member is spending the majority of his or her time correcting erroneously or ambiguously entered data, another option in which input address validation is performed closer to the level of patient contact might be worth considering, but would require more individuals trained in geocoding—specifically the address validation portion (see Section 6.4 for more detail)—at more locations (e.g., hospitals and doctor's offices). These types of tradeoffs will need to be weighed carefully and discussed between both the central registry and facility before a decision is made.

At a higher level, it also is useful to consider the roles of each of the two organizations involved: (1) the data submitters, and (2) the central registries. It can be argued that, ideally, the role of the data submitter is simply to gather the best raw data they can while they are in contact with the patient (although this may not be what actually occurs in practice for a variety of reasons), while the central registries are responsible for ensuring a standardized process for turning the raw data into its spatial counterpart. Even if the data submitters perform geocoding locally before submitting the data to the central registries, the geocoded results may be discarded and the geocoding process applied again upon consolidation by the central registry to maintain consistency (in terms of geocoding quality due to the geocoding process used) amongst all geocodes kept at the central registry. However, even in currently existing and used standards, diagnosing facilities (and/or data submitters) are responsible for

some of the spatial fields in a record (e.g., the county), so the lines between responsibilities have already been blurred for some time.

Best Practices 7 – When to geocode

Policy Decision	Best Practice
Where and when is geocoding used?	Geocoding should be performed as early as possible (i.e., as soon as the address data become available), wherever the data are obtained. Metadata should describe where the geocoding took place: • Diagnosing facility • Central registry. Metadata should describe when the geocoding took place: • Upon intake • Upon transfer from a single registry • Upon consolidation from multiple registries • Every time it is used for analysis.

2.4 SUCCESS STORIES

The use of geocoding and geocoded data in health- and cancer-related research has a long, vivid, and exciting history, stretching back many years (e.g., the early attempts in Howe [1986]). The use of automated geocoding to facilitate spatial analyses in cancer research has enabled entirely new modes of inquiry that were not possible or feasible prior to automated geocoding. Several exemplary applications are noted here to illustrate the potential of the technique and the success that can be achieved. For a more comprehensive review of research studies that have utilized geocoding as a fundamental component, see the recent review article by Rushton et al. (2006).

Epidemiological investigation into the links between environmental exposure and disease incidence rely heavily on geocoded data and are particularly sensitive to the accuracy that can be obtained through the different methods and data sources that can be employed. For example, a whole series of studies investigating ambient pesticide exposure in California's Central Valley all have used geocoding as the fundamental component for identifying the locations of individuals living near pesticide application sites (e.g., Bell et al. 2001; Rull and Ritz 2003; Rull et al. 2001, 2006a, 2006b; Reynolds et al. 2005; Marusek et al. 2006; Goldberg et al. 2007b; and Nuckols et al. 2007). Due to the rapid distance decay inherent in these environmental factors, a high level of spatial accuracy was necessary to obtain accurate exposure estimates.

Likewise, in a currently ongoing study, Cockburn et al. (2008) have uncovered evidence that the risk of mesothelioma with proximity to the nearest freeway (assessing the possible impact of asbestos exposure from brake and clutch linings) is two-fold higher for residences within 100 m of a freeway than those over 500 m away, using linear-interpolation geocoding based on TIGER/Line files (U.S. Census Bureau 2008d). However, when comparing distances to freeways obtained from TIGER/Line file geocodes to those obtained from a parcel-based interpolation approach, it was shown that 24 percent of the data points had parcel-based geocode freeway distances in excess of 500 m greater than those derived from TIGER/Line files. This means that up to 24 percent of the data were misclassified in the original

analysis. If the misclassification varied by case/control status (under examination), then the true relative risk is likely very different from what was observed (biased either to or away from the null).

In addition to its role in exposure analysis, geocoding forms a fundamental component of research studies investigating distance and accessibility to care (Armstrong et al. 2008). These studies typically rely on geocoded data for both a subject's address at diagnosis (dxAddress) and the facility at which they were treated. With these starting and ending points, Euclidean, Great Circle, and/or network distance calculations can be applied to determine both the distance and time that a person must travel to obtain care. Studies using these measures have investigated such aspects as disparities in screening and treatment (Stitzenberg et al. 2007), the affects of distance on treatment selection and/or outcomes (e.g., Nattinger et al. 2001, Stefoski et al. 2004, Voti et al. 2005, Feudtner et al. 2006, and Lianga et al. 2007), and for targeting regions for prevention and control activities (e.g., Rushton et al. 2004).

3. GEOGRAPHIC INFORMATION SCIENCE FUNDAMENTALS

This section introduces the fundamental geographic principles used throughout the remainder of the document, as well as common mistakes that often are encountered.

3.1 GEOGRAPHIC DATA TYPES

In general, GIS data are either vector- or raster-based. **Vector-based data** consist of **vector objects or features** and rely on points and discrete line segments to specify the locations of real-world entities. The latter are simply phenomena or things of interest in the world around us (i.e., a specific street like Main Street) that cannot be subdivided into phenomena of the same kind (i.e., more streets with new names). Vector data provide information relative to where everything occurs—they give a location to every object—but vector objects do not necessarily fill space, because not all locations need to be referenced by objects. One or more attributes (like street names in the aforementioned example) can be assigned to individual objects to describe what is where with vector-based data. **Raster-based data,** in contrast, divide the area of interest into a regular grid of cells in some specific sequence, usually row-by-row from the top left corner. Each cell is assigned a single value describing the phenomenon of interest. Raster-based data provide information relative to what occurs everywhere—they are space filling because every location in an area of interest corresponds to a cell in the raster—and as a consequence, they are best suited for representing things that vary continuously across the surface of the Earth.

Most geocoding applications work with vector-based GIS data. The fundamental primitive is the **point,** a 0-dimensional (0-D) object that has a position in space but no length. Geographic objects of increasing complexity can be created by connecting points with straight or curved lines. A **line** is a 1-D geographic object having a length and is composed of two or more 0-D point objects. Lines also may contain other descriptive attributes that are exploited by geocoding applications such as direction, whereby one end point or node is designated as the start node and the other is designated as the end node. A **polygon** is a geographic object bounded by at least three 1-D line objects or segments with the requirement that they must start and end at the same location (i.e., node). These objects have a length and width, and from these properties one can calculate the area. Familiar 2D shapes such as squares, triangles, and circles are all polygons in vector-based views of the world around us.

Most GIS software supports both vector- and raster-based views of the world, and any standard GIS textbook can provide further information on both the underlying principles and strengths and weaknesses of these complementary data models. The key aspects from a geocoding perspective relative to the methods used to: (1) determine and record the locations of these objects on the surface of the Earth, and (2) calculate distance because many geocoding algorithms rely on one or more forms of linear interpolation.

3.2 GEOGRAPHIC DATUMS AND GEOGRAPHIC COORDINATES

The positions or locations of objects on the surface of Earth are represented with one or more **coordinate systems.** Specifying accurate x and y coordinates for objects is

fundamental for all GIS software and location-based services. However, many different coordinate systems are used to record location, and one often needs to transform data in a GIS from one reference system to another.

There are three basic options: (1) global systems, such as latitude and longitude, are used to record position anywhere on Earth's surface; (2) regional or local systems that aim to provide accurate positioning over smaller areas; and (3) postal codes and cadastral reference systems that record positions with varying levels of precision and accuracy. The reference system to be used for a particular geocoding application and accompanying GIS project will depend on the purpose of the project and how positions were recorded in the source data.

There usually is a geodetic datum that underpins whatever reference system is used or chosen. Most modern tools (i.e., GPS receivers) and data sources (i.e., U.S. Geological Survey National Map, U.S. Census Bureau TIGER/Line files) rely on the North American Datum of 1983 (NAD-83). This and other datums in use in various parts of the world provide a reference system against which horizontal and/or vertical positions are defined. It consists of an ellipsoid (a model of the size and shape of Earth that accounts for the slight flattening at the poles and other irregularities) and a set of point locations precisely defined with reference to that surface.

Geographic coordinates, which specify locations in terms of latitude and longitude, constitute a very popular reference system. The Prime Meridian (drawn through Greenwich, England) and Equator serve as reference planes to define latitude and longitude. **Latitude** is the angle from the plane at the horizontal center of the ellipsoid, the Equator, to the point on the surface of the ellipsoid (at sea level). **Longitude** is the angle between the plane at the vertical center of the ellipsoid, the meridian, and the point on the surface of the ellipsoid. Both are recorded in degrees, minutes, and seconds or decimal degrees.

3.3 MAP PROJECTIONS AND REGIONAL REFERENCE SYSTEMS

Several different systems are used regionally to identify geographic positions. Some of these are true coordinate systems, such as those based on the Universal Transverse Mercator (UTM) or Universal Polar Stereographic (UPS) map projections. Others, such as the Public Land Survey System (PLSS) used widely in the Western United States, simply partition space into blocks. The systems that incorporate some form of map projection are preferred if the goal is to generate accurate geocoding results. A **map projection** is a mathematical function to transfer positions on the surface of Earth to their approximate positions on a flat surface (i.e., a computer monitor or paper map). Several well-known projections exist; the differences between them generally are determined by which property of the Earth's surface they seek to maintain with minimal distortion (e.g., distance, shape, area, and direction). Fortunately, a great deal of time and effort has been expended to identify the preferred map projections in many/most parts of the world.

Hence, the State Plane Coordinate System (SPC) was developed by U.S. scientists in the 1930s to provide local reference systems tied to a national geodetic datum. Each state has its own SPC system with specific parameters and projections. Smaller states such as Rhode Island use a single SPC zone; larger states such as California and Texas are divided into several SPC zones. The SPC zone boundaries in the latter cases typically follow county boundaries. The initial SPC system was based on the North American Datum of 1927 (NAD-27) and the coordinates were recorded in English units (i.e., feet). Some maps using NAD-27 coordinates are still in use today.

Improvements in the measurements of both the size and shape of Earth and of positions on the surface of Earth itself led to numerous efforts to refine these systems, such that the

1927 SPC system has been replaced for everyday use by the 1983 SPC system. The latter is based on the NAD-83 and the coordinates are expressed in metric units (i.e., meters). The 1983 SPC system used Lambert Conformal Conic projections for regions with larger east-west than north-south extents (e.g., Nebraska, North Carolina, and Texas); the Transverse Mercator projections were used for regions with larger north-south extents (e.g., Illinois and New Hampshire). There are exceptions—Florida, for example, uses the Lambert Conformal Conic projection in its north zone and the Transverse Mercator projection in its west and east zones. Alaska uses a completely different Oblique Mercator projection for the thin diagonal zone in the southeast corner of the state.

The choice of map projection and the accompanying coordinate system may have several consequences and is a key point to keep in mind during any aspect of the geocoding process because distance and area calculations required for geocoding rely on them. The most common mistake made from not understanding or realizing the distinctions between different coordinate systems occurs during distance calculations. Latitude and longitude record angles and the utilization of Euclidean distance functions to measure distances in this coordinate system is not appropriate. Spherical distance calculations should be used in these instances. The simpler Euclidean calculations are appropriate at a local scale because the distortion caused by representing positions on a curved surface on a flat computer monitor and/or paper map is minimized. Some special care may be needed if/when the distance calculations extend across two or more SPC zones given the way positions are recorded in northings and eastings relative to some local origin. Some additional information on these types of complications can be gleaned from standard GIS textbooks. Best practices relating to geographic fundamentals are listed in Best Practices 8.

Best Practices 8 – Geographic fundamentals

Policy Decision	Best Practice
What information should be kept about attributes of a reference dataset?	All metadata should be maintained about the type and lineage of the reference data (e.g., coordinate system, projection).
What coordinate system should reference data be kept in?	Reference data should be kept in a Geographic Coordinate System using the North American Datum of 1983 (NAD- 1983) and projected when it needs to be displayed or have distance-based calculations performed. If a projected coordinate system is required, an appropriate one for the location/purpose should be used.
What projection should be used to project reference data?	An appropriate projection should be chosen based on the geographic extent of the area of interest and/or what the projected data are going to be used for. For further information, see any basic GIS textbook. In general: • For most cancer maps, use an equal area projection. • For maps with circular buffers, use a conformal projection. • For calculating distances, use a projection that minimizes distance error for the area of interest.
What distance calculations should be used?	In a projected coordinate space, planar distance metrics should be used. In a non-projected (geographic) coordinate space, spherical distance metrics should be used.

Part 2: The Components of Geocoding

Geocoding is an extremely complicated task involving multiple processes and datasets all simultaneously working together. Without a fundamental understanding of how these pieces all fit together, intelligent decisions regarding them are impossible. This part of the document will first look at the geocoding process from a high level, and subsequently perform a detailed examination of each component of the process.

This page is left blank intentionally.

4. ADDRESS GEOCODING PROCESS OVERVIEW

This section identifies types of geocoding process and outlines the high-level geocoding process, illustrating the major components and their interactions.

4.1 TYPES OF GEOCODING PROCESSES

Now that geocoding has been defined and placed within the context of the larger concept of spatial analysis, the technical background that makes the process possible will be presented. The majority of the remainder of this document will focus on the predominant type of geocoding performed throughout the cancer research community, software-based geocoding. **Software-based geocoding** is a geocoding process in which a significant portion of the components are software systems. From this point forward unless otherwise stated, the term "geocoder" will refer to this particular arrangement.

The software-based geocoding option is presently by far the most economical option available to registries and is the most commonly used option. This document will seek to inform specific decisions that must be made with regard to software-based geocoding. However, information will be relevant to other geocoder processes that utilize other tools (e.g., GPS devices or identification and coordinate assignment from aerial imagery). The accuracy and metadata reporting discussions in particular will be applicable to all types of geocoding process definitions.

In the following sections, the fundamental components of the geocoding process will be introduced. The discussion will provide a high-level description of the components in the geocoding process and their interactions will be offered to illustrate the basic steps that a typical geocoder performs as it produces output from the input provided. Each of these steps, along with specific issues and best practice recommendations related to them, will be described in greater detail in the sections that follow. Additional introductory material on the overall geocoding process, components, and possible sources of error can be found in Armstrong and Tiwari (2008) and Boscoe (2008). The theoretical background presented in the following sections can be grounded by reviewing the case study of the detailed specific geocoding practices and products used in the New Jersey State Cancer Registry (NJSCR) (as well as several other registries) available in Abe and Stinchcomb (2008).

4.2 HIGH-LEVEL GEOCODING PROCESS OVERVIEW

At the highest level, most generalized geocoding processes involve three separate yet related components: (1) the descriptive locational input data (e.g., addresses); (2) the geocoder; and (3) the spatial output data. These high-level relationships are illustrated in Figure 2.

The **input data** to the geocoding process can be any descriptive locational textual information such as an address or building name. The **output** can be any form of valid spatial data such as latitude and longitude. **Geocoding** is the process used to convert the input into the output, which is performed by the **geocoder.**

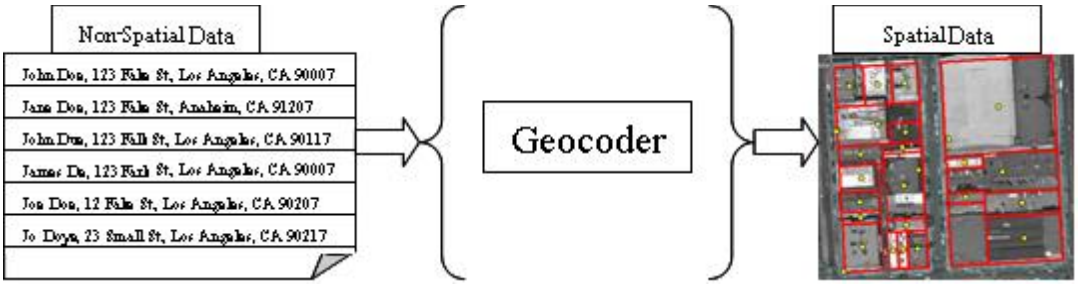


Figure 2 – High-level data relationships

4.3 SOFTWARE-BASED GEOCODERS

A software-based geocoder is composed of two fundamental components. These are the **reference dataset** and the **geocoding algorithm,** each of which may be composed of a series of sub-components and operations. The geocoding process with these new relationships is depicted in Figure 3.

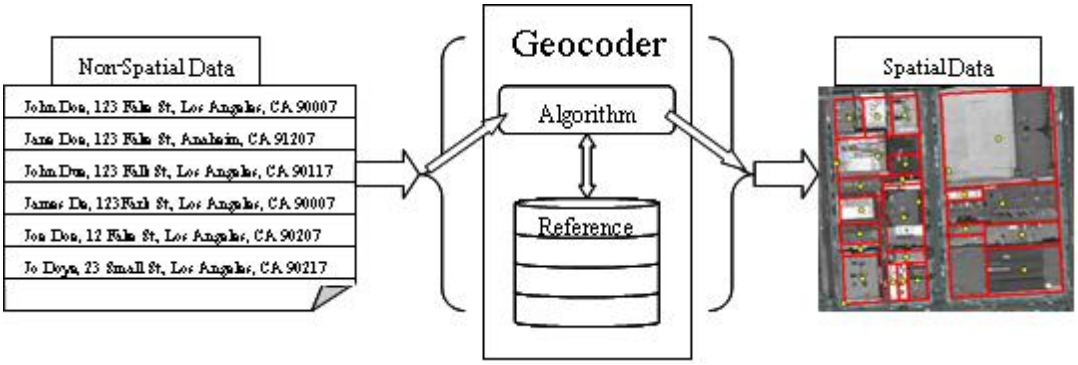


Figure 3 - Schematic showing basic components of the geocoding process

It is likely that the actual software implementation of a geocoder will vary in the nature of the components chosen and conceptual representation of the geocoding system. Each registry will have its own **geocoding requirements,** or set of limitations, constraints, or concerns that influence the choice of a particular geocoding option. These may be technical, budgetary, legal, or policy related and will necessarily guide the choice of a geocoding process. Best practices related to determining geocoding requirements are listed in Best Practices 9. Even though the geocoding requirements may vary between registries, the NAACCR standards for data reporting spatial fields as defined in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008) should be followed by all registries to ensure uniformity across registries.

Best Practices 9 – Geocoding requirements

Policy Decision	Best Practice
What considerations should affect the choice of geocoding requirements?	Technical, budgetary, legal, and policy constraints should influence the requirements of a geocoding process.
When should requirements be reviewed and/or changed?	Ideally, requirements should be revisited annually but registries may have constraints that extend or shorten this time period.

The general process workflow shown in Figure 4 represents a generalized abstraction of the geocoding process. It illustrates the essential components that should be common to any geocoder implementation and sufficient for registries with few requirements, while being detailed enough to illustrate the decisions that must be made at registries with many detailed requirements. This conceptualization also is sufficient to illustrate the generalized steps and requirements that geocoder vendors will need to accommodate to work with registries.

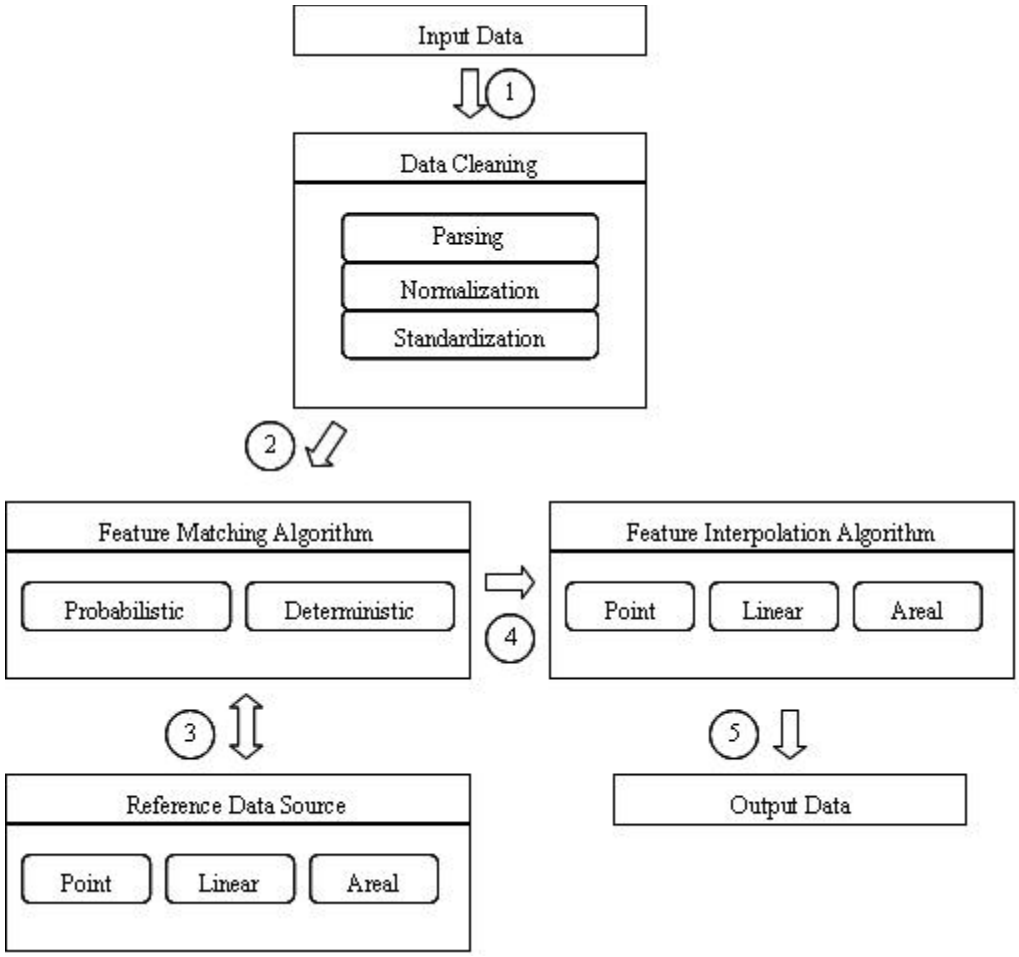


Figure 4 – Generalized workflow

4.4 INPUT DATA

Input data are the descriptive locational texts that are to be turned into computer-useable spatial data by the process of geocoding. As indicated earlier (Table 3), the wide variety of possible forms and formats of input data is the main descriptor of a geocoder's flexibility, as well as a contributing factor to the overall difficulty of implementing geocoding.

4.4.1 Classifications of input data

Input data can first be classified into two categories, relative and absolute. **Relative input data** are textual location descriptions which, by themselves, are not sufficient to produce an output geographic location. These produce **relative geocodes** that are geographic locations relative to some other reference geographic locations (i.e., based on an interpolated distance along or within a reference feature in the case of line vectors and areal units, respectively). Without reference geographic locations (i.e., the line vector or areal unit), the output locations for the input data would be unobtainable.

Examples of these types of data include "Across the street from Togo's" and "The northeast corner of Vermont Avenue and 36th Place." These are not typically considered valid address data for submission to a central registry, but they nonetheless do occur. The latter location, for instance, cannot be located without identifying both of the streets as well as the cardinal direction in which one must head away from their exact intersection. Normal postal street addresses also are relative. The address "3620 South Vermont Avenue" is meaningless without understanding that "3620" denotes a relative geographic location somewhere on the geographic location representing "Vermont Avenue." It should be noted that in many cases, geocoding platforms do not support these types of input and thus may not be matchable, but advances in this direction are being made.

Absolute input data are textual location descriptions which, by themselves, are sufficient to produce an output geographic location. These input data produce an **absolute geocode** in the form of an absolute known location or an offset from an absolute known location. Input data in the form of adequately referenced placenames, USPS ZIP Codes, or parcel identifiers are examples of the first because each can be directly looked up in a data source (if available) to determine a resulting geocode.

Locations described in terms of linear addressing systems also are absolute by definition. For example, the Emergency 911-based (E-911) geocoding systems being mandated in rural areas of the United States are (in many cases) absolute because they use distances from known mileposts on streets as coordinates. These mileposts are a linear addressing system because each represents an absolute known location. It should be noted that in some cases, this may not be true because the only implementation action taken to adhere to the E-911 system was street renaming or renumbering.

With these distinctions in mind, it is instructive at this point to classify and enumerate several commonly encountered forms of input data that a geocoder can and must be able to handle in one capacity or another, because these may be the only information available in the case in which all other fields in a record are null. This list is presented in Table 4.

Table 4 – Common forms of input data with corresponding NAACCR fields and example values

Type	NAACCR Field(s)	Example
Complete postal address	2330: dxAddress - Number and Street 70: dxAddress - City 80: dxAddress - State 100: dxAddress - Postal Code	3620 S Vermont Ave, Unit 444, Los Angeles, CA 90089
Partial postal address	2330: dxAddress - Number and Street	3620 Vermont
USPS PO box	2330: dxAddress - Number and Street 70: dxAddress - City 80: dxAddress - State 100: dxAddress - Postal Code	PO Box 1234, Los Angeles CA 90089-1234
Rural Route	2330: dxAddress - Number and Street 70: dxAddress - City 80: dxAddress - State	RR12, Los Angeles CA
City	70: dxAddress - City	Los Angeles
County	90: County at dx	Los Angeles County
State		CA
USPS ZIP Code, USPS ZIP+4 (United States Postal Service 2008a)	100: dxAddress - Postal Code	90089-0255
Intersection	2330: dxAddress - Supplemental	Vermont Avenue and 36th Place
Named place	2330: dxAddress - Supplemental	University of Southern California
Relative	2330: dxAddress - Supplemental	Northeast corner of Vermont Ave and 36th Pl
Relative	2330: dxAddress - Supplemental	Off Main Rd

From this list, it is apparent that most input data are based on postal addressing systems, administrative units, named places, coordinate systems, or relative descriptions that use one of the others as a referent. Input data in the form of postal addresses, or portions thereof, are by far the most commonly encountered, and as such this document will focus almost exclusively on this input data type. Significant problems may appear when processing postal address data because they are among the "noisiest" forms of data available. As used here, "noisy" refers to the high degree of variability in the way they can be represented, and to the fact that they often include extraneous data and/or are missing required elements. To overcome these problems, geocoders usually employ two techniques known as **address normalization** and **address standardization.**

4.4.2 Input data processing

Address normalization organizes and cleans input data to increase its efficiency for use and sharing. This process attempts to identify the component pieces of an input address (e.g., street number, street name, or USPS ZIP Code) within the input string. The goal is to identify the correct pieces in the input data so that it will have the highest likelihood of being successfully assigned a geocode by the geocoder. In Table 5, several forms of the same address are represented to illustrate the need for address normalization.

Table 5 – Multiple forms of a single address

Sample Address
3620 South Vermont Avenue, Unit 444, Los Angeles, CA 90089-0255
3620 S Vermont Ave, #444, Los Angeles, CA 90089-0255
3620 S Vermont Ave, 444, Los Angeles, 90089-0255
3620 Vermont, Los Angeles, CA 90089

Address standardization converts an address from one normalized format into another. It is closely linked to normalization and is heavily influenced by the performance of the normalization process. Standardization converts the normalized data into the correct format expected by the subsequent components of the geocoding process. Address standards may be used for different purposes and may vary across organizations because there is no single, set format; however, variability in formats presents a barrier to data sharing among organizations. Interoperability assumes an agreement to implement a standardized format. In Table 6, several existing or proposed address standards are listed. Best practices related to input data are listed in Best Practices 10.

Table 6 – Existing and proposed address standards

Organization	Standard
USPS	Publication 28 (United States Postal Service 2008d)
Urban and Regional Information Systems Association (URISA)/United States Federal Geographic Data Committee (FGDC)	Street Address Data Standard (United States Federal Geographic Data Committee 2008b)

Best Practices 10 – Input data (high level)

Policy Decision	Best Practice
What type of input data can and should be geocoded?	At a minimum, NAACCR standard address data should be able to be geocoded. Ideally, any type of descriptive locational data, both relative and absolute, in any address standard should be an acceptable type of input and geocoding can be attempted: • Any form of postal address • Intersections • Named places • Relative locations.
What type of relative input data can and should be geocodable?	At a minimum, postal street addresses. Ideally, relative directional descriptions.
What type of absolute input data can and should be geocodable?	At a minimum, E-911 locations (if they are absolute).
What type of normalization can and should be performed?	Any reproducible technique that produces certifiably valid results should be considered a valid normalization practice: • Tokenization • Abbreviation (introduction/substitution).
What type of standardization can and should be performed?	Any reproducible technique that produces certifiably valid results should be considered a valid standardization practice.

4.5 REFERENCE DATASETS

The **reference dataset** is the underlying geographic database containing geographic features that the geocoder can use to generate a geographic output. This dataset stores all of the information the geocoder knows about the world and provides the base data from which the geocoder calculates, derives, or obtains geocodes. Interpolation algorithms (discussed in the next section) perform computations on the reference features contained in these datasets to estimate where the output of the geocoding process should be placed (using the attributes of the input address).

Reference datasets are available in many forms and formats. The sources of these data also vary greatly from local government agencies (e.g., tax assessors) to national governmental organizations (e.g., the Federal Geographic Data Committee [FGDC]). Each must ultimately contain valid spatial geographic representations that either can be returned directly in response to a geocoder query (as the output) or be used by other components of the geocoding process to deduce or derive the spatial output. A few examples of the numerous types of geographic reference data sources that may be incorporated into the geocoder process are listed in Table 7, with best practices listed in Best Practices 11.

Table 7 – Example reference datasets

Type	Example
Vector line file	U.S. Census Bureau's TIGER/Line (United States Census Bureau 2008c)
Vector polygon file	Los Angeles (LA) County Assessor Parcel Data (Los Angeles County Assessor 2008)
Vector point file	Australian Geocoded National Address File (G-NAF) (Paull 2003)

Best Practices 11 – Reference data (high level)

Policy Decision	Best Practice
What types of reference datasets can and should be supported by a geocoder?	Linear-, point-, and polygon-based vector reference datasets should be supported by a geocoding system.

4.6 THE GEOCODING ALGORITHM

The **geocoding algorithm** is the main computational component of the geocoder. This algorithm can be implemented in a variety of ways, especially if trends about the input data or reference dataset can be determined *a priori*.

Generally speaking, any algorithm must perform two basic tasks. The first, **feature matching,** is the process of identifying a geographic feature in the reference dataset corresponding to the input data to be used to derive the final geocode output for an input. A **feature-matching algorithm** is an implementation of a particular form of feature matching. These algorithms are highly dependent on both the type of reference dataset utilized and the attributes it maintains about its geographic features. The algorithm's chances of selecting the correct feature vary with the number of attributes per feature. A substantial part of the overall quality of the output geocodes rests with this component because it is responsible for identifying and selecting the reference feature used for output derivation.

The next task, **feature interpolation,** is the process of deriving a geographic output from a reference feature selected by feature matching. A **feature interpolation algorithm** is an implementation of a particular form of feature interpolation. These algorithms also are highly dependent on the reference dataset in terms of the type of data it contains and the attributes it maintains about these features.

If one were to have a reference dataset containing valid geographic points for every address in one's study area (e.g., the ADDRESS-POINT [Higgs and Martin 1995a, Ordnance Survey 2008] and G-NAF [Paull 2003] databases), the feature interpolation algorithm essentially returns this spatial representation directly from the reference dataset. More often, however, the interpolation algorithm must estimate where the input data should be located with reference to a feature in the reference dataset. Typical operations include linear or areal interpolation (see Section 8) when the reference datasets are street vectors and parcel polygons, respectively.

Best Practices 12 – Geocoding algorithm (high level)

Policy Decision	Best Practice
What type of geocoding can and should be performed?	At a minimum, software-based geocoding should be performed.
What forms of feature matching should the geocoding algorithm include?	The geocoding algorithm should consist of feature-matching algorithms consistent with the forms of reference data the system supports. Both probability-based and deterministic methods should be supported.
What forms of feature interpolation should the geocoding algorithm include?	The geocoding algorithm should consist of feature interpolation algorithms consistent with the forms of reference data the system supports (e.g., linear-based interpolation if linear-based reference datasets are used).

4.7 OUTPUT DATA

The last component of the geocoder is the actual **output data,** which are the valid spatial representations derived from features in the reference dataset. As defined in this document, these data can have many different forms and formats, but each must contain some type of valid spatial attribute.

The most common format of output is points described with geographic coordinates (latitude, longitude). However, the accuracy of these spatial representations suffers when they are interpolated, due to data loss during production. Alternate forms can include multi-point representations such as polylines or polygons. As noted, registries must at least report the results in the formats explicitly defined in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008).

When using these output data, one must consider the geographic resolution they represent in addition to the type of spatial geometry. For example, a point derived by areal interpolation from a polygon parcel boundary should not be treated as equivalent to a point derived from the aerial interpolation of a polygon USPS ZIP Code boundary (note that a USPS ZIP Code is not actually an areal unit; more details on this topic can be found in Section 5.1.4). These geocoder outputs, while in the same format and produced through the same process, do not represent data at the same geographic resolution and must be differentiated.

Best Practices 13 – Output data (high level)

Policy Decision	Best Practice
What forms and formats can and should be returned as output data?	The geocoding process should be able to return any valid geographic object as output. At a minimum, outputs showing the locations of point should be supported.
What can and should be returned as output?	The geocoding process should be able to return the full feature it matched to (e.g., parcel polygon if matching to a parcel reference dataset), in addition to an interpolated version.

4.8 METADATA

Without proper metadata describing each component of the geocoding process and the choices that were made at each step, it is nearly impossible to have any confidence in the quality of a geocode. With this in mind, it is recommended that all geocodes contain all relevant information about all components used in the process as well as all decisions that each component made. Table 8**,** Table 9, and Table 10 list example geocoding component, process, and record metadata items. These lists are not a complete enumeration of every metadata item for every combination of geocoding component, method, and decision, nor do they contain complete metadata items for all topics described in this document. These lists should serve as a baseline starting point from which registries, geocode developers, and/or vendors can begin discussion as to which geocode component information needs documentation. Details on each of the concepts listed in the following tables are described later in the document. Component- and decision-specific metadata items for these and other portions of the geocoding process are listed in-line through this document where applicable. The creation and adoption of similar metadata tables describing the complete set of geocoding topics covered in this document would be a good first step toward the eventual cross-registry standardization of geocoding processes; work on this task is currently underway.

Table 8 – Example geocoding component metadata

Component	Item	Example Value
Input data	Normalizer	Name of normalizer
	Normalizer version	Version of normalizer
	Normalizer strategy	Substitution-based
		Context-based
		Probability-based
	Standardizer	Name of standardizer
	Standardizer version	Version of standardizer
	Standard	Name of standard
Reference dataset	Dataset type	Lines
		Points
		Polygons
	Dataset name	TIGER/Line files
	Dataset age	2008
	Dataset version	2008b
Feature matching	Feature matcher	Name of feature matcher
	Feature matcher version	Version of feature matcher
	Feature-matching strategy	Deterministic
		Probabilistic
Feature interpolation	Feature interpolator	Name of feature interpolator
	Feature interpolator version	Version of feature interpolator
	Feature interpolator strategy	Address-range interpolation
		Uniform lot interpolation
		Actual lot interpolation
		Geometric centroid
		Bounding box centroid
		Weighted centroid

Table 9 – Example geocoding process metadata

Component	Decision	Example Value
Substitution-based normalization	Substitution table	USPS Publication 28 abbreviations
	Equivalence function	Exact string equivalence
		Case-insensitive equivalence
Probabilistic feature matching	Confidence threshold	95%
	Probability function	Match-unmatch probability
	Attribute weights	Weight values
Uniform lot interpolation	Dropback distance	6m
	Dropback direction	Reference feature orthogonal

Table 10 – Example geocoding record metadata

Component	Decision	Example Value
Substitution-based normalization	Original data	3620 So. Vermont Av
	Normalized data	3620 S Vermont Ave
Probabilistic feature matching	Match probability	95%
	Unmatch probability	6%
Uniform lot interpolation	Number of lots on street	6
	Lot width	Street length proportional

This page is left blank intentionally.

5. ADDRESS DATA

This section presents an in-depth, detailed examination of the issues specifically related to address data including the various types that are possible, estimates of their accuracies, and the relationships between them.

5.1 TYPES OF ADDRESS DATA

Postal address data are the most common form of input data encountered. They can take many different forms, each with its own inherent strengths and weaknesses. These qualities are directly related to the amount of information that is encoded. There also are specific reasons for the existence of each, and in some cases, plans for its eventual obsolescence. Several of the commonly encountered types will be described through examples and illustrations. Each example will highlight differences in possible resolutions that can be represented and first-order estimates of expected levels of accuracy.

5.1.1 City-Style Postal Addresses

A **city-style postal address** describes a location in terms of a numbered building along a street. This address format can be described as consisting of a number of attributes that when taken together uniquely identify a postal delivery site. Several examples of traditional postal addresses for both the United States and Canada are provided in Table 11.

Table 11 – Example postal addresses

Example Address
2121 West White Oaks Drive, Suite C, Springfield, IL, 62704-6495
1600 Pennsylvania Ave NW, Washington, DC 20006
Kaprielian Hall, Unit 444, 3620 S. Vermont Ave, Los Angeles, CA, 90089-0255
490 Sussex Drive, Ottawa, Ontario K1N 1G8, Canada

One of the main benefits of this format is the highly descriptive power it provides (i.e., the capability of identifying locations down to sub-parcel levels). In the United States, the attributes of a city-style postal address usually include a house number and street name, along with a city, state, and USPS ZIP Code. Each attribute may be broken down into more descriptive levels if they are not sufficient to uniquely describe a location. For example, unit numbers, fractional addresses, and/or USPS ZIP+4 Codes (United States Postal Service 2008a) are commonly used to differentiate multiple units sharing the same property (e.g., 3620 Apt 1, 3620 Apt 6E, 3620 ½, or 90089-0255 [which identifies Kaprielian Hall]). Likewise, pre- and post-directional attributes are used to differentiate individual street segments when several in the same city have the same name and are within the same USPS ZIP Code. This case often occurs when the origin of the address range of a street is in the center of a city and expands outward in opposite directions (e.g., the 100 North [longer arrow pointing up and to the left] and 100 South [shorter arrow pointing down and to the right] Sepulveda Boulevard blocks, as depicted in Figure 5).

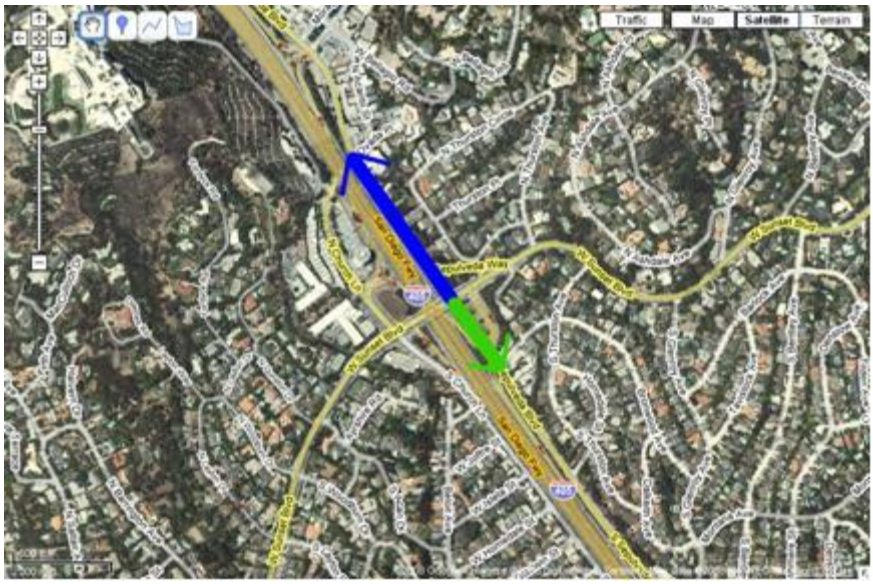


Figure 5 – Origin of both the 100 North (longer arrow pointing up and to the left) and 100 South (shorter arrow pointing down and to the right) Sepulveda Boulevard blocks (Google, Inc. 2008b)

Also, because this form of input data is so ubiquitous, suitable reference datasets and geocoders capable of processing it are widely available at many different levels of accuracy, resolution, and cost. Finally, the significant body of existing research explaining geocoding processes based upon this format make it an enticing option for people starting out.

However, several drawbacks to using data in the city-style postal address format exist. These drawbacks are due to the multitude of possible attributes that give these addresses their descriptive power. When attributes are missing, not ordered correctly, or if extraneous information has been included, significant problems can arise during feature matching. These attributes also can introduce ambiguity when the same values can be used for multiple attributes. For instance, directional and street suffix indicators used as street names can cause confusion as in "123 North South Street" and "123 Street Road." Similar confusion also may arise in other circumstances when numbers and letters are used as street name values as in "123 Avenue 2" and "123 N Street." Non-English-based attributes are commonly encountered in some parts of the United States and Canada (e.g., "123 Paseo del Rey") which further complicates the geocoding process.

A final, more conceptual problem arises due to a class of locations that have ordinary city-style postal addresses but do not receive postal delivery service. An example of this is a private development or gated community. These data may sometimes be the most difficult cases to geocode because postal address-based reference data are truly not defined for them and systems relying heavily on postal address-based normalization or standardization may fail to process them. This also may occur with minor civil division (MCD) names (particularly townships) that are not mailing address components.

5.1.2 Post Office Box Addresses

A USPS **post office (PO) box address** designates a physical storage location at a U.S. post office or other mail-handling facility. By definition, these types of data do not represent residences of individuals, and should not be considered as such. Conceptually, a USPS PO box address removes the street address portion from an address, leaving only a USPS ZIP

Code. Thus, USPS PO box data in most cases can never be geocoded to street-level accuracy. Exceptions to this include the case of some limited mobility facilities (e.g., nursing homes), for which a USPS PO box can be substituted with a street address using lookup tables and aliases. Also, the postal codes used in Canada serve a somewhat similar purpose but are instead among the most accurate forms of input data because of the organization of the Canadian postal system.

In the majority of cases though, it is difficult to determine *anything* about the level of accuracy that can be associated with USPS PO box data in terms of how well they represent the residential location of an individual. As one example, consider the situation in which a person rents a USPS PO box at a post office near their place of employment because it is more convenient than receiving mail at their residence. If the person works in a completely different city than where they live, not even the city attribute of the USPS PO box address can be assumed to correctly represent the person's residential location (or state for that matter when, for example, commuters go to Manhattan, NY, from New Jersey or Connecticut). Similarly, personal mail boxes may be reported and have the same lack of correlation with residence location. Being so frequently encountered, a substantial body of research exists dedicated to the topic of USPS PO boxes and their effect on the geocoding process and studies that use them (e.g., Hurley et al. 2003, Shi 2007, and the references within).

5.1.3 Rural Route and Highway Contract Addresses

A **Rural Route (RR) or Highway Contract (HC) address** identifies a stop on a postal delivery route. This format is most often found in rural areas and is of the form "RR 16 Box 2," which indicates that mail should be delivered to "Box 2" on the rural delivery route "Number 16." These delivery locations can be composed of several physical cluster boxes at a single drop-off point where multiple residents pick up their mail, or they can be single mailboxes at single residences.

Historically, numerous problems have occurred when applying a geocoding process to these types of addresses. First and foremost, an RR by definition is a route traveled by the mail carrier denoting a path, not a single street (similar to a USPS ZIP Code, as will be discussed later). Until recently, it was therefore impossible to derive a single street name from a numbered RR portion of an RR address. Without a street name, feature matching to a reference street dataset is impossible (covered in Section 8.1). Further, the box number attribute of an RR address did not include any data needed for linear-based feature interpolation. There was no indication if a box was not standalone, nor did it relate to and/or inform the relative distance along a reference feature. Thus, it was unquantifiable and unusable in a feature interpolation algorithm.

Recently, however, these difficulties have begun to be resolved due to the continuing implementation of the E-911 service across the United States. In rural areas where RR addresses had historically been the predominant addressing system, any production of the required E-911 geocodes from address geocoding was impossible (for the reasons just mentioned). To comply with E-911 regulations, local governments therefore assigned geocodes to the RR addresses (and their associated phone numbers) based on the existing linear-based referencing system of street mileposts. This led to the creation and availability of a system of absolute geocodes for RR addresses.

Also, for these areas where E-911 had been implemented, the USPS has taken the initiative to create the Locatable Address Conversion System (LACS) database. The primary role of this database is to enable RR to city-style postal street address conversion (United States Postal Service 2008c). The availability of this conversion tool enables a direct link between

an RR postal address and the reference datasets capable of interpolation-based geocoding that require city-style postal addresses. The USPS has mandated that all Coding Accuracy Support System (CASS) certified software providers must support the LACS database to remain certified (United States Postal Service 2008b), so RR to city-style address translation is available now for most areas, but at a cost. Note that USPS CASS-certified systems are only certified to parse and standardize address data into valid USPS data. This certification is in no way a reflection of any form of certification of a geocode produced by the system.

5.1.4 USPS ZIP Codes and U.S. Census Bureau ZCTAs

The problems arising from the differences between the USPS ZIP Codes and the U.S. Census Bureau's ZCTA are widely discussed in the geocoding literature, and so they will only be discussed briefly in this document. In general, the common misunderstanding is that the two refer to the same thing and can be used interchangeably, despite their published differences and the fact that their negative effects on the geocoding process have been widely publicized and documented in the geocoding literature (e.g., Krieger et al. 2002b, Hurley et al. 2003, Grubesic and Matisziw 2006). USPS ZIP Codes represent delivery routes rather than regions, while a ZCTA represents a geographic area. For an excellent review of USPS ZIP Code usage in the literature and a discussion of the differences, effects, and the multitude of ways they can be handled, see Beyer et al. (2008).

Best practices relating to the four types of input postal address data just described are listed in Best Practices 14.

Best Practices 14 – Input data types

Policy Decision	Best Practice
What types of input address data can and should be supported?	Any type of address data should be considered valid geocoding input (e.g., city-style and rural route postal addresses).
What is the preferred input data specification?	If possible, input data should be formatted as city-style postal addresses.
Should USPS PO box data be accepted for geocoding?	If possible, USPS PO box data should be investigated to obtain more detailed information and formatted as city-style postal addresses.
Should RR or HC data be accepted for geocoding?	If possible, RR and HC data should be converted into city-style postal addresses.
Should USPS ZIP Code and/or ZCTA data be accepted for geocoding?	If possible, USPS ZIP Code and/or ZCTA data should be investigated for more detailed information and formatted as a city-style postal address. If USPS ZIP Code and/or ZCTA data must be used, special care needs to be taken when using the resulting geocodes in research, see Beyer et al. (2008) for additional guidance.
Should an input address ever be abandoned and not used for geocoding?	If the potential level of resulting accuracy is too low given the input data specification and the reference features that can be matched, lower level portions of the input data should be used (e.g., USPS ZIP Code, city).

5.2 First-Order Estimates

The various types of input address data are capable of describing different levels of information, both in their best and worst cases. First-order estimates of these values one can expect to achieve in terms of geographic resolution are listed in Table 12.

Table 12 – First-order accuracy estimates

Data type	Best Case	Worst Case
Standard postal address	Sub-parcel	State
USPS PO box	USPS ZIP Code centroid	State
Rural route	Sub-parcel	State
U.S. National Grid	1 m^2	1,000 m^2

5.3 Postal Address Hierarchy

As noted in Section 5.1.1, city-style postal addresses are the most common form encountered in the geocoding process and are extremely valuable given the hierarchical structure of the information they contain. This implicit hierarchy often is used as the basis for multi-resolution geocoding processes that allow varying levels of geographic resolution in the resulting geocodes based on where a match can be made in the hierarchy. This relates directly to the ways in which people communicate and understand location, and is chiefly responsible for enabling the geocoding process to capture this same notion.

The following city-style postal address has all possible attributes filled in (excluding multiple street type suffixes), and will be used to illustrate this progression through the scales of geographic resolution as different attribute combinations are employed:

3620 ½ South Vermont Avenue East, Unit 444, Los Angeles, CA, 90089-0255

The possible variations of this address in order of decreasing geographic resolution (with 0 ranked as the highest) are listed in Table 13. Also listed are the best possible and most probable resolutions that could be achieved, along with the ambiguity introduced at each resolution. Selected resolutions are also displayed visually in Figure 6. The table and figure underscore two observations: (1) the elimination of attributes from city-style postal addresses degrades the best possible accuracy quite rapidly, and (2) different combinations of attributes will have a significant impact on the geographic resolution or granularity of the resulting geocode. More discussion on the strengths and weaknesses of arbitrarily ranking geographic resolutions is presented in Section 15.1.

Table 13 – Resolutions, issues, and ranks of different address types

Address	Best Resolution	Probable Resolution	Ambiguity	Rank
3620 North Vermont Avenue, Unit 444, Los Angeles, CA, 90089-0255	3D Sub-parcel-level	Sub-parcel-level	none	0
3620 North Vermont Avenue, Los Angeles, CA, 90089-0255	Parcel-level	Parcel-level	unit, floor	1
3620 North Vermont Avenue, Los Angeles, CA, 90089	Parcel-level	Parcel-level	unit, floor, USPS ZIP Code	2
3620 Vermont Avenue, Los Angeles, CA, 90089	Parcel-level	Street-level	unit, floor, street, USPS ZIP Code	3
Vermont Avenue, Los Angeles, CA, 90089	Street-level	USPS ZIP Code-level	building, unit, floor, street, USPS ZIP Code	4
90089	USPS ZIP Code-level	USPS ZIP Code-level	building, unit, floor, street, city	5
Vermont Avenue, Los Angeles, CA	City-level	City-level, though small streets may fall entirely into a single USPS ZIP Code	building, unit, floor, street, USPS ZIP Code	6
Los Angeles, CA	City-level	City-level	building, unit, floor, street, USPS ZIP	7
Vermont Avenue, CA	State-level	State-level	building, unit, floor, street, USPS ZIP, city	8
CA	State-level	State-level	building, unit, floor, street, USPS ZIP, city	8

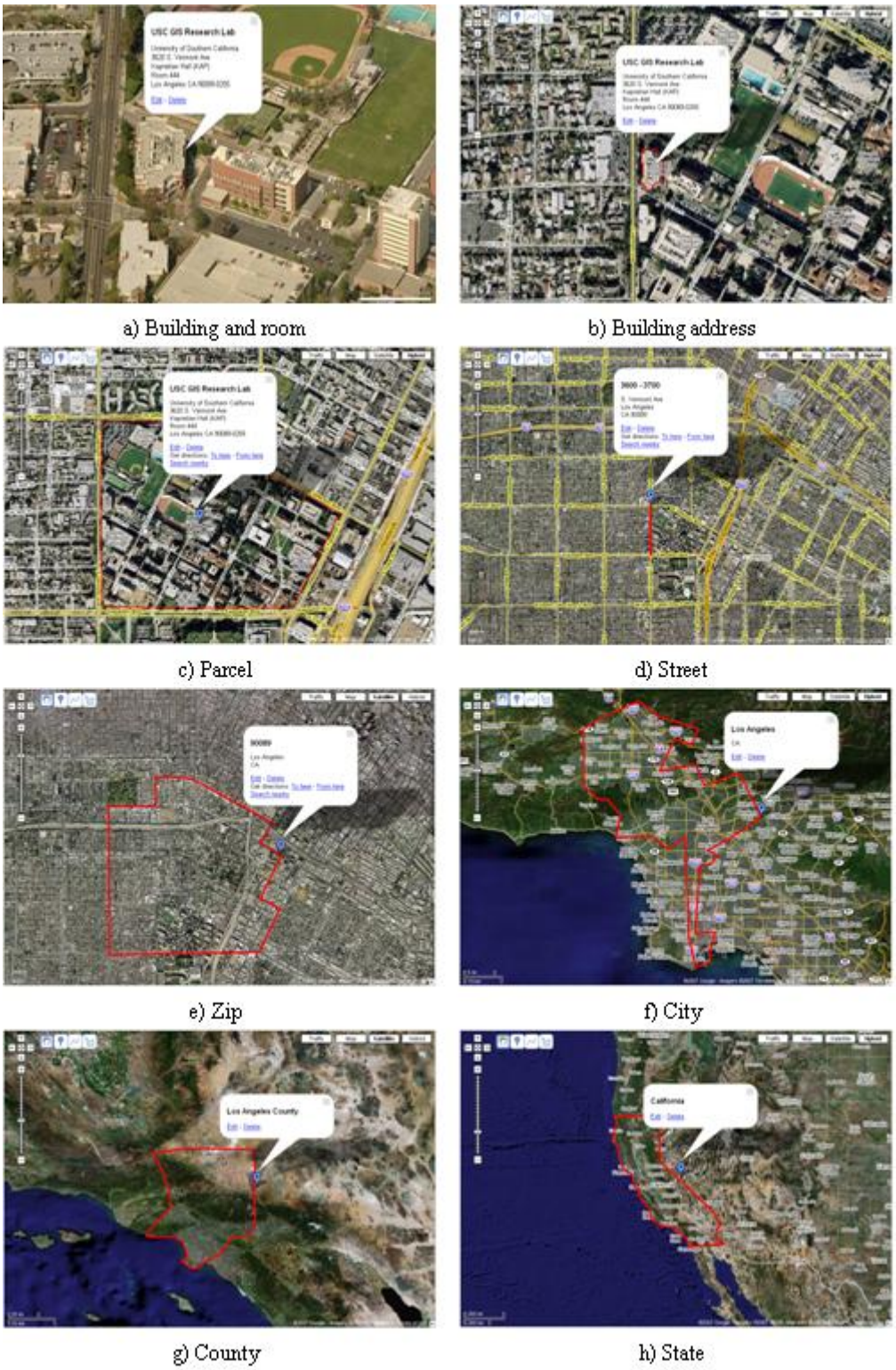


Figure 6 – Geographic resolutions of different address components (Google, Inc. 2008b)

This page is left blank intentionally.

6. ADDRESS DATA CLEANING PROCESSES

This section presents a detailed examination of the different types of processes used to clean address data and discusses specific implementations.

6.1 ADDRESS CLEANLINESS

The "cleanliness" of input data is perhaps the greatest contributing factor to the success or failure of a successful geocode being produced. As Zandbergen concludes, "improved quality control during the original capture of input data is paramount to improving geocoding match rates" (2008, pp. 18). Address data are notoriously "dirty" for several reasons, including simple data entry mistakes and the use of non-standard abbreviations and attribute orderings. The addresses listed in Table 14 all refer to the same address, but are in completely different formats, exemplifying why various address-cleaning processes are required. The address-cleaning processes applied to prepare input address data for processing will be detailed in the next sections.

Table 14 – Example postal addresses in different formats

3620 North Vermont Avenue, Unit 444, Los Angeles, CA, 90089-0255
3620 N Vermont Ave, 444, Los Angeles, CA, 90089-0255
3620 N. VERMONT AVE., UNIT 444, LA, CA
N Vermont 3620, Los Angeles, CA, 90089

6.2 ADDRESS NORMALIZATION

Address normalization is the process of identifying the component parts of an address such that they may be transformed into a desired format. This first step is critical to the cleaning process. Without identifying which piece of text corresponds to which address attribute, it is impossible to subsequently transform them between standard formats or use them for feature matching. The typical component parts of a city-style postal address are displayed in Table 15.

Table 15 – Common postal address attribute components

Number	3620
Prefix Directional	N
Street Name	Vermont
Suffix Directional	
Street Type	Ave
Unit Type	Unit
Unit Number	444
Postal Name (Post Office name, USPS default or acceptable name for given USPS ZIP Code)	
USPS ZIP Code	90089-0255
State	CA

The normalization algorithm must attempt to identify the most likely address attribute to associate with each component of the input address. Decades of computer science research have been invested into this difficult parsing problem. Many techniques can be applied to this problem, some specifically developed to address it and others that were developed for other purposes but are nonetheless directly applicable. These approaches range in their level of sophistication; examples from the simplistic to highly advanced will now be described.

6.2.1 Substitution-Based Normalization

Substitution-based normalization makes use of lookup tables for identifying commonly encountered terms based on their string values. This is the most popular method because it is the easiest to implement. This simplicity also makes it applicable to the fewest number of cases (i.e., only capable of substituting correct abbreviations and eliminating [some] extraneous data).

In this method, **tokenization** converts the string representing the whole address into a series of separate **tokens** by processing it left to right, with embedded spaces used to separate tokens. The original order of input attributes is highly critical because of this linear sequential processing. A typical system will endeavor to populate an internal representation of the parts of the street address listed in Table 15, in the order presented. A set of **matching rules** define the valid content each attribute can accept and are used in conjunction with **lookup tables** that list synonyms for identifying common attribute values.

As each token is encountered, the system tries to **match** it to the next empty attribute in its internal representation, in a sequential order. The lookup tables attempt to identify known token values from common abbreviations such as directionals (e.g., "n" being equal to "North," with either being valid). The matching rules limit the types of values that can be assigned to each attribute. To see how it works, the following address will be processed, matching it to the order of attributes listed in Table 15:

"3620 Vermont Ave, RM444, Los Angeles, CA 90089"

In the first step, a match is attempted between the first token of the address, "3620," and the internal attribute in the first index, "number." This token satisfies the matching rule for this internal attribute (i.e., that the data must be a number), and it is therefore accepted and assigned to this attribute. Next, a match is attempted between the second word, "Vermont," and the address attribute that comprises the second index, the pre-directional. This time, the match will fail because the matching rule for this attribute is that data must be a valid form of a directional, and this word is not. The current token "Vermont" then is attempted to be matched to the next attribute (index 3, street name). The matching rule for this has no restrictions on content, so the token is assigned. The next token, "Ave," has a match attempted with the valid attributes at index 4 (the post-directional), which fails. Another match is attempted with the next address attribute at the next index (5, street type), which is successful, so it is assigned. The remainder of the tokens subsequently are assigned in a similar manner.

It is easy to see how this simplistic method can become problematic when keywords valid for one attribute such as "Circle" and "Drive" are used for others as in "123 Circle Drive West," with neither in the expected position of a street suffix type. Best practices related to substitution-based normalization are listed in Best Practices 15.

Best Practices 15 – Substitution-based normalization

Policy Decision	Best Practice
When should substitution-based normalization be used?	Substitution-based normalization should be used as a first step in the normalization process, especially if no other more advanced methods are available.
Which matching rules should be used in substitution-based normalization?	Any deterministic set of rules that create reproducible results that are certifiably valid should be considered acceptable.
Which lookup tables (substitution synonyms) should be used in substitution-based normalization?	At a minimum, the USPS Publication 28 synonyms should be supported (United States Postal Service 2008d)
Which separators should be used for tokenization?	At a minimum, whitespace should be used as a token separator.
What level of token matching should be used for determining a match or non-match?	At a minimum, an exact character-level match should be considered a match.

6.2.2 Context-Based Normalization

Context-based normalization makes use of syntactic and lexical analysis to identify the components of the input address. The main benefit of this less commonly applied method is its support for reordering input attributes. This also makes it more complicated and harder to implement. It has steps very similar to those taken by a programming language compiler, a tool used by programmers to produce an executable file from plain text source code written in a high-level programming language.

The first step, **scrubbing,** removes illegal characters and white space from the input datum. The input string is scanned left to right and all invalid characters are removed or replaced. Punctuation marks (e.g., periods and commas) are all removed and all white-space characters are collapsed into a single space. All characters then are converted into a single common case, either upper or lower. The next step, **lexical analysis,** breaks the scrubbed string into typed tokens. Tokenization is performed to convert the scrubbed string into a series of tokens using single spaces as the separator. The order of the tokens remains the same as the input address. These tokens then are assigned a type based on their character content such as numeric (e.g., "3620"), alphabetic (e.g., "Vermont"), and alphanumeric (e.g., "RM444"). The final step, **syntactic analysis,** places the tokens into a parse tree based on a grammar. This **parse tree** is a data structure representing the decomposition of an input string into its component parts. The **grammar** is the organized set of rules that describe the language, in this case possible valid combinations of tokens that can legitimately make up an address. These are usually written in **Backus-Naur form** (BNF), a notation for describing grammars as combinations of valid components. See the next page for an example of an address described in BNF, in which a postal address is composed of two components: (1) the street-address-part, and (2) the locality-part. The street-address-part is composed of a house-number, a street-name-part, and an optional suite-number and suite-type, which would be preceded by a comma if they existed. The remaining components are composed in a similar fashion:

<postal-address> ::= <street-address-part> <locality-part>
<street-address-part> ::= <house-number> <street-name-part> {"," <suite-number> <suite-type>}
<street-name-part> ::= {<pre-directional>} <street-name> <street-type> {<post-directional>}
<locality-part> ::= <town-name> "," <state-code> <USPS-ZIP-Code> {"+" <ZIP-extension>}

The difficult part of context-based normalization is that the tokens described thus far have only been typed to the level of the characters they contain, not to the domain of address attributes (e.g., street name, post-directional). This level of domain-specific token typing can be achieved using lookup tables for common substitutions that map tokens to address components based on both character types and values. It is possible that a single token can be mapped to more than one address attribute. Thus, these tokens can be rearranged and placed in multiple orders that all satisfy the grammar. Therefore, constraints must be imposed on them to limit the erroneous assignments. Possible options include using an iterative method to enforce the original order of the tokens as a first try, then relaxing the constraint by allowing only tokens of specific types to be moved in a specific manner, etc. Also, the suppression of certain keywords can be employed such that their importance or relevance is minimized.

This represents the difficult part of performing context-based normalization—writing these relaxation rules properly, in the correct order. One must walk a fine line and carefully consider what should be done to which components of the address and in what order, otherwise the tokens in the input address might be moved from their original position and seemingly produce "valid" addresses that misrepresent the true address. Best practices related to context-based normalization are listed in Best Practices 16.

6.2.3 Probability-Based Normalization

Probability-based normalization makes use of statistical methods to identify the components of the input address. It derives mainly from the field of **machine learning**, a subfield of computer science dealing with algorithms that induce knowledge from data. In particular, it is an example of **record linkage,** the task of finding features in two or more datasets that essentially refer to the same feature. These methods excel at handling the difficult cases; those that require combinations of substitutions, reordering, and removal of extraneous data. Being so powerful, they typically are very difficult to implement, and usually are seen only in research scenarios.

These algorithms essentially treat the input address as unstructured text that needs to be semantically annotated with the appropriate attributes from the target domain (i.e., address attributes). The key to this approach is the development of an optimal **reference set**, which is the set of candidate features that may possibly match an input feature. This term should not to be confused with reference datasets containing the reference features, even though the reference set will most likely be built from them. The reference set defines the search space of possible matches that a feature-matching algorithm processes to determine an appropriate match. In most cases, the complexity of performing this search (i.e., processing time) grows linearly with the size of the reference set. In the worst case, the search space can be composed of the entire reference dataset, resulting in non-optimal searching. The intelligent use of **blocking schemes**, or strategies designed to narrow the set of candidate values (O'Reagan and Saalfeld 1987, Jaro 1989), can limit the size of the search space.

Best Practices 16 – Context-based normalization

Policy Decision	Best Practice
When should context-based normalization be used?	If the correct software can be acquired or developed, context-based normalization should be used.
Which characters should be considered valid and exempt from scrubbing?	All alpha-numeric characters should be considered valid. Forward slashes, dashes, and hyphens should be considered valid when they are between other valid characters (e.g., 1/2 or 123-B).
What action should be taken with scrubbed characters?	Non-valid (scrubbed) characters should be removed and not replaced with any character.
Which grammars should be used to define the components of a valid address?	Any grammar based on existing addressing standards can be used (e.g., OASIS xAL Standard [Organization for the Advancement of Structured Information Standards 2008] or the proposed URISA/FGDC address standard [United States Federal Geographic Data Committee 2008b]). The grammar chosen should be representative of the address data types the geocoding process is likely to see.
What level of token matching should be used for determining a match or non-match?	Only exact case-insensitive character-level matching should be considered a match.
How far from their original position should tokens within branches of a parse tree be allowed to move?	Tokens should be allowed to move no more than two positions of their original location.

After creating a reference set, matches and non-matches between input address elements and their normalized attribute counterparts can be determined. The input elements are scored against the reference set individually as well as collectively using several measures. These scores are combined into vectors and their likelihood as matches or non-matches is determined using such tools as support vector machines (SVMs), which have been trained on a representative dataset. For complete details of a practical example using this method, see Michelson and Knoblock (2005). Best practices related to probability-based normalization are listed in Best Practices 17.

Best Practices 17 – Probability-based normalization

Policy Decision	Best Practice
When should probability-based normalization be used?	If the output certainty of the resulting geocodes meets an acceptable threshold, probability-based normalization should be considered a valid option. Experiments should be run to determine what an appropriate threshold should be for a particular registry. These experiments should contrast the probability of getting a false positive versus the repercussions such an outcome will cause.
What level of composite score should be considered a valid match?	This will depend on the confidence that is required by the consumers of the geocoded data. At a minimum, a composite score of 95% or above should be considered a valid match.

6.3 ADDRESS STANDARDIZATION

More than one address standard may be required or in use at a registry for other purposes during or outside of the geocoding process. Therefore, after attribute identification and normalization, transformation between common address standards may be required. The difficult portion of this process is writing the **mapping functions**—the algorithms that translate between a normalized form and a target output standard. These functions transform attributes into the desired formats by applying such tasks as abbreviation substitution, reduction, or expansion, and attribute reordering, merging, or splitting. These transformations are encoded within the mapping functions for each attribute in the normalized form.

Mapping functions must be defined *a priori* for each of the potential standards that the geocoder may have to translate an input address into, and there are commonly many. To better understand this, consider that during feature matching, the input address must be in the same standard as that used for the reference dataset before a match can be attempted. Therefore, the address standard used by every reference dataset in a geocoder must be supported (i.e., a mapping function is required for each). With the mapping functions defined *a priori*, the standardization process can simply execute the appropriate transformation on the normalized input address and a properly standardized address ready for the reference data source will be produced.

In addition to these technical requirements for address standard support, registries must select an address standard for their staff to report and in which to record the data. Several existing and proposed address standards were listed previously in Table 6. NAACCR recommends that when choosing an address standard, registries abide by the data standards in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008) which reference United States Postal Service *Publication 28 - Postal Addressing Standards* (United States Postal Service 2008d) and the Canadian equivalent, *Postal Standards: Lettermail and Incentive Lettermail* (Canada Post Corporation 2008). Best practices related to address standardization are listed in Best Practices 18.

Best Practices 18 – Address standardization

Policy Decision	Best Practice
Which input data standards should be supported for standardization?	At a minimum, all postal address standards for all countries for which geocoding are to be performed should be supported.
Which address standard should be used for record representation?	Only a single address standard should be used for recording standardized addresses along with a patient/tumor record. This should be the standard defined in the NAACCR publication *Data Standards for Cancer Registries, Volume II*. All input data should be standardized according to these guidelines.
Which mapping functions should be used?	Mapping functions for all supported address standards should be created or obtained.

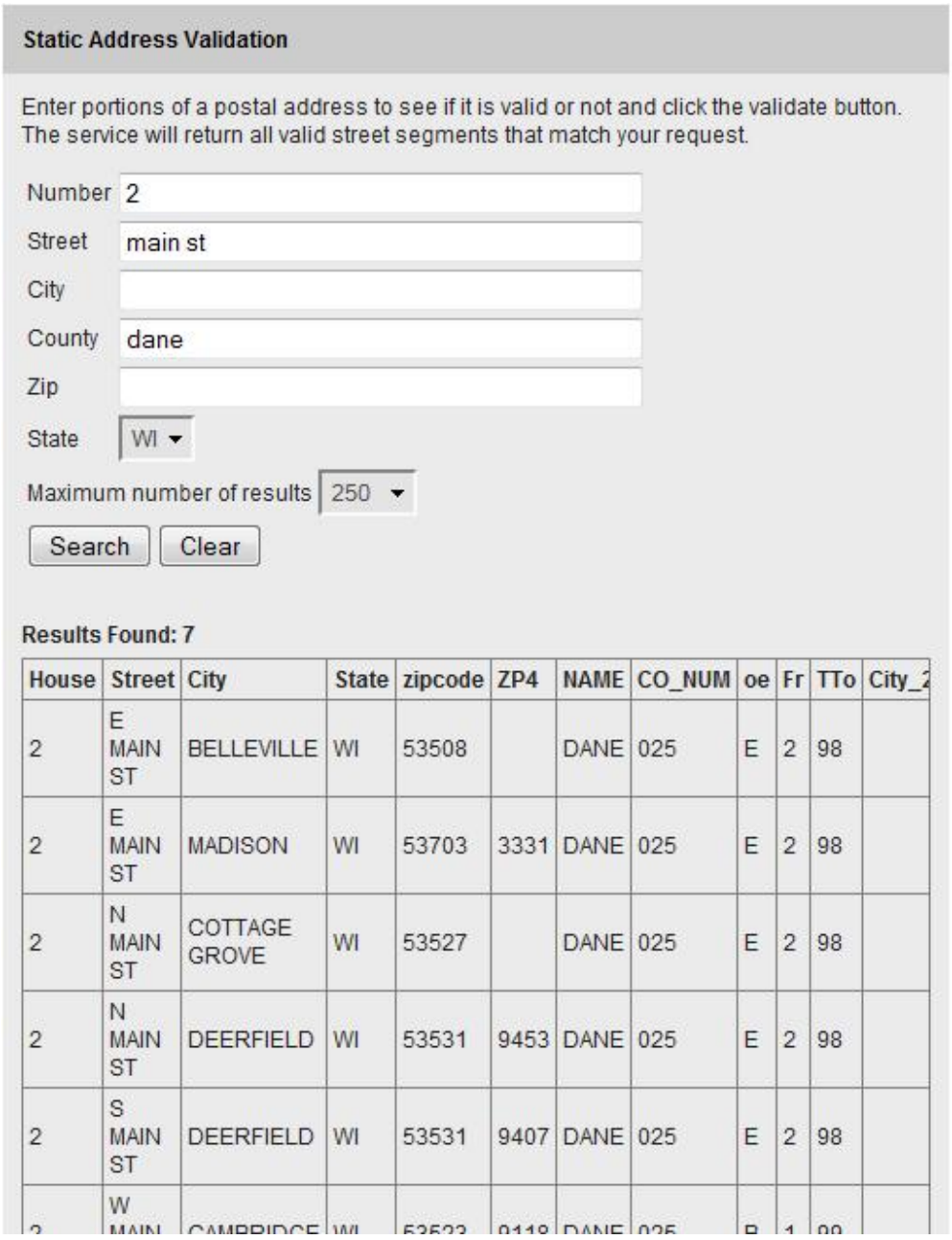
6.4 ADDRESS VALIDATION

Address validation is another important component of address cleaning that determines whether an input address represents a location that actually exists. This should always be attempted because it has a direct effect on the accuracy of the geocode produced for the input data in question, as well as other addresses that may be related to it (e.g., when performing linear-interpolation as discussed in Section 9.2). Performing address validation as close to the data entry as possible is the surest way to improve all aspects of the quality of the resulting geocode. Note that even though some addresses may validate, they still may not be geocodable due to problems or shortcomings with the reference dataset (note that the reverse also is true), which will be covered in more detail in Section 13.

In the ideal case, this validation will take place not at the central registry, but at the hospital. This practice currently is being implemented in several states (e.g., Kentucky, North Carolina, and Wisconsin), and is beginning to look like a feasible option, although regulations in some areas may prohibit it. The most commonly used source is the USPS ZIP+4 database (United States Postal Service 2008a), but others may be available for different areas and may provide additional help.

The simplest way to attempt address validation is to perform feature matching using a reference dataset containing **discrete features.** Discrete features are those in which a single feature represents only a single, real-world entity (e.g., a point feature) as opposed to a feature that represents a range or series of real-world entities (e.g., a line feature), as described in Section 7.2.3. A simple approach would be to use a USPS CASS-certified product to validate each of the addresses, but because of bulk mailers CASS systems are prohibited from validating segment-like reference data, and parcel or address points reference data must be used. In contrast, **continuous features** can correspond to multiple real-world objects, such as street segment, which has an address range that can correspond to several addresses. An example of this can be seen in the address validation application shown in Figure 7, which can be found on the USC GIS Research Laboratory Web site (https://webgis.usc.edu). This image shows the USC Static Address Validator (Goldberg 2008b), a Web-based address validation tool that uses the USPS ZIP+4 database to search for all valid addresses that match the address entered by the user. Once the user clicks search, either zero, one, or more than one potential address will be returned to indicate to the user that the information they entered did not match any addresses, matched an exact address, or matched multiple addresses. This information will allow the user to validate the address in question by determining and cor-

recting any attributes that are wrong or incomplete that could potentially lead to geocoding errors had the non-validated address been used directly.



Static Address Validation

Enter portions of a postal address to see if it is valid or not and click the validate button.
The service will return all valid street segments that match your request.

Number 2
Street main st
City
County dane
Zip
State WI
Maximum number of results 250
Search Clear

Results Found: 7

House	Street	City	State	zipcode	ZP4	NAME	CO_NUM	oe	Fr	TTo	City_
2	E MAIN ST	BELLEVILLE	WI	53508		DANE	025	E	2	98	
2	E MAIN ST	MADISON	WI	53703	3331	DANE	025	E	2	98	
2	N MAIN ST	COTTAGE GROVE	WI	53527		DANE	025	E	2	98	
2	N MAIN ST	DEERFIELD	WI	53531	9453	DANE	025	E	2	98	
2	S MAIN ST	DEERFIELD	WI	53531	9407	DANE	025	E	2	98	
2	W MAIN	CAMBRIDGE	WI	53523	9118	DANE	025	R	1	99	

Figure 7 – Example address validation interface (https://webgis.usc.edu)

If feature matching applied to a reference dataset of discrete features succeeds, the matched feature returned will be in one of two categories: a true or false positive. A **true positive** is the case when an input address is returned as being true, and is in fact true (e.g., it actually exists in the real world). A **false positive** is the case when an input address is returned as being true, and is in fact false (e.g., it does not actually exist in the real world). If feature matching fails (even after attempting attribute relaxation as described in Section 8.3.1) the input address will fall again into one of two categories, a true or false negative. A

true negative is the case when an input address is returned as being false, and is in fact false. A **false negative** is the case when an input address is returned as being false, and is in fact true (e.g., it does actually exist in the real world).

Both false positives and negatives also can occur due to temporal inaccuracy of reference datasets. False positives occur when the input address is actually invalid but appears in the reference dataset (e.g., it has not yet been removed). False negatives occur when the input address exists, but is not present in the reference dataset (e.g., it has not yet been added). To address these concerns, the level of confidence for the temporal accuracy of a reference dataset needs to be determined and utilized. To assess this level of confidence, a registry will need to consider the frequency of reference dataset update as well as address lifecycle management in the region and characteristics of the region (e.g., how old is the reference set, how often is it updated, and how frequently do addresses change in the region). More details on the roots of temporal accuracy in reference datasets are described in Section 13.3.

Common reference data sources that can be used for address verification are listed in Table 16. Although parcel data have proven very useful for address data, it should be noted that in most counties, assessors are under no mandate to include the **situs address** of a parcel (the actual physical address associated with the parcel) in their databases. In these cases, the mailing address of the owner may be all that is available, but may or may not be the actual address of the actual parcel. As such, E-911 address points may be an alternative and better option for performing address validation. Best practices related to address validation are listed in Best Practices 19. Recent work by Zandbergen (2008) provides further discussion on the affect discrete (address point- or parcel-based) versus continuous (address range-based street segments) reference datasets has on achievable match rates.

Table 16 – Common address verification data sources

USPS ZIP+4 (United States Postal Service 2008a)
U.S. Census Bureau Census Tracts
County or municipal assessor parcels

There appears to be general consensus among researchers and registries that improving address data quality at the point of collection should be a task that is investigated, with its eventual implementation into existing data entry systems a priority. It is as-of-yet unclear how utilizing address validation tools like the USC Web-based address validator shown in this section may or may not slow down the data entry process because there have been no published reports detailing registry and/or staff experiences, time estimates, or overall cost increases. However, preliminary results presented at the 2008 NAACCR Annual Meeting (Durbin et al. 2008) on the experience of incorporating a similar system into a data entry system used by the State of Kentucky seem to indicate that the time increases are manageable, with proper user interface design having a large impact. More research is needed on this issue to determine the cost and benefits that can be obtained using these types of systems and the overall impact that they will have on resulting geocode quality.

Best Practices 19 – Address validation

Policy Decision	Best Practice
When should address validation be used?	If a trusted, complete address dataset is available, it should be used for address validation during both address standardization and feature matching and interpolation.
Which data sources should be used for address validation?	The temporal footprint of the address validation source should cover the period for which the address in question was supposed to have existed in the dataset. If an assessor parcel database is available, this should be used as an address validation reference dataset.
What should be done with invalid addresses?	If an address is found to be invalid during address standardization, it should be corrected. If an invalid address is not correctable, it should be associated with the closest valid address.
What metadata should be maintained?	If an address is corrected or assigned to the closest valid address, the action taken should be recorded in the metadata, and the original address should be kept as well.

7. REFERENCE DATASETS

This section identifies and describes the different types of reference datasets and the relationships between them.

7.1 REFERENCE DATASET TYPES

Vector-based data, such as the U.S. Census Bureau's TIGER/Line files (United States Census Bureau 2008c), are the most frequently encountered reference datasets because their per-feature representations allow for easy feature-by-feature manipulation. The pixel-based format of raster-based data, such as digital orthophotos, can be harder to work with and generally make them less applicable to geocoding. However, some emerging geocoding processes do employ raster-based data for several specific tasks including feature extraction and correction (as discussed later).

7.2 TYPES OF REFERENCE DATASETS

The following sections offer more detail on the three types of vector-based reference datasets—linear-, areal unit-, and point-based—frequently used in geocoding processes, organized by their degree of common usage in the geocoding process. The descriptions of each will, for the most part, be generalizations applicable to the whole class of reference data. Also, it should be noted that the true accuracy of a data source can only be determined with the use of a GPS device, or in some cases imagery, and these discussions again are generalizations about classes of data sources. An excellent discussion of the benefits and drawbacks of geocoding algorithms based on each type of reference dataset is available in Zandbergen (2008).

7.2.1 Linear-Based Reference Datasets

A **linear-based (line-based) reference dataset** is composed of linear-based data, which can either be simple-line or polyline vectors. The type of line vector contained typically can be used as a first-order estimate of the descriptive quality of the reference data source. Reference datasets containing only simple straight-line vectors usually will be less accurate than reference datasets containing polyline vectors for the same area (e.g., when considering the shortest possible distance between two endpoints). Curves typically are represented in these datasets by breaking single straight-line vectors into multiple segments (i.e., polylines). This scenario is depicted in Figure 8, which shows a polyline more accurately describing the shape of a street segment than a straight line.

a) Low-resolution line

b) High-resolution line

c) Comparison of two representations

Figure 8 – Vector reference data of different resolutions (Google, Inc. 2008b)

Line-based datasets underpin typical conceptions of the geocoding process and are by far the most cited in the geocoding literature. Most are usually representations of street **networks (graphs),** which are an example of a topologically connected set of nodes and edges. The **nodes (vertices)** are the endpoints of the line segments in the graph and the **edges**

(**arcs**) are lines connecting the endpoints. The term "network" refers to the **topological connectivity** resulting from reference features sharing common endpoints, such that it is possible to traverse through the network from feature to feature. Most literature commonly defines a graph as $G = (V, E)$, indicating that the graph G is composed of the set of vertices V and the set of edges E. The inherent topological connectedness of these graphs enables searching. Dijkstra's (1959) shortest path algorithm is frequently used for route planning, and several well known examples of street networks are provided in Table 17. Further details of street networks, alternative distance estimations, and their application to accessibility within the realm of cancer prevention and control can be found in Armstrong et al. (2008) and the references within.

Table 17 – Common linear-based reference datasets

Name	Description	Coverage
U.S. Census Bureau's TIGER/Line files (United States Census Bureau 2008c)	Street centerlines	U.S.
NAVTEQ Streets (NAVTEQ 2008)	Street centerlines	Worldwide
Tele Atlas Dynamap, MultiNet (Tele Atlas 2008a, c)	Street centerlines	Worldwide

The first dataset listed, the U.S. Census Bureau's TIGER/Line files, is the most commonly used reference dataset in geocoding. The next two are competing products that are commercial derivatives of the TIGER/Line files for the United States. All three products essentially provide the same type of data, with the commercial versions containing improvements over the TIGER/Lines files in terms of reference feature spatial accuracy and the inclusion of more aspatial attributes. The accuracy differences between products can be stunning, as can the differences in their cost.

Commercial companies employ individuals to drive GPS-enabled trucks to obtain GPS-level accuracy for their polyline street vector representations. They also often include areal unit-based geographic features (polygons) (e.g., hospitals, parks, water bodies), along with data that they have purchased or collected themselves. These data collection tasks are not inexpensive, and these data therefore are usually very expensive, typically costing on the order of tens of thousands of dollars. However, part of the purchase price usually includes yearly or quarterly updates to the entire reference dataset, resulting in very temporally accurate reference data.

In contrast, new releases of the TIGER/Line files have historically corresponded to the decennial Census, resulting in temporal accuracy far behind their commercial counterparts. Also, even though support for polyline representations is built into the TIGER/Line file data format, most features contained are in fact simple-lines, with very few areal unit-based features included. However, while the commercial versions are very expensive, TIGER/Line files are free (or, to avoid time-consuming downloading, are available for reasonable fees on DVD), making them an attractive option. Also, beginning in 2002, updates to the TIGER/Line files have been released once and now twice per year, resulting in continually improving spatial and temporal accuracy. In some areas, states and municipalities have created much higher quality line files; these eventually will be or already have been incorporated into the TIGER/Line files. Beginning in 2007, the U.S. Census Bureau has released

MAF-TIGER files to replace annual TIGER/Line files; these merge the U.S. Census Bureau's Master Address File (MAF) with TIGER databases to create a relational database management system (RDBMS) (United States Census Bureau 2008b).

Recent studies have begun to show that in some areas, the TIGER/Line files are essentially as accurate as commercial files (Ward et al. 2005), and are becoming more so over time. Some of this change is due to the U.S. Census Bureau's MAF-TIGER/Line file integration and adoption of the new American Community Survey (ACS) system (United States Census Bureau 2008a), which itself includes a large effort focused on improving the TIGER/Line files, others are due to pressure from the FGDC. These improvements are enabling greater public participation and allowing local-scale knowledge with higher accuracy of street features and associated attributes (e.g., address ranges), to inform and improve the national-scale products.

All of the products listed in Table 17 share the attributes listed in Table 18. These represent the attributes typically required for feature matching using linear-based reference datasets. Note that most of these attributes correspond directly to the components of city-style postal address-based input data.

Table 18 – Common postal address linear-based reference dataset attributes

Attribute	Description
Left side street start address number	Beginning of the address range for left side of the street segment
Right side street start address number	Beginning of the address range for right side of the street segment
Left side street end address number	End of the address range for left side of the street segment
Right side street end address number	End of the address range for right side of the street segment
Street prefix directional	Street directional indicator
Street suffix directional	Street directional indicator
Street name	Name of street
Street type	Type of street
Right side ZCTA	ZCTA for addresses on right side of street
Left side ZCTA	ZCTA for addresses on left side of street
Right side municipality code	A code representing the municipality for the right side
Left side municipality code	A code representing the municipality for the left side
Right side county code	A code representing the county for the right side
Left side county code	A code representing the county for the left side
Feature class code	A code representing the class of the feature

In street networks, it is common for each side of the reference feature to be treated separately. Each can be associated with different address ranges and ZCTAs, meaning that one side of the street can be in one ZCTA while the other is in another ZCTA (i.e., the street forms the boundary between two ZCTAs). The address ranges on each side do not necessary need to be related, although they most commonly are. Attributes of lower geographic resolutions than the ZCTA (city name, etc.) usually are represented in the form of a code

(e.g., a Federal Information Processing Standard [FIPS] code [National Institute of Standards and Technology 2008]), and also applied to each side of the street independently.

All features typically include an attribute identifying the class of feature it is, e.g., a major highway without a separator, major highway with a separator, minor road, tunnel, freeway onramp. These classifications serve many functions including allowing for different classes of roads to be included or excluded during the feature-matching process, and enabling first-order estimates of road widths to be assumed based on the class of road, typical number of lanes in that class, and typical lane width. In the TIGER/Line files released before March 2008, these are represented by a Feature Classification Code (FCC), which has subsequently been changed to the MAF/TIGER Feature Class Code (MTFCC) in the upgrade to MAF-TIGER/Line files (United States Census Bureau 2008c). In the more advanced commercial versions, additional information such as one-way roads, toll roads, etc., are indicated by binary true/false values for each possible attribute.

7.2.2 Polygon-Based Reference datasets

A **polygon-based reference dataset** is composed of polygon-based data. These datasets are interesting because they can represent both the most accurate and inaccurate forms of reference data. When the dataset represents true building footprints, they can be the most accurate data source one could hope for when they are based from surveys; they have less or unknown accuracy when derived from photographs. Likewise, when the polygons represent cities or counties, the dataset quickly becomes less appealing. Most polygon-based datasets only contain single-polygon representations, although some include polygons with multiple rings. Three-dimensional reference datasets such as building models are founded on these multi-polygon representations.

Polygon reference features often are difficult and expensive to create initially. But when available, they typically are on the higher side of the accuracy spectrum. Table 19 lists some examples of polygon-based vector reference datasets, along with estimates of their coverages and costs.

Table 19 – Common polygon-based reference datasets

Source	Description	Coverage	Cost
Tele Atlas (2008c), NAVTEQ (2008)	Building footprints, parcel footprints	Worldwide, but sparse	Expensive
County or municipal Assessors	Building footprints, parcel footprints	U.S., but sparse	Relatively inexpensive but varies
U.S. Census Bureau	Census Block Groups, Census Tracts, ZCTA, MCD, MSA, Counties, States	U.S.	Free

The highest quality dataset one can usually expect to encounter are building footprints. These data typically enable the geocoding process to return a result with an extremely high degree of accuracy, with automated geocoding results of higher quality generally only obtainable through the use of 3-D models such as that shown in Figure 9. Three-dimensional models also are built from polygon representations but are even less commonly encountered.

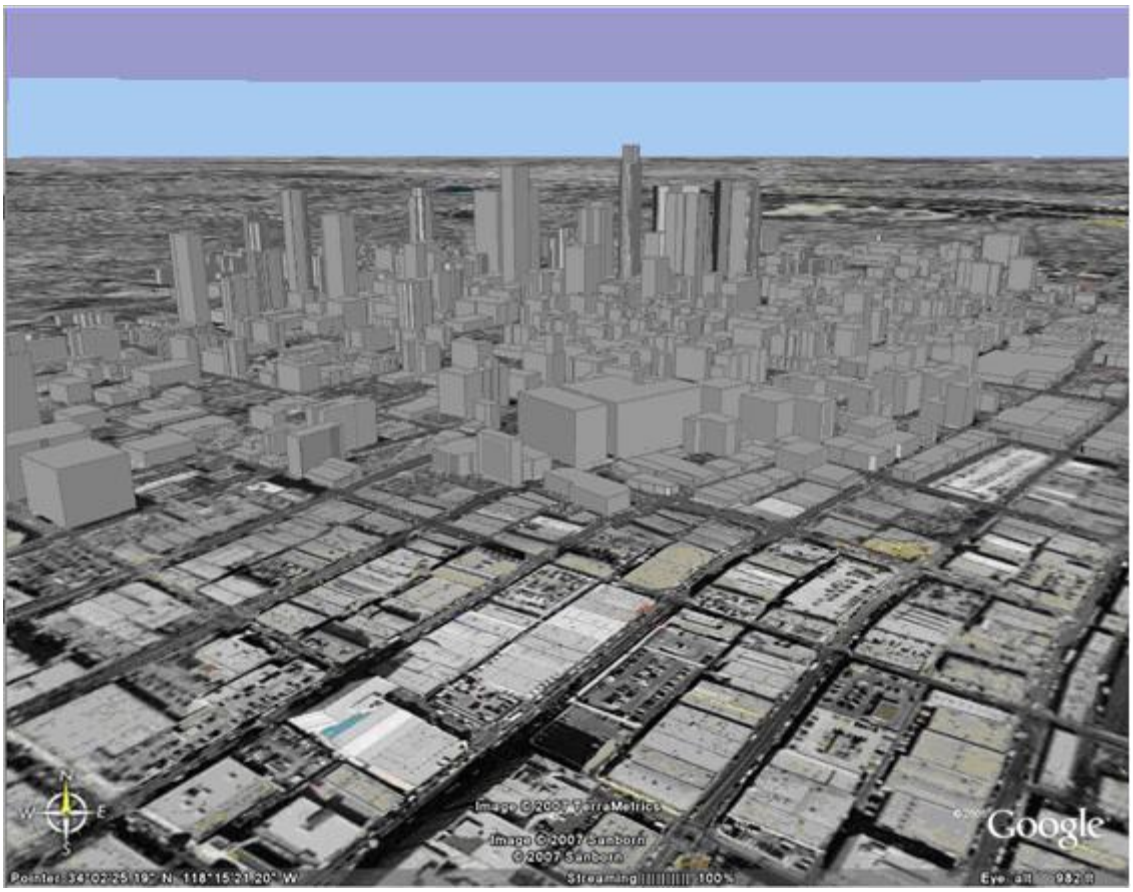


Figure 9 – Example 3D building models (Google, Inc. 2008a)

Although both building footprints and 3-D polygons are becoming more commonplace in commercial mapping applications (e.g., Microsoft Virtual Earth and Google Maps having both for portions of hundreds of cities worldwide), these datasets often are difficult or costly to obtain, typically requiring a substantial monetary investment. They are most often found for famous or public buildings in larger cities or for buildings on campuses where the owning organization has commissioned their creation. It is quite rare that building footprints will be available for every building in an entire city, especially for residential structures, but more and more are becoming available all the time.

A person attempting to gather reference data can become frustrated because although maps depicting building footprints are widely available, digital copies of the underlying datasets can be difficult, if not impossible, to obtain. This happens frequently with paper maps created for insurance purposes (e.g., Sanborn Maps), and static digital images such the USC Campus Map (University of Southern California 2008) shown in Figure 10. In many cases, it is obvious that digital geographic polygon data serve as the basis for online interactive mapping applications as in the UCLA Campus Map shown in Figure 11 (University of California, Los Angeles 2008), but often these data are not made available to the general public for use as a reference dataset within a geocoding process.

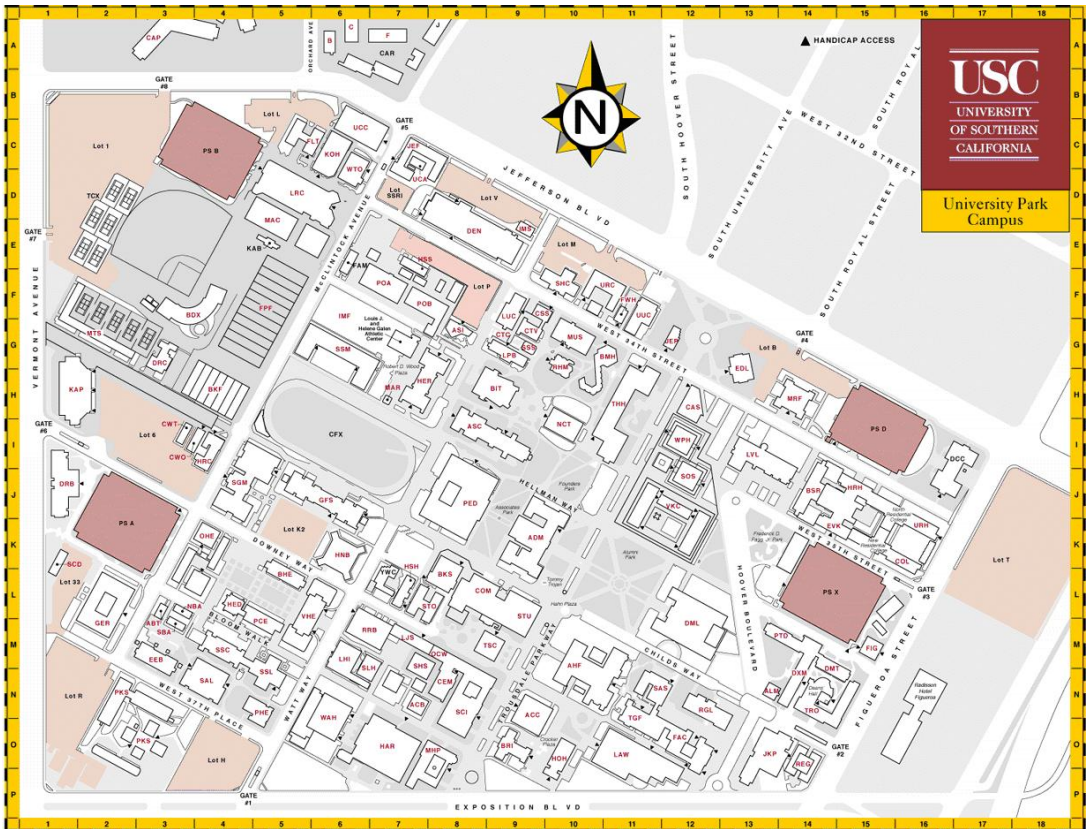


Figure 10 – Example building footprints in raster format (University of Southern California 2008)

In contrast to building footprints, parcel boundaries are available far more frequently. These are descriptions of property boundaries, usually produced by local governments for taxation purposes. In most cases they are legally binding and therefore often are created with survey-quality accuracy, as shown in Figure 12. However, it should be noted that only a percentage of the total actually are produced from surveying, with others being either derived from imagery or legacy data. Therefore, "legally-binding" does not equate to "highly accurate" in every instance.

These data are quickly becoming available for most regions of the United States, with some states even mandating their creation and dissemination to the general public at low cost (e.g., California [(Lockyer 2005]). Also, the U.S. FGDC has an initiative underway to create a national parcel file for the entire country within a few years (Stage and von Meyer 2005). As an example of their ubiquitous existence, the online site Zillow (Zillow.com 2008) appears to have obtained parcel data for most of the urban areas in the United States.

The cost to obtain parcels usually is set by the locality and can vary dramatically from free (e.g., Sonoma County, CA [County of Sonoma 2008]) to very expensive (e.g., $125,000 for the Grand Rapids, MI Metropolitan Area [Grand Valley Metropolitan Council 2008]). Also, because they are created for tax purposes, land and buildings that are not subject to local taxation (e.g., public housing, state-owned residential buildings, or residences on military bases or college campuses) may be omitted. The attributes which these parcel-based reference datasets have in common are listed in Table 20.

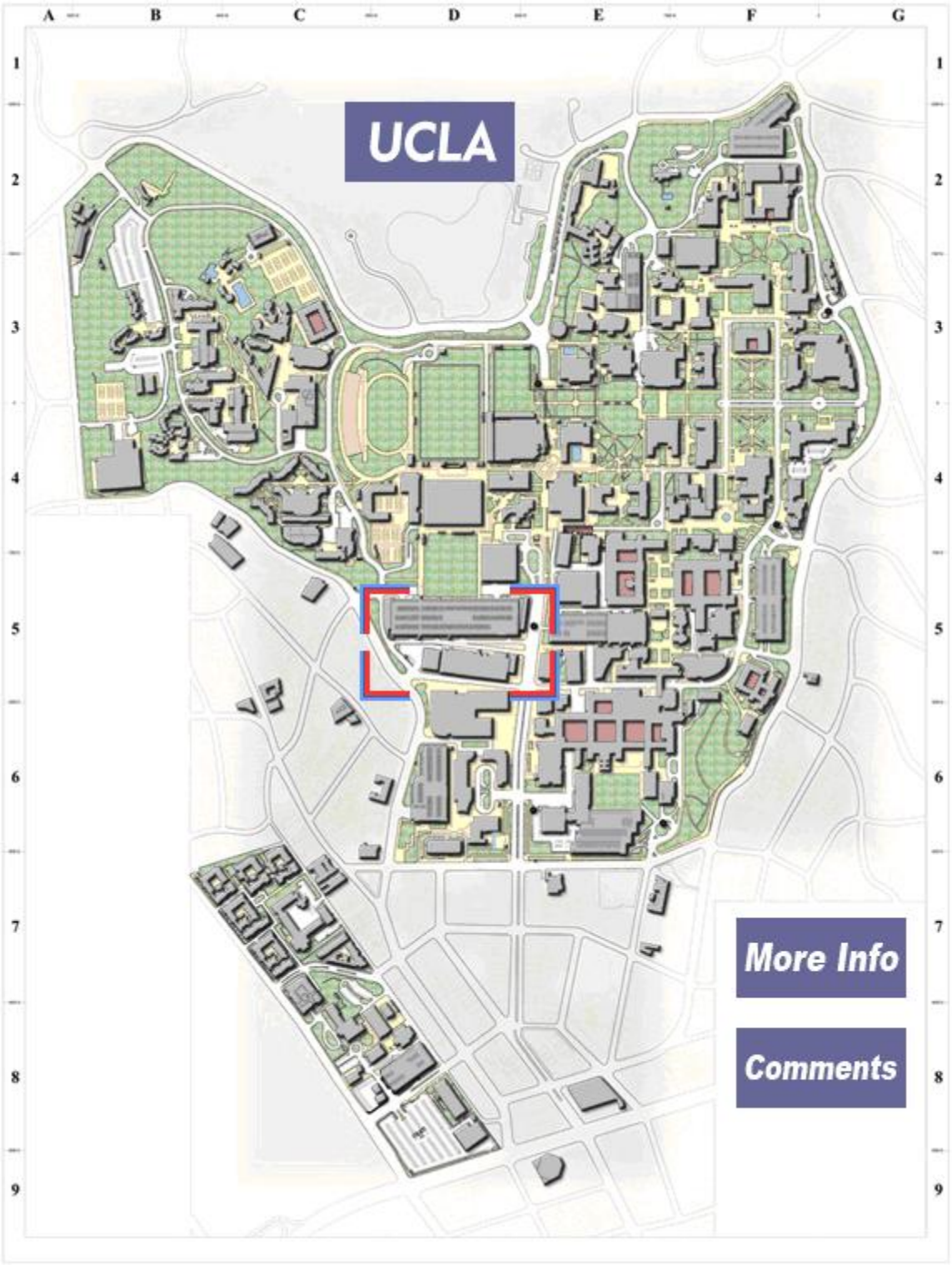


Figure 11 – Example building footprints in digital format (University of California, Los Angeles 2008)

Figure 12 – Example parcel boundaries with centroids

Table 20 – Common polygon-based reference dataset attributes

Attribute	Description
Name	The name of feature used for search
Polygon Coordinates	Set of polylines in some coordinate system
Index code/identifier	Code to identify the polygon within the reference data system

Similar to point-based reference features, parcel-based reference features are **discrete** (i.e., they typically describe a single real-world geographic feature). Thus, a feature-matching algorithm usually will either find an exact match or none at all. Unlike point features, these parcel-based features are complex geographic types, so spatial operations can be performed on them to create new data such as a centroid (i.e., interpolation). Also, the address associated with a parcel may be the mailing address of the owner, not the **situs address,** or address associated with the physical location of the parcel. The benefits and drawbacks of various centroid calculations are detailed in Section 9.3.

Again, similar to point-based reference datasets, lower-resolution versions of polygon-based reference datasets are readily obtainable. For example, in addition to their centroids, the U.S. Census Bureau also freely offers polygon representations of MCDs, counties, and states. The low resolution of these polygon features may prohibit their direct use as spatial output, but they do have valuable uses. In particular, they are extremely valuable as the spatial boundaries of spatial queries when a feature-matching algorithm is looking for a line-based reference feature within another reference dataset. They can serve to limit (clip) the spatial domain that must be searched, thus speeding up the result, and should align well with

U.S. Census Bureau Census Tract (CT), Census Block Group (CBG), etc. files from the same release.

7.2.3 Point-Based Reference Datasets

A **point-based reference dataset** is composed of point-based data. These are the least commonly encountered partly because of their usability, and partly because of the wide ranges in cost and accuracy. The usability of geographic points (in terms of interpolation potential) is almost non-existent because a point represents the lowest level of geographic complexity. They contain no attributes that can be used for the interpolation of other objects, in contrast to datasets composed of more complex objects (e.g., lines) that do have attributes suitable for deriving new geographic objects (e.g., deriving a point from a line using the length attribute). Their usability is further reduced because most are composed of discrete features; however, they are sometimes used in research studies.

Although this is beneficial for improving the precision of the geocoder (i.e., it will only return values for input addresses that actually exist), it will lower the match rate achieved (more details on match rate metrics are described in Section 14.2). This phenomenon is in contrast to linear-based reference datasets that can handle values within ranges for a feature to be matched. This scenario produces the exact opposite effect of the point-based reference set—the match rate rises, but precision falls. See Zandbergen (2008) for a detailed analysis of this phenomenon.

The cost of production and accuracy of point-based reference datasets can range from extremely high costs and high accuracy when using GPS devices, such as the address points available for some parts of North Carolina, to extremely low-cost and variable accuracy when building a cache of previously geocoded data (as described in Section 13.4). Several examples of well-known, national-scale reference datasets are listed in Table 21, and Abe and Stinchcomb (2008, pp. 123) note that commercial vendors are beginning to produce and market point-level address data. The attributes listed in Table 22 are common to all products listed in Table 21. These form the minimum set of attributes required for a feature-matching algorithm to successfully match a reference in a point-based reference dataset.

Table 21 – Point-based reference datasets

Supplier	Product	Description	Coverage
Government	E-911 Address Points	Emergency management points for addresses	Portions of U.S.
Government	Postal Codes	Postal Code centroids	U.S./Canada
Government	Census MCD	Minor Civil Division centroids	U.S.
Government	Geographic Names Information System (United States Board on Geographic Names 2008)	Gazetteer of geographic features	U.S.
Government	GeoNames (United States National Geospatial-Intelligence Agency 2008)	Gazetteer of geographic features	World, excepting U.S.
Academia	Alexandria Digital Library (2008)	Gazetteer of geographic features	World

Table 22 – Minimum set of point-based reference dataset attributes

Attribute	Description
Name	The name of the feature used for the search.
Point coordinates	A pair of values for the point in some coordinate system.

The United Kingdom and Australia currently have the highest quality point-based reference datasets available, containing geocodes for every postal address in the country. Their creation processes are well documented throughout the geocoding literature (Higgs and Martin 1995a, Churches et al., 2002, Paull, 2003), as are numerous studies performed to validate and quantify their accuracy (e.g., Gatrell 1989). In contrast, neither the United States nor Canada can currently claim the existence of a national-scale reference dataset containing accurate geocodes for all addresses in the country. The national-scale datasets that are available instead contain lower-resolution geographic features. In the United States, these datasets are mostly available from the U.S. Census Bureau (e.g., ZCTAs, centroids, and points representing named places such as MCDs). These two datasets in particular are distributed in conjunction with the most common linear-based reference data source used, the U.S. Census Bureau TIGER/Line files (United States Census Bureau 2008c). USPS ZIP Codes are different than U.S. Census Bureau ZCTAs and their (approximate) centroids are available from commercial vendors (covered in more detail in Section 5.1.4). Higher resolution point data have been created by individual localities across the United States, but these can be difficult to find in some locations unless one is active or has connections in the locality. Best practices relating to reference dataset types are listed in Best Practices 20.

7.3 REFERENCE DATASET RELATIONSHIPS

The implicit and explicit relationships that exist between different reference dataset types are similar to the components of postal address input data. These can be both structured spatially hierarchical relationships and lineage-based relationships. An example of the first is the hierarchical relationships between polygon-based features available at different geographic resolutions of Census delineations in the TIGER/Line files. Census blocks are at the highest resolution, followed by CBG, CT, ZCTA, county subdivisions, counties, and/or other state subdivisions, etc. The spatially hierarchical relationships between these data types are important because data at lower resolutions represent an aggregation of the features at the higher level. When choosing a reference feature for interpolation, one can safely change from selecting a higher resolution representation to a lower one (e.g., a block to a block group) without fear of introducing erroneous data (e.g., the first digit of the block is the block group code). The inverse is not true because lower-resolution data are composed of multiple higher-resolution features (e.g., a block group contains multiple blocks). When attempting to increase the resolution of the feature type matched to, there will be a level of ambiguity introduced as to which is the correct higher resolution feature that should be selected.

Best Practices 20 – Reference dataset types

Policy Decision	Best Practice
What reference datasets formats can and should be used?	Any reference dataset format should be supported by a geocoding process, both vector- and raster-based. At a minimum, vector-based must be supported.
What vector-based reference dataset types can and should be used?	Any vector-based reference dataset type should be supported by a geocoding process (e.g., point-, linear-, and polygon-based). At a minimum, linear-based must be supported.
Which data source should be obtained?	A registry should obtain the most accurate reference dataset they can obtain given their budgetary and technical constraints. Cost may be the influencing factor as to which data source to use. There may be per-product limitations, so all choices and associated initiations should be fully investigated before acquisition.
When should a new data source be obtained?	A registry should keep their reference dataset up-to-date as best they can within their means. The update frequency will depend on budgetary constraints and the frequency with which vendors provide updates.
Should old data sources be discarded?	A registry should retain historical versions of all their reference datasets.
Where can reference data be obtained?	Local government agencies and the FGDC should be contacted to determine the types, amounts, and usability of reference datasets available. Commercial firms (e.g., Tele Atlas [2008c] and NAVTEQ [2008]) also can be contacted if needs cannot be met by public domain data.
How should reference data sources be kept?	Registries should maintain lists of reference datasets applicable to their area across all resolutions (e.g., TIGER/Lines [United States Census Bureau 2008c] – national, county government roads – regional, parcel databases – local).

Examples of derivational lineage-based relationships include the creation of NAVTEQ (2008) and Tele Atlas (2008c) as enhanced derivatives of the TIGER/Line files and **geocode caching,** in which the output of a feature interpolation method is used to create a point-based reference dataset (as described in Section 13.4). In either of these cases, the initial accuracy of the original reference dataset is a main determinant of the accuracy of later generations. This effect is less evident in the case of TIGER/Line file derivatives because of the continual updating, but is completely apparent in reference datasets created from cached results. Best practices related to these spatial and derivational reference dataset relationships are listed in Best Practices 21.

Best Practices 21 – Reference dataset relationships

Policy Decision	Best Practice
Should primary or derivative reference datasets be used (e.g., TIGER/Lines or NAVTEQ)?	Primary source reference datasets should be preferred to secondary derivatives unless significant improvements have been made and are fully documented and can be proven.
Should lower-resolution aggregate reference data be used over original individual features (e.g., block groups instead of blocks)?	Moving to lower resolutions (e.g., from block to block group) should only be done if feature matching is not possible at the higher resolution due to uncertainty or ambiguity.

In addition to the inter-reference dataset relationships among different datasets, intra-reference dataset relationships are at play between features within a single dataset. This can be seen by considering various **holistic metrics** used to describe datasets, which are characteristics describing values over an entire dataset as a whole. **Atomic metrics,** in contrast, describe characteristics of individual features in a dataset. For example, datasets commonly purport the holistic metric "average horizontal spatial accuracy" as a single value (e.g., 7 m in the case of the TIGER/Line files). However, it is impossible to measure the horizontal spatial accuracy of every feature in the entire set, so where did this number come from? These holistic measures are calculated by choosing a representative sample and averaging their values to derive a metric. For this reason, holistic metrics usually are expressed along with a **confidence interval** (CI), which is a measurement of the percentage of data values that are within a given range of values. This is the common and recommended practice for describing the quality of spatial data, according to the FGDC data standards.

For example, stating that the data are accurate to 7 m with a CI of 95 percent means that for a particular subset of individual features that were tested out of all the possible features, roughly 95 percent fall within 7 m. The creator of the dataset usually does not (and usually cannot) guarantee that each and every feature within the dataset has this same value as its accuracy (which would make it an atomic metric). Although a data consumer generally can trust CIs associated with holistic metrics, they must remain aware of the potential for individual features to vary, sometimes being much different than those reported for the entire set. This phenomenon is commonly most pronounced in the differences in values for feature metrics seen in different geographic regions covered by large datasets (e.g., feature accuracy in rural versus urban areas).

Another aspect related to atomic and holistic feature completeness and accuracy is **geographical bias**. In one sense, this describes the observation that the accuracy of geographic features may be a function of the area in which they are located. Researchers are beginning to realize that geocodes produced with similar reported qualities may not actually have the same accuracy values when they are produced for different areas. The accuracy of the geocoding process as a whole has been shown to be highly susceptible to specific properties of the reference features, such as the length of the street segments (Ratcliffe 2001, Cayo and Talbot 2003, Bakshi et al. 2004) that are correlated with characteristics such as the rural or urban character of a region (e.g., smaller/larger postal code/parcel areas and the likelihood of USPS PO box addresses areas [Skelly et al. 2002, Bonner et al. 2003, McElroy et al. 2003, Ward et al. 2005]). Likewise, the preponderance of change associated with the reference features and input data in rapidly expanding areas will undoubtedly affect the geocoding process in different ways in different areas depending on the level of temporal dynamism of the local

built environment. This notion is partially captured by the newly coined term **cartographic confounding** (Oliver et al., 2005). Best practices relating to reference dataset characteristics are listed in Best Practices 22.

Best Practices 22 – Reference dataset characteristics

Policy Decision	Best Practice
Should holistic or atomic metrics be used to describe the accuracy of a reference dataset?	If the geographic variability of a region is low or the size of the region covered is small (e.g., city scale), the holistic metrics for the reference dataset should be used. If the geographic variability of a region is high or the size of the region covered is large (e.g., national scale), the accuracy of individual reference features within the area of the input data should be considered over the holistic measures.
Should geographic bias be considered a problem?	If the geographic variability of a region is high or the size of the region covered is large (e.g., national scale), geographic bias should be considered as a possible problem.

8. FEATURE MATCHING

This section investigates the components of a feature-matching algorithm, detailing several specific implementations.

8.1 THE ALGORITHM

Many implementations of feature-matching algorithms are possible and available, each with their own benefits and drawbacks. At the highest and most general level, the feature-matching algorithm performs a single simple role. It selects the correct reference feature in the reference dataset that represents the input datum. The chosen feature then is used in the feature interpolation algorithm to produce the spatial output. This generalized concept is depicted in Figure 13. The matching algorithms presented in this section are **non-interactive matching algorithms** (i.e., they are automated and the user is not directly involved). In contrast, **interactive matching algorithms** involve the user in making choices when the algorithm fails to produce an exact match by either having the user correct/refine the input data or make a subjective, informed decision between two equally likely options.

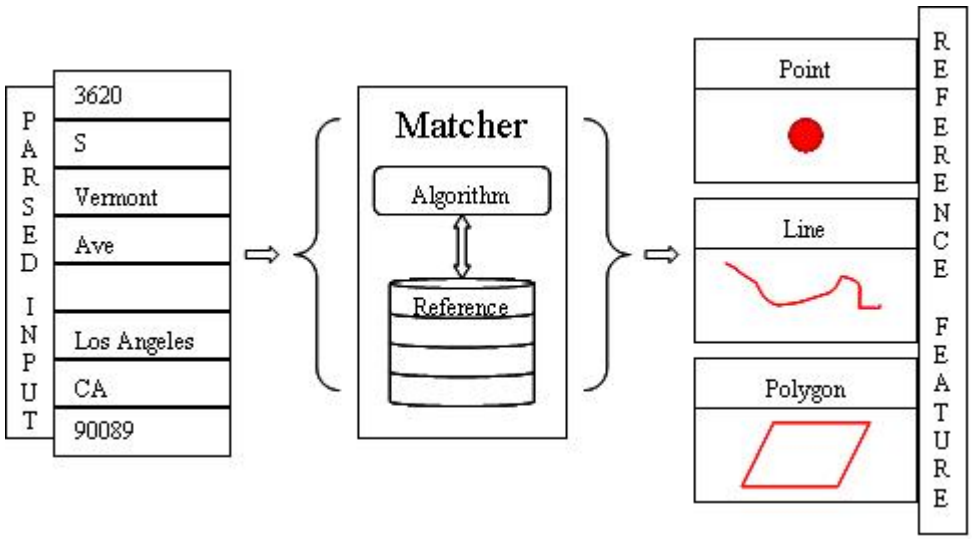


Figure 13 – Generalized feature-matching algorithm

8.1.1 SQL Basis

The form taken by feature-matching algorithms is dictated by the storage mechanism of the reference dataset. Therefore, because most reference datasets are stored as traditional relational database structures, most matching algorithms usually operate by producing and issuing queries defined using the Structured Query Language (SQL). These SQL queries are defined in the following format:

```
SELECT   <selection attributes>
FROM     <data source>
WHERE    <attribute constraints>
```

The **selection attributes** are the attributes of the reference feature that should be returned from the reference dataset in response to the query. These typically include the identifiable attributes of the feature such as postal address components, the spatial geometry of the reference feature such as an actual polyline, and any other desired descriptive aspatial qualities such as road width or resolution. The **data sources** are the relational table (or tables) within the reference dataset that should be searched. For performance reasons (e.g., scalability), the reference dataset may be separated into multiple tables (e.g., one for each state) within a national-scale database. The **attribute constraints** form the real power of the query, and consist of zero, one, or more predicates. A **predicate** is an attribute/value pair defining what the value of an attribute *must* be for a feature to be selected. Multiple predicates can be linked together with "AND" and "OR" statements to form conjunctions and disjunctions. Nesting of predicates also is supported through the use of parentheses.

To satisfy a query, the relational database engine used to store the reference dataset will ensure that Boolean Logic is employed to evaluate the attribute constraints against each feature in the reference dataset, returning only those that evaluate to true statements. The following example would enforce the condition that only reference features whose 'name' attribute was equal to 'Vermont' and had a 'type' attribute equal to either 'AVE' or 'ST' would be returned.

```
SELECT   <attributes>
FROM     <data source>
WHERE    name='Vermont' and (type='AVE' or type='ST')
```

Case sensitivity relates to whether or not a database differentiates between the case of alphabetic characters (i.e., upper-case or lower-case) when evaluating a query against reference features, and if enforced can lead to many false negatives. This is platform dependent and may be a user-settable parameter. Best practices related to SQL-type feature matching are listed in Best Practices 23.

Best Practices 23 – SQL-like feature matching

Policy Decision	Best Practice
What level of training does staff need to perform feature matching?	At a minimum, staff should be trained to understand how to create and work with simple database applications such as Microsoft Access databases.
Should case-sensitivity be enforced?	Case-sensitivity should not be enforced in feature matching. All data should be converted to upper case as per NAACCR data standards.

8.2 CLASSIFICATIONS OF MATCHING ALGORITHMS

Feature-matching algorithms generally can be classified into two main categories: deterministic and probabilistic. A **deterministic matching method** is based on a series of rules that are processed in a specific sequence. These can be thought of as binary operations; a feature is either matched or it is not. In contrast, a **probabilistic matching method** uses a computational scheme to determine the likelihood, or probability, that a feature matches and returns this value for each feature in the reference set.

It should be noted that each normalization process from Section 6.2 can be grouped into these two same categories. Substitution-based normalization is deterministic, while context- and probability-based are probabilistic. Address normalization can be seen as a higher-resolution version of the feature-matching algorithm. Whereas feature-matching maps the entire set of input attributes from the input data to a reference feature, address normalization matches each component of the input address to its corresponding address attribute. These processes are both linking records to a reference set—actual features in the case of feature matching and address attributes in the case of normalization. Note that Boscoe (2008) also can be consulted for a discussion of portions of the matching techniques presented in this section.

8.3 DETERMINISTIC MATCHING

The main benefit of deterministic matching is the ease of implementation. These algorithms are created by defining a series of rules and a sequential order in which they should be applied. The simplest possible matching rule is the following:

> "Match all attributes of the input address to the corresponding attributes of the reference feature."

This rule will either find and return a perfect match, or it will not find anything and subsequently return nothing; a binary operation. Because it is so restrictive, it is easy to imagine cases when this would fail to match a feature even though the feature exists in reality (i.e., false negatives). As one example, consider a common scenario in which the reference dataset contains more descriptive attributes than the input address, as is seen in the following two example items. The first is an example postal address with only the attributes street number and name defined. The second (Table 23) depicts a reference feature that is more descriptive (i.e., it includes the pre-directional and suffix attributes_:

> "3620 Vermont"

Table 23 – Attribute relation example, linear-based reference features

From	To	Pre-directional	Name	Suffix
3600	3700	South	Vermont	Ave

In both of these cases, the restrictive rule would fail to match and no features would be returned when one or two (possibly) should have been.

8.3.1 Attribute Relaxation

In practice, less restrictive rules than the one previously listed tend to be created and applied. **Attribute relaxation**, the process of easing the requirement that all street address

attributes must exactly match a feature in the reference data source to obtain a matching street feature, often is applied to create these less restrictive rules. It generally is only applied in deterministic feature matching because probabilistic methods can account for attribute discrepancies through the weighting process. Relaxation is commonly performed by removing or altering street address attributes in an iterative manner using a predefined order, thereby increasing the probability of finding a match while also increasing the probability of error. Employing attribute relaxation, the rule previously defined could become:

> "Match all attributes *which exist* in the input address to the corresponding attributes of the reference feature"

In this case, missing attributes in the input data will not prohibit a match and the feature "3600-3700 South Vermont Ave" can be matched and returned. This example illustrates how to allow attributes present in the reference features to be missing in input data, but there is nothing stopping a matching algorithm from allowing the disconnect the other way around, with attributes missing from the reference dataset but present in the input data. However, this example also shows how ambiguity can be introduced. Take the same relaxed matching rule and apply it to the features listed in Table 24 and two matches would be returned. More detail on feature-matching ambiguity is provided in Section 14.

Table 24 – Attribute relation example, ambiguous linear-based reference features

From	To	Pre-directional	Name	Suffix
3600	3700	South	Vermont	Ave
3600	3700		Vermont	Pl

It is important to reiterate that relaxation algorithms should be implemented in an iterative manner, relaxing attributes in a specific order through a pre-defined series of steps and passes (Levine and Kim 1998). A **pass** relaxes a single (or multiple) attributes within a step. These passes start with the least descriptive attributes (those whose removal creates the least amount of error) and progress upward through more and more descriptive attributes. A **step** relaxes a single (or multiple) attributes at once, such that: (1) the resulting certainty of the relaxed address effectively moves to another level of geographic resolution, the (2) ambiguity introduced increases exponentially, or (3) the complexity of an interactive exhaustive disambiguation increases linearly.

Within each step, several passes should be performed. These passes should relax the different attributes individually and then in conjunction, until no more combinations can be made without resulting in a step to another level of geographic resolution. The order in which they are relaxed can be arbitrary and will have minimal consequence because steps are the real influencing factor. Note that relaxing the house number increases the ambiguity linearly because n = number of houses on street, while relaxing all other attributes increases the ambiguity exponentially because n = the number of possible new segments that can be included.

The preferred order of steps and passes is displayed in Table 25 through Table 27 (the pass ordering has been arbitrarily selected). The ambiguity column describes the domain of potential matches that could all equally be considered likely. The relative exponent and magnitude of ambiguity column is an estimate that shows how the magnitude of ambiguity should be calculated and the order of the derived exponent of this ambiguity (in

parentheses). The relative magnitude of spatial error column is an estimate of the total area within which the correct address should be contained and the exponent of this ambiguity (in parentheses). The worst-case resolution column lists the next level of accuracy that could be achieved when disambiguation is not possible and assumes that the lower-order attributes below those that are being relaxed are correct. Note that the last two rows of Table 26 could belong to either pass 5 or 6 because the ambiguity has increased exponentially and the search complexity has increased linearly, but the effective level of geographic certainty remains the same (USPS ZIP Code).

Table 25 – Preferred attribute relaxation order with resulting ambiguity, relative magnitudes of ambiguity and spatial error, and worst-case resolution, passes 1 – 4

Step	Pass	Relaxed Attribute	Ambiguity	Relative Exponent and Magnitude of Ambiguity	Relative Magnitude of Spatial Error	Worst-Case Resolution
1	1	none	none	(0) none	certainty of address location	single address location
2	1	number	multiple houses on single street	(0) # houses on street	length of street	single street
3	1	pre	single house on multiple streets	(1) # streets with same name and different pre	bounding area of locations containing same number house on all streets with the same name	USPS ZIP Code
3	2	post		(1) # streets with same name and different post		
3	3	type		(1) # streets with same name and different type		
4	1	number, pre	multiple houses on multiple streets	(2) # houses on street * # streets with same name and different pre	bounding area of all streets with the same name	
4	2	number, type		(2) # houses on street * # streets with same name and different type		
4	3	number, post		(2) # houses on street * # streets with same name and different post		

Table 26 – Preferred attribute relaxation order with resulting ambiguity, relative magnitudes of ambiguity and spatial error, and worst-case resolution, pass 5

Step	Pass	Relaxed Attribute	Ambiguity	Relative Magnitude of Ambiguity	Relative Magnitude of Spatial Error	Worst-Case Resolution
5	1	pre, type	single house on multiple streets	(2) # streets with same name and different pre * # streets with same name and different type	bounding area of locations containing same number house on all streets with the same name	USPS ZIP Code
5	2	pre, post		(2) # streets with same name and different pre * # streets with same name and different post		
5	3	post, type		(2) # streets with same name and different pre * # streets with same name and different type		
5	5	number, pre, type	multiple houses on multiple streets	(2) # houses on street * # streets with same name and different pre * # streets with same name and different type	bounding area of all streets with the same name	
5	6	number, pre, post		(2) # houses on street * # streets with same name and different pre * # streets with same name and different post		
5	7	number, post, type		(2) # houses on street * # streets with same name and different post * # streets with same name and different type		
5	8	number, pre, post, type		(2) # houses on street * # streets with same name and different pre * # streets with same name and different post * # streets with same name and different type		
5/6	9	pre, post, type	single house on multiple streets	(3) # streets with same name and different pre * # streets with same name and different post * # streets with same name and different type	bounding area of locations containing same number house on all streets with the same name	USPS ZIP Code
5/6	10	number pre, post, type	single house on multiple streets	(3) # houses on street * # streets with same name and different pre * # streets with same name and different post * # streets with same name and different type	bounding area of all streets with the same name	USPS ZIP Code

Table 27 – Preferred attribute relaxation order with resulting ambiguity, relative magnitudes of spatial error, and worst-case resolution, pass 6

Step	Pass	Relaxed Attribute	Ambiguity	Relative Magnitude of Ambiguity	Relative Magnitude of Spatial Error	Worst-Case Resolution
6	2	pre, type, USPS ZIP Code	single house on multiple streets in multiple USPS ZIP Codes	(3) # streets with same name and different pre * # streets with same name and different type * # USPS ZIP Codes that have those streets	bounding area of locations containing same number house on all streets with the same name in all USPS ZIP Codes	city
6	3	pre, post, USPS ZIP Code		(3) # streets with same name and different pre * # streets with same name and different post * # USPS ZIP Codes that have those streets		city
6	4	post, type, USPS ZIP Code		(3) # streets with same name and different post * # streets with same name and different type * # USPS ZIP Codes that have those streets		city
6	4	number, pre, type, USPS ZIP Code	multiple houses on multiple streets in multiple USPS ZIP Codes	(3) # houses on street * # streets with same name and different pre * # streets with same name and different type * # USPS ZIP Codes that have those streets	bounding area of all streets with the same name in all USPS ZIP Codes	city
6	5	number, pre, post, USPS ZIP Code		(3) # houses on street *# streets with same name and different pre * # streets with same name and different post * # USPS ZIP Codes that have those streets		city
6	6	number, post, type, USPS ZIP Code		(3) # houses on street *# streets with same name and different post * # streets with same name and different type * # USPS ZIP Codes that have those streets		city
6	4	number, pre, type, post, USPS ZIP Code		(3) # houses on street *# streets with same name and different pre * # streets with same name and different post * # streets with same name and different type * # USPS ZIP Codes that have those streets		city

An example of the first few iterations of the algorithm is depicted in Figure 14. This diagram shows how each step moves the certainty of the result to a lower geographic resolution. It should be noted that the authors who originally developed these attribute relaxation techniques recommend never relaxing the street name attribute (Levine and Kim 1998). In their case, this action led to the introduction of a great deal of error due to the similarity in different Hawaiian street names. Best practices relating to deterministic feature matching are listed in Best Practices 24.

Step	Pass								
1	1	100	W	4th	St	N	Los Angeles	CA	90013
2	1		W	4th	St	N	Los Angeles	CA	90013
3	1	100		4th	St	N	Los Angeles	CA	90013
3	2	100	W	4th		N	Los Angeles	CA	90013
3	3	100	W	4th	St		Los Angeles	CA	90013
4	1			4th	St	N	Los Angeles	CA	90013
4	2		W	4th		N	Los Angeles	CA	90013
4	3		W	4th	St		Los Angeles	CA	90013
5	1	100		4th		N	Los Angeles	CA	90013
5	2	100		4th	St		Los Angeles	CA	90013
5	3	100	W	4th			Los Angeles	CA	90013
5	4			4th		N	Los Angeles	CA	90013
5	5			4th	St		Los Angeles	CA	90013
5	6		W	4th			Los Angeles	CA	90013
6	1	100		4th		N	Los Angeles	CA	
6	2	100		4th	St		Los Angeles	CA	
6	3	100	W	4th			Los Angeles	CA	
6	4	100		4th		N	Los Angeles	CA	
6	5	100		4th	St		Los Angeles	CA	
6	6	100	W	4th			Los Angeles	CA	
6	7			4th		N	Los Angeles	CA	
6	8			4th	St		Los Angeles	CA	
6	9		W	4th			Los Angeles	CA	
6	10			4th		N	Los Angeles	CA	
6	11			4th	St		Los Angeles	CA	
6	12		W	4th			Los Angeles	CA	

Figure 14 – Example relaxation iterations

Best Practices 24 – Deterministic feature matching

Policy Decision	Best Practice
When should deterministic matching be used?	Deterministic matching should be the first feature-matching type attempted.
What types of matching rules can and should be used?	Any deterministic set of rules can be used, but they should always be applied in the same order.
What order of matching rules can and should be applied?	Rules should be applied in order of decreasing restrictiveness, starting from the most restrictive such that tightly restrictive rules are applied first, and progressively less restrictive rules are applied subsequently upon a previous rule's failure.
Should attribute relaxation be allowed?	Attribute relaxation should be allowed when using deterministic feature matching.
What order should attributes be relaxed?	Attribute relaxation should occur as the series of steps and passes as listed in this document.

8.4 PROBABILISTIC MATCHING

Probabilistic matching has its roots in the fields of probability and decision theory, and has been employed in geocoding processes since the outset (e.g., O'Reagan and Saalfeld 1987, Jaro 1989). The exact implementation details can be quite messy and mathematically complicated, but the concept in general is quite simple.

The **unconditional probability (prior probability)** is the probability of something occurring, given that no other information is known. Mathematically, the unconditional probability, *P*, of an event, *e*, occurring is notated *P(e)*, and is equivalent to (1 - the probability of the event not occurring), that is, $P(e) = 1 - P(\neg e)$.

In contrast, the **conditional probability** is the probability of something occurring, given that other information is known. Mathematically, having obtained additional information, *I*, the conditional probability, *P*, of event *e* occurring given that *I* is true, $P(i \mid e)$, defined as the probability of *I* and *e* occurring together divided by the probability that *e* that occurs alone as in Equation 1.

$$P(i \mid e) = \frac{P(i \wedge e)}{P(e)}$$

Equation 1 – Conditional probability

In probabilistic matching, the **match probability** is a degree of belief ranging from 0 to 1 that a feature matches. These systems report this degree of belief that a feature matches (the easy part) based on and derived from some criteria (the hard part). A degree of belief of 0 represents a 0 percent chance that it is correct, while a 1 represents a 100 percent chance. The **confidence threshold** is the probability cutoff point determined by the user above which a feature is accepted and below which it is rejected. To harness the power of these probabilities and achieve feature results that would not otherwise be obtainable, the use of this approach requires the acceptance of a certain level of risk that an answer could be wrong.

There are many forms of probabilistic feature matching, as the entire field of record linkage is focused on this task. Years of research have been devoted to this problem, with

particular interest paid to health and patient records (e.g., Winkler 1995, Blakely and Salmond 2002). In this section, to illustrate the basic concepts and present a high-level overview, one common approach will be presented: attribute weighting.

8.4.1 Attribute weighting

Attribute weighting is a form of probabilistic feature matching in which probability-based values are associated with each attribute, and either subtract from or add to the **composite score** for the feature as a whole. Then, the composite score is used to determine a match or non-match. In this approach each attribute of the address is assigned two probabilities, known as weights. These **weights** represent the level of importance of the attribute, and are a combination of the matched and unmatched probabilities. The **matched probability** is the probability of two attributes matching, m, given that the two records match, M. Mathematically, this is denoted as the conditional probability $P(m|M)$. This probability can be calculated with statistics over a small sample of the total dataset in which the input datum and the reference feature do actually match. The **error rate,** α, denotes instances in which the two attributes do not actually match, even though the two records do match. Thus, $P(m|M) = 1 - \alpha$. In the existing literature, the full probability notation usually is discarded, and $P(m|M)$ is simply written as m. It should be noted that α generally is high.

The **unmatched probability** is the probability that the two attribute values match, m, given that the two records themselves do not match, $\neg M$. Mathematically, this denoted by the conditional probability $P(m|\neg M)$. This second probability represents the likelihood that the attributes will match at random, and can be calculated with statistics over a small sample of the total dataset for which the input data and the reference do not match. Again, $P(m|\neg M)$ usually is denoted simply as u. It should be noted that u generally is low for directionals, but is higher for street names.

From these two probabilities m and u, frequency indices for agreement, f_a , and disagreement, f_d , can be calculated and used to compute the positive and negative weights for agreement and disagreement, w_a , and, w_d , as in Equation 2.

$$f_a = \frac{m}{u}, f_d = \frac{1-m}{1-u}$$
$$w_a = \log_2(f_a), w_d = \log_2(f_d)$$

Equation 2 – Agreement and disagreement probabilities and weights

These weights are calculated for each of the attributes in the reference dataset *a priori*. Composite scores for input data are created on-the-fly during feature matching by summing the attribute weights of the individual input attributes as compared against the reference feature attributes. Where an agreement is found, w_a , it is added to the score, and where a disagreement is found, w_d , it is subtracted. This composite score is the probability used to determine if the feature is a match (i.e., if it is above the confidence threshold). Excellent descriptions of this and other more advanced record linkage algorithms can be found in Jaro (1989), Blakely and Salmond (2002), Meyer et al. (2005), and Boscoe (2008) as well as in the references contained within each. Best practices related to probabilistic feature matching are listed in Best Practices 25.

Best Practices 25 – Probabilistic feature matching

Policy Decision	Best Practice
When should probabilistic matching be used?	Probabilistic matching should be used when deterministic feature matching fails, and if the consumers of the data are comfortable with the confidence threshold.
What confidence threshold should be considered acceptable?	At a minimum, a 95% confidence threshold should be acceptable.
What metadata should be maintained?	The metadata should describe the match probability.
How and when should match probabilities for different attributes be calculated?	Match probabilities for different attributes should be calculated *a priori* for the reference dataset by using a computational approach that randomly selects records and iterates continuously until the rate stabilizes.
How and when should unmatch probabilities for different attributes be calculated?	Unmatch probabilities for different attributes should be calculated *a priori* for the reference dataset by using a computational approach that randomly selects records and iterates continuously until the rate stabilizes.
How and when should confidence thresholds be re-evaluated?	Confidence thresholds should continuously be re-evaluated based on the frequency with which attribute values are encountered.
How and when should composite weights be re-evaluated?	Composite weights should continuously be re-evaluated based on the frequency with which attribute values are encountered.

8.5 STRING COMPARISON ALGORITHMS

Any feature-matching algorithm requires the comparison of strings of character data to determine matches and non-matches. There are several ways this can be attempted, some more restrictive or flexible in what they are capable of matching than others. The first, **character-level equivalence,** enforces that each character of two strings must be exactly the same. In contrast, **essence-level equivalence** uses metrics capable of determining if two strings are "essentially" the same. This allows for minor misspellings in the input address to be handled, returning reference features that "closely match" what the input may have "intended." These techniques are applicable to both deterministic and probabilistic matching algorithms because relaxing the spelling of attributes using different string matching algorithms is a form of attribute relaxation. In all cases, careful attention must be paid to the accuracy effects when these techniques are employed because they can and do result in incorrect features being returned.

Word stemming is the simplest version of an essence-level equivalence technique. These algorithms reduce a word to its root (stem), which then is used for essence-level equivalence testing. The **Porter Stemmer** (Porter 1980) is the most famous of these. It starts by removing common suffixes (e.g., "-ed," "-ing,") and additionally applies more complex rules for specific substitutions such as "-sses" being replaced with "-ss." The algorithm is fairly straightforward and run as a series of steps. Each progressive step takes into account what has been done before, as well as word length and potential problems with a stem if a suffix is removed.

Phonetic algorithms provide an alternative method for encoding the essence of a word. These algorithms enable essence-level equivalence testing by representing a word in terms of how it sounds when it is pronounced (i.e., phonetically). The goal of these types of algorithms is to produce common representations for words that are spelled differently, yet sound the same. The **Soundex** algorithm is the most famous of this class of algorithms. It has existed since the late 1800s and originally was used by the U.S. Census Bureau. The algorithm is very simple and consists of the following steps:

1) Keep the first letter of the string
2) Remove all vowels and the letters y, h, and w, unless they are the first letter
3) Replace all letters after the first with numbers based on a known table
4) Remove any numbers which are repeated in a row
5) Return the first four characters, padded on the right with zeros if there are less than four.

Producing an encoded form of any information necessarily loses information (unless they are defined as exact equivalents). Stemming and phonetic algorithms, while efficient and precise, still suffer from this fact and can produce inaccurate results in the context of matching street names. In particular, two originally unrelated attribute values can become related during the process. Table 28 presents examples of words encoded by both algorithms that result in ambiguities.

Table 28 – String comparison algorithm examples

Original	Porter Stemmed	Soundex
Running Ridge	run ridg	R552 R320
Runs Ridge	run ridg	R520 R320
Hawthorne Street	hawthorn street	H650 S363
Heatherann Street	heatherann street	H650 S363

To minimize these negative effects or data loss, feature-matching algorithms can attempt string comparisons as a two-step process. The first pass can use an essence-level comparison to generate a set of candidate reference features. The second pass then can generate a probability-based score for each of the candidates using the original text of the attributes, not the essence-level derivations. The values from the second pass then can be used to determine the likelihood of correctness. Best practices related to string comparison algorithms are listed in Best Practices 26.

Best Practices 26 – String comparison algorithms

Policy Decision	Best Practice
When and how should alternative string comparison algorithms be used?	Alternative string comparison algorithms should be used when no exact feature matches can be identified. A two-step approach should be used to compare the original input with the essence-level equivalence match to determine the match and unmatched probabilities (as in the probability-based feature-matching approach).
What types of string comparison algorithms can and should be used?	Both character- and essence-level string comparisons should be supported.
When should character-level string equivalence be used?	Character-level equivalence should always be attempted first on every attribute.
When and how should essence-level string equivalence be used?	Essence-level equivalence should only be attempted if character-level equivalence fails. Essence-level equivalence should only be attempted on attributes other than the street name. Only one essence-level equivalence algorithm should be applied at a time. They can be tried in succession but one should not process the output of the other (i.e., they should both start with the raw data). Metadata should describe the calculated essence of the string used for comparison, and strings that it was matched to in the reference dataset.
What types of essence-level algorithms should be used?	Both stemming and phonetic algorithms should be supported by the geocoding process.
Which word-stemming algorithms should be used?	At a minimum, the Porter Stemmer (Porter 1980) should be supported by the geocoding process.
Which phonetic algorithms should be used?	At a minimum, the Soundex algorithm should be supported by the geocoding process.

9. FEATURE INTERPOLATION

This section examines each of the feature interpolation algorithms in depth.

9.1 FEATURE INTERPOLATION ALGORITHMS

Feature interpolation is the process of deriving an output geographic feature from geographic reference features (e.g., deriving a point for an address along a street center-line or the centroid of a parcel). A **feature interpolation algorithm** is an implementation of a particular form of feature interpolation. One can distinguish between separate classes of feature interpolation algorithms for linear- and areal unit-based reference feature types. Each implementation is tailored to exploit the characteristics of the reference feature types upon which it operates.

It is useful to point out that interpolation is only *ever* required if the requested output geographic format is of lower geographic complexity than the features stored in the reference dataset. If a geocoding process uses a line-based reference dataset and is asked to produce a line-based output, no interpolation is necessary because the reference feature is returned in its native form. Likewise, a polygon-based reference dataset should return a native polygon representation if the output format requests it.

Linear-based interpolation is most commonly encountered, primarily because linear-based reference datasets currently are the most prevalent. The advantages and disadvantages of each type of interpolation method will be explored in this section.

9.2 LINEAR-BASED INTERPOLATION

Linear-based feature interpolation operates on segments lines (or **polylines**, which are a series of connected lines) and produces an estimation of an output feature using a computational process on the spatial geometry of the line. This approach was one of the first implemented, and as such, is detailed dozens of times in the scientific literature and the user manuals of countless geocoding platforms. With this abundance of information on the topic and data sources readily available (see Table 17), the discussion presented here will outline only the high-level details, focusing on identifying assumptions used in the process that affect the results and ways that they can be overcome.

For the purpose of this discussion, it will be assumed that the input data and the reference feature are correctly matched. In essence, linear-based interpolation attempts to estimate where along the reference feature the spatial output—in this case a point—should be placed. This is achieved by using the number attribute of the input address data to identify the proportion of the distance down the total length of the reference feature where the spatial output should be placed. The reference feature attribute used for this operation is the **address range,** which describes the valid range of addresses on the street (line segment) in terms of start and end addresses (and also serves to make street vectors continuous geographic objects).

The **address parity** (i.e., even or odd) is an indication of which side of the street an input address falls. This simplistic case presumes **binary address parity** for the reference street segment (i.e., one side of the street is even and the other is odd), which may not be the

case. More accurate reference data sometimes account for different parities on the same side of the street as necessary (**non-binary address parity**) and a more advanced geocoding algorithm can take advantage of these attributes. A common parity error for a reference data source is for an address to be listed as if it occurs on both sides of the street. An equally common address range error in a reference data source is for an address range to be reversed. This can mean that the address is on the wrong sides of the street, that the address range start and end points of the street have been reversed, or a combination of both. These should be considered reference data source errors, not interpolation errors, although they are commonly viewed that way.

In an effort to continue with the simplest possible case, interpolation will be performed on a simple line-based reference feature made up of only two points (i.e., the start, or origin, and end, or destination). The distance from the start of the street segment where the spatial output should be placed, d, is calculated as a proportion of the total street length, l, the number of the input address, a, and the size of the address range, r, which is equal to one-half the difference between the start address and end address of the address range, r_s and r_e respectively, as in Equation 3.

$$r = \frac{Abs(r_s - r_e)}{2}, \quad d = l\left(\frac{a}{r}\right)$$

Equation 3 – Size of address range and resulting distance from origin

Using the distance that the output point should be located from the origin of the street vector, it is possible to calculate the actual position where the spatial output should be placed. This is achieved is through the following calculation with the origin of the street denoted x_0,y_0, the destination is denoted x_1,y_1, and the output location is denoted x_2,y_2, as in Equation 4. Note that although the Earth is an ellipsoid and spherical distance calculations would be the most accurate choice, planar calculations such as Equation 4 are most commonly employed because the error they introduce is negligible for short distances such as most typical street segments.

$$x_2 = d(x_1 - x_0),$$
$$y_2 = d(y_1 - y_0)$$

Equation 4 – Resulting output interpolated point

This calculated position will be along the centerline of the reference feature, corresponding to the middle of the street. Thus, a **dropback** usually is applied to move the output location away from the centerline toward and/or beyond the sides of the street where the buildings probably are located in city-style addresses.

Experiments have been performed attempting to determine the optimal direction and length for this dropback but have found that the high variability in street widths and directions prohibits consistent improvements (Ratcliffe 2001, Cayo and Talbot 2003). Therefore, in practice, an orthogonal direction usually is chosen along with a standard distance. However, it is likely that better results could be achieved by inspecting the MTFCC of a road to determine the number of lanes and multiplying by the average width per lane. Best practices related to these fundamental components of the linear-based interpolation methods are listed in Best Practices 27.

Best Practices 27 – Linear-based interpolation

Policy Decision	Best Practice
When should linear-based interpolation be used?	Linear-based interpolation should be used when complete and accurate point- or polygon-based reference datasets are not available. Linear-based interpolation should be used when input data cannot be directly linked with a point-based reference dataset and must be matched to features representing multi-entity features.
When and how should the parameters for linear interpolation be chosen?	The parameters used for linear-based interpolation should be based upon the attributes available in the reference dataset.
What parity information should be used for linear-based feature interpolation?	At a minimum, binary parity should be used. If more information is available in the reference dataset regarding the parity of an address it should be used (e.g., multiple address ranges per side of street).
What linear-interpolation function should be used?	At a minimum, planar interpolation should be used. If a spherical interpolation algorithm is available it should be used.
Should the same dropback value and direction always be used?	The same dropback value and direction should always be used based on the width of the street as determined by: • Number of lanes • MTFCC codes • Average width per lane
Which dropback value and direction can and should be used?	An *a priori* dropback value of one-half the reference street's width (based on the street classification code and average classification street widths) should be applied in an orientation orthogonal to the primary direction of the street segment to which the interpolated output falls.

When performing linear-based interpolation in the manner just described, several assumptions are involved and new geocoding methods are aimed at eliminating each (e.g., Christen and Churches [2005] and Bakshi et al. [2004]). The **parcel existence assumption** is that all addresses within an address range actually exist. The **parcel homogeneity assumption** is that each parcel is of exactly the same dimensions. The **parcel extent assumption** is that addresses on the segment start at one endpoint of the segment and completely fill the space on the street all the way to the other endpoint. These concepts are illustrated in Figure 15. Additionally, the **corner lot assumption/problem** is that when using a measure

of the length of the segment for interpolation, it is unknown how much real estate may be taken up along a street segment by parcels from other intersecting street segments (around the corner), and the actual street length may be shorter than expected. **Address-range feature interpolation** is subject to all of these assumptions (Bakshi et al. 2004).

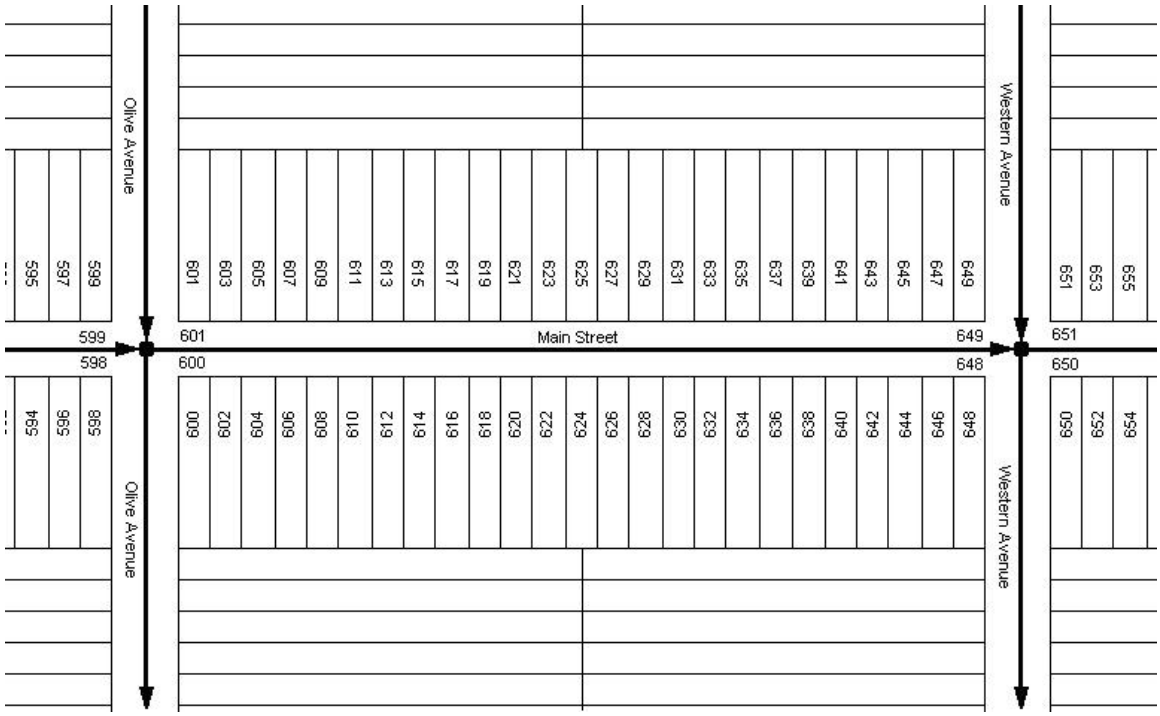


Figure 15 – Example of parcel existence and homogeneity assumptions

Recent research has attempted to address each of these assumptions by incorporating additional knowledge into the feature interpolation algorithm about the true characteristics of the reference feature (Bakshi et al. 2004). First, by determining the true number of buildings along a reference feature, the parcel existence assumption can be alleviated. By doing this, the distance to the proper feature can be calculated more accurately. However, this approach still assumes that each of the parcels is of the same size, and is thus termed **uniform lot feature interpolation.** This is depicted in Figure 16.

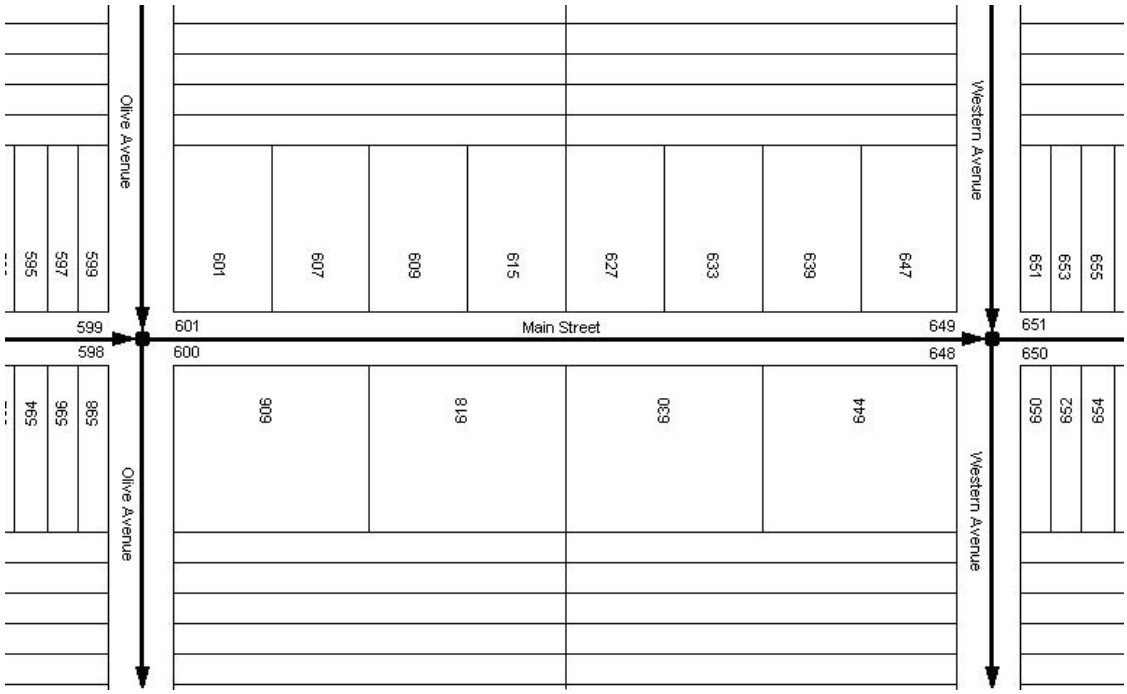


Figure 16 – Example of uniform lot assumption

If the actual parcel sizes are available, the parcel homogeneity assumption can be overcome and the actual distance from the origin of the street segment can be calculated directly by summing the distances of each parcel until the correct one is reached, and is thus termed **actual lot feature interpolation**. This is depicted in Figure 17.

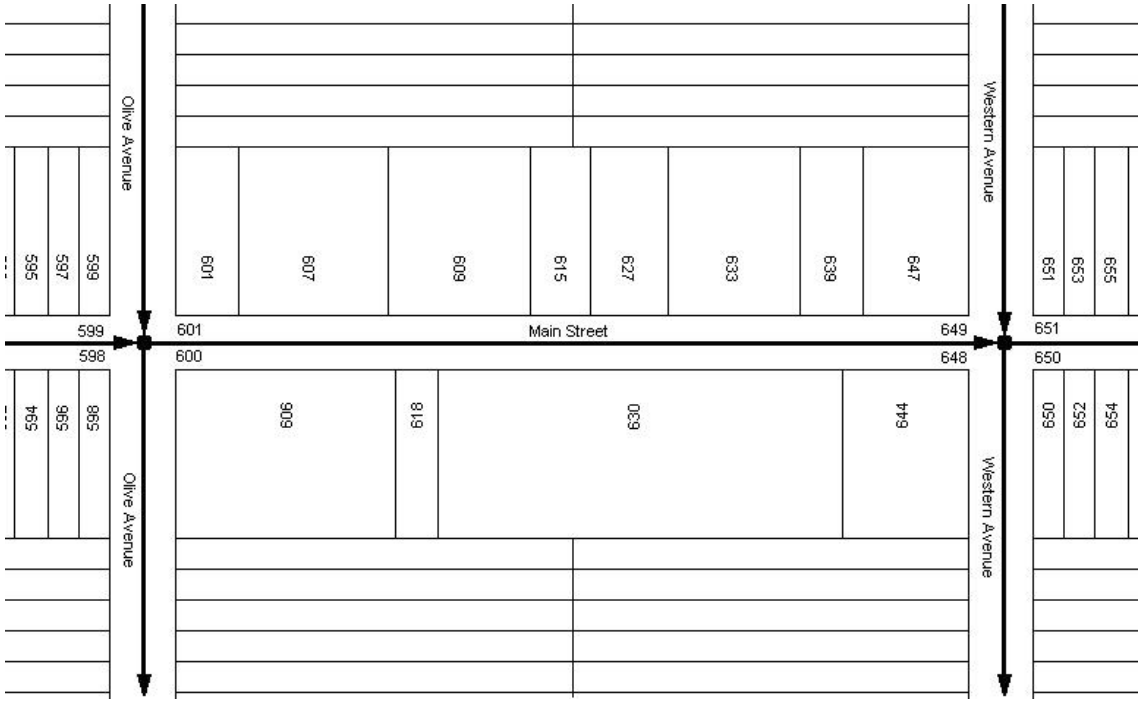


Figure 17 – Example of actual lot assumption

However, the distance is still calculated using the parcel extent assumption that the addresses on a block start exactly at the endpoint of the street. This obviously is not the case because the endpoint of the street represents the intersection of centerlines of intersecting streets. The location of this point is in the center of the street intersection, and therefore the actual parcels of the street cannot start for at least one-half the width of the street (i.e., where the curb starts). This is depicted in Figure 18.

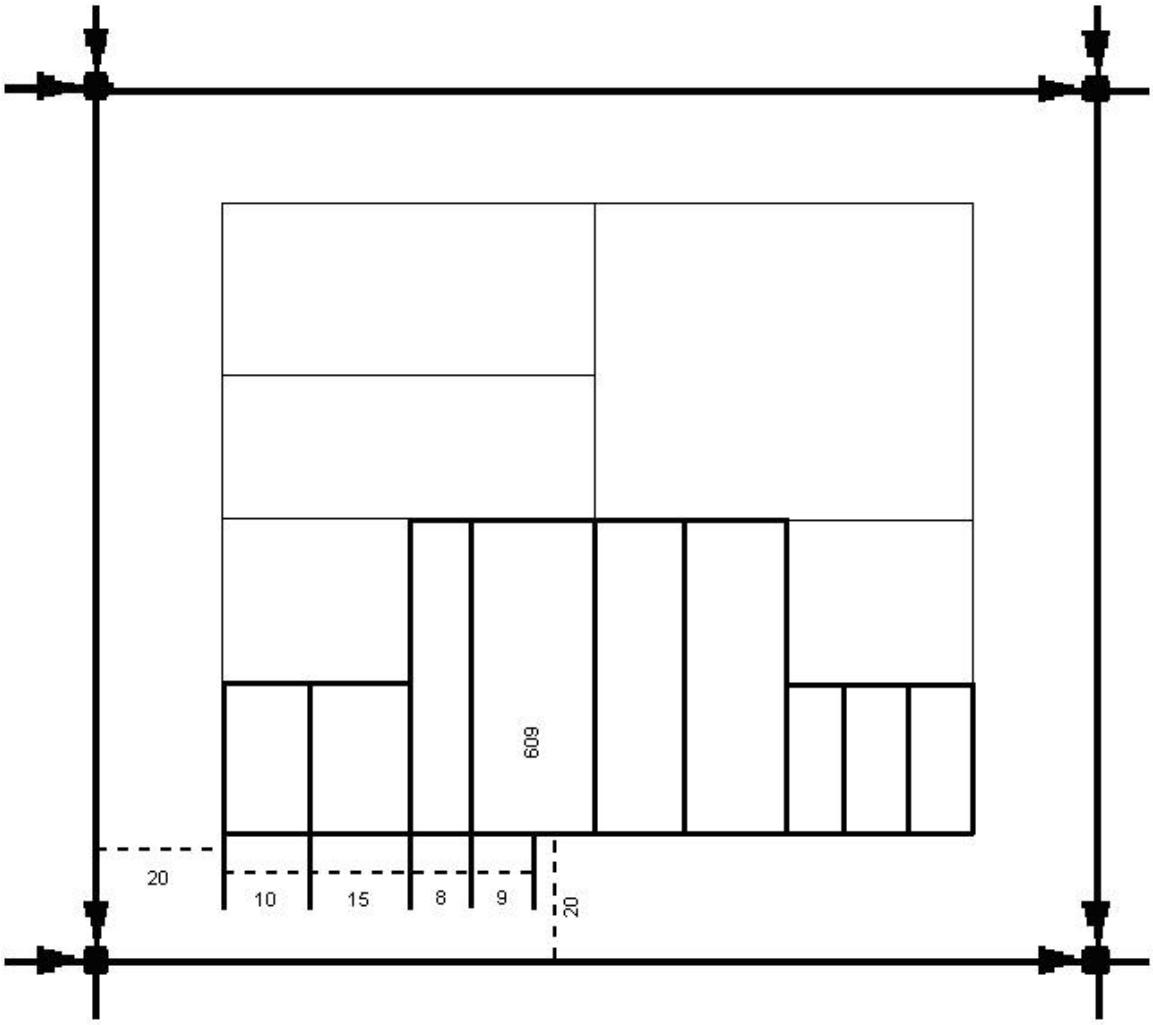


Figure 18 – Example of street offsets

The **corner lot problem** can be overcome in a two-step manner. First, the segments that make up the block must be determined. Second, an error-minimizing algorithm can be run to determine the most likely distribution of the parcels for the whole block based on the length of the street segments, the sizes of lots, and the combinations of their possible layouts. This distribution then can be used to derive a better estimate of the distance from the endpoint to the center of the correct parcel. This is depicted in Figure 19. None of the approaches discussed thus far can overcome the assumption that the building is located at the centroid of the parcel, which may not be the case.

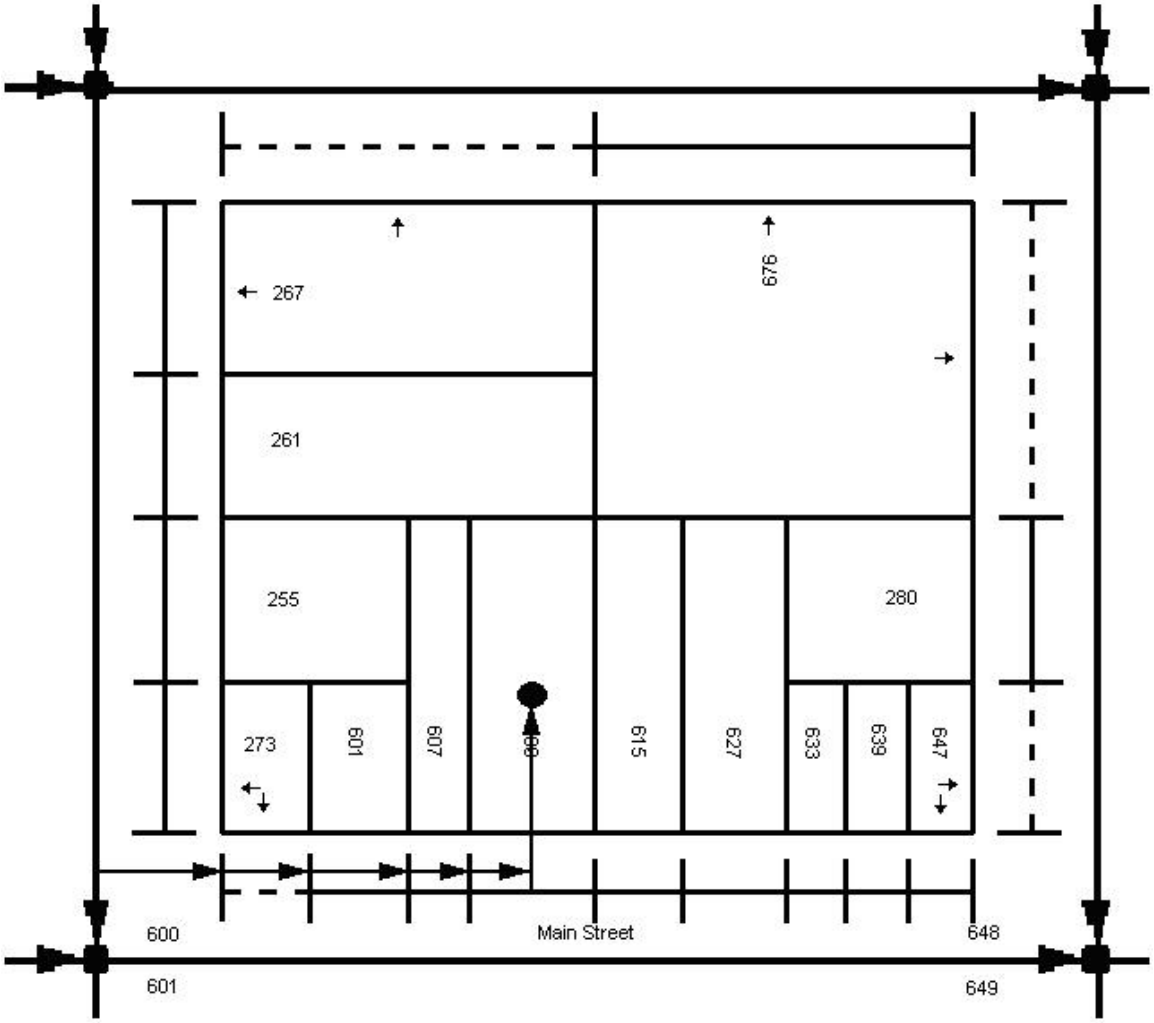


Figure 19 – Example of corner lot problem

These small performance gains in the accuracy of the linear-based feature interpolation algorithm may hardly seem worth the effort, but this is not necessarily the case. Micro-scale spatial analyses, although not currently performed with great frequency or regularity, are becoming more and more prevalent in cancer- and health-related research in general. For example, a recent study of exposure to particulate matter emanating from freeways determined that the effect of this exposure is reduced greatly as one moves small distances away from the freeway, on the order of several meters (i.e., high-distance decay). Thus, if the accuracy of the geocoding process can be improved by just a few meters, cases can more accurately be classified as exposed or not (Zandbergen 2007) and more accurate quantifications of potential individual exposure levels can be calculated, as has been attempted with pesticides (Rull and Ritz 2003, Nuckols et al. 2007) for example. Best practices related to linear-based interpolation assumptions are listed in Best Practices 28.

Best Practices 28 – Linear-based interpolation assumptions

Policy Decision	Best Practice
When and how should linear-based interpolation assumptions be overcome?	If data are available and/or obtainable, all assumptions that can be overcome should be.
Where can data be obtained to overcome linear-based interpolation assumptions?	Local government organizations should be contacted to obtain information on the number, size, and orientation of parcels as well as address points.
When and how can the parcel existence assumption be overcome?	If an address verifier is available, it should be used to verify the existence of parcels before interpolation is performed.
When and how can the parcel homogeneity assumption be overcome?	If the parcel dimensions are available, these should be used to calculate the interpolated output location.
When and how can the parcel extent assumption be overcome?	If the street widths are known or can be derived from the attributes of the data (street classification and average classification widths), these should be used to buffer the interpolation range geometry before performing interpolation.
When and how can the corner lot problem be overcome?	If the layout and sizes of parcels for the entire block are available, they should be used in conjunction with the lengths of the street segments that compose the blocks to determine an error-minimizing arrangement which should be used for linear-based interpolation.

9.3 AREAL UNIT-BASED FEATURE INTERPOLATION

Areal unit-based feature interpolation uses a computational process to determine a suitable output from the spatial geometry of polygon-based reference features. This technique has a unique characteristic—the possibility to be both very accurate or very inaccurate, depending on the geographic scale of the reference features used. For instance, areal unit-based interpolation on parcel-level reference features should produce very accurate results compared to linear-based interpolation for the same input feature. However, areal unit-based interpolation at the scale of a typical USPS ZIP Code would be far less accurate in comparison to a linear-based interpolation for the same input data (noting again that USPS ZIP Codes are not actually areal units; see Section 5.1.4).

Centroid calculations (or an approximation thereof) are the usual interpolation performed on areal unit-based reference features. This can be done via several methods, with each emphasizing different characteristics. The simplest method is to take the centroid of the bounding box of the feature and often is employed in cases for which complex computations are too expensive. A somewhat more complicated approach, the center-of-mass or geographic centroid calculation, borrows from physics and simply uses the shape and area to compute the centroid. This does not take into account any aspatial information about the contents of the areal unit that might make it more accurate. At the resolution of an urban

parcel, this has been shown to be fairly accurate because the assumption that a building is in the center of a parcel is mostly valid, as long as the parcels are small (Ratcliffe 2001).

However, as parcels increase in size (e.g., as the reference dataset moves from an urban area characterized by small parcels to a rural area characterized by larger parcels) this assumption becomes less and less valid and the centroid calculation becomes less accurate. In particular, on very large parcels such as farms or campuses, the center of mass centroid becomes very inaccurate (Stevenson et al. 2000, Durr and Froggatt 2002). In contrast, algorithms that employ a weighted centroid calculation sometimes are more accurate when applied to these larger parcels. These make use of the descriptive quality of the aspatial attributes associated with the reference feature (e.g., population density surfaces) to move the centroid toward a more representative location.

To achieve this, the polygon-based features can be intersected with a surface created from finer resolution features to associate a series of values for each location throughout the polygon. This weight surface can be derived from both raster-based and individual feature reference data. In either case, the weighted centroid algorithm runs on top of this surface to calculate the position of the centroid from the finer resolution dataset, either the raster cell values in the first case or the values of the appropriate attribute for individual features. For example, in a relatively large areal unit such as a ZCTA (granted that not all ZCTAs are large), a weighted centroid algorithm could use information about a population distribution to calculate a more representative and probable centroid. This will produce a centroid closer to where the most people actually live, thereby increasing the probability that the geocode produced is closer to where the input data really are. This surface could be computed from a raster dataset with cell values equaling population counts or from a point dataset with each point having a population count attribute; essentially a method of looking at a point dataset as a non-uniformly distributed raster dataset. See Beyer et al. (2008) for a detailed evaluation of multiple weighting schemes. Best practices related to areal unit-based interpolation are listed in Best Practices 29.

Best Practices 29 – Areal unit-based interpolation

Policy Decision	Best Practice
When should areal unit-based interpolation methods be used?	Areal unit-based interpolation should be used over linear-based alternatives when the spatial resolution of the areal unit-based reference features is higher than that of the corresponding linear-based counterparts. Areal unit-based interpolation should be used when more accurate means have been tried and failed, and it is the only option left. Areal unit interpolation should not be used if metadata about the accuracy of the features is not available.
When and which areal unit-based interpolation methods should be used?	At a minimum, geographic (center-of-mass) centroid calculations should be used. If appropriate information is available, weighted centroid approximations should be used. Feature-bounding box centroids should not be used.
Which additional data sources should be used for areal unit-based centroid approximations?	Population density should be used for weighted centroid calculation for areal unit-based reference datasets containing reference features of lower resolution than parcels (e.g., USPS ZIP Codes).
What metadata should be maintained?	If weighted centroids are calculated, the metadata for the datasets used in the calculation, identifiers for the grid cells containing the values used for calculation, and aggregates for the values used in the calculation (e.g., mean, min, max, range) should be recorded along with the geocoded record.

10. OUTPUT DATA

This section briefly discusses issues related to geocoded output data.

10.1 DOWNSTREAM COMPATIBILITY

The definition of geocoding presented earlier was specifically designed to encompass and include a wide variety of data types as valid output from the geocoding process. Accordingly, it is perfectly acceptable for a geocoding process to return a point, line, polyline, polygon, or some other higher-complexity geographic object. What must be considered, however, is that the output produced inevitably will need to be transmitted to and consumed and/or processed by some downstream component (e.g., a spatial statistical package). These requirements, capabilities, and limitations of the eventual data consumer and transmission mechanisms need to be considered when assessing an appropriate output format. In most cases, these constraints will tend to lean towards the production of simple points as output data.

10.2 DATA LOSS

When examining the available output options from a data loss perspective, one may consider a different option. Take the ambiguity problems inherent in moving from a lower resolution geographic feature to a higher one described earlier (Section 7.3), for example. The high-resolution data can always be abstracted to lower resolution later if necessary, but once converted they cannot be unambiguously converted back to their higher-resolution roots. For example, a parcel centroid can always be computed from a parcel boundary, but the other direction is not possible if new data are discovered that could have influenced the assignment of the centroid. Therefore, it may be advisable to always return and store the spatial output of the geocoding process at the highest level of geographic resolution possible. There is a risk associated with this process because of the temporal staleness problems that can occur with geocode caches (e.g., if the parcel boundaries change over time).

Best Practices 30 – Output data

Policy Decision	Best Practice
What geographic format should output data take?	At a minimum, output data should be a geographic point with a reference to the ID of the reference feature used. If other processes can handle it, the full geometry of the reference feature also should be returned.

This page is left blank intentionally.

Part 3: The Many Metrics for Measuring Quality

Notions of "quality" vary among the scientific disciplines. This term (concept) has become particularly convoluted when used to describe the geocoding process. In the information and computational sciences, the "quality" of a result traditionally refers to the notions of precision (accuracy) and recall (completeness), while in the geographical sciences these same terms take on different (yet closely related) meanings. Although the factors that contribute to the overall notion of geocode quality are too numerous and conceptually diverse to be combined into a single value, this is how it is generally described. The very nature of the geocoding process precludes the specification of any single quality metric capable of sufficiently describing the geocoded output. This part of the document will elaborate on the many metrics that can affect different aspects of quality for the resulting geocode.

This page is left blank intentionally.

11. QUALITY METRICS

This section explores several contributing factors to spatial accuracy within different components and at different levels of the geocoding process.

11.1 ACCURACY

Researchers must have a clear understanding of the quality of their data so that they can decide its fitness-for-use in their particular study. Each study undoubtedly will have its own unique data quality criteria, but in general the metrics listed in Table 29 could be used as a guide to develop a level of confidence about the quality of geocodes. Several aspects of confidence are listed along with their descriptions, factors, and example evaluation criteria.

Further research is required to determine exactly if, how, or when these metrics could be combined to produce a single "quality" metric for a geocode. The topics in this table will be covered in more detail in the following sections. An excellent review of current studies looking at geocoding quality and its effects on subsequent studies is available in Abe and Stinchcomb (2008). Research on how spatial statistical models can be used in the presence of geocodes (and geocoded dataset) of varying qualities (e.g., Zimmerman [2008]) is emerging as well.

Table 29 – Metrics for deriving confidence in geocoded results

Metric	Description	Example Factors	Example Criteria (> better than)
Precision	How close is the location of a geocode to the true location?	Interpolation algorithm used	Uniform Lot > address range
		Interpolation algorithm assumptions	Less assumptions > more assumptions
		Reference feature geometry size	Smaller > larger
		Reference feature geometry accuracy	Higher > lower
		Matching algorithm certainty	Higher > lower
Certainty	How positive can one be that the geocode produced represents the correct location?	Matching algorithm used	Deterministic > probabilistic
		Matching algorithm success	Exact match > non-exact match High probability > low probability Non-ambiguous match > ambiguous match
		Reference feature geometry accuracy	Higher > lower
		Matching algorithm relaxation amount	None > some
		Matching algorithm relaxation type	Attribute transposition > Soundex
Reliability	How much trust can be placed in the process used to create a geocode?	Transparency of the process	Higher > lower
		Reputability of software/vendor	Higher > lower
		Reputability of reference data	Higher > lower
		Reference data fitness for area	Higher > lower
		Concordance with other sources	Higher > lower
		Concordance with ground truthing	Higher > lower

12. SPATIAL ACCURACY

This section explores several contributing factors to spatial accuracy within different components and at different levels of the geocoding process.

12.1 SPATIAL ACCURACY DEFINED

The **spatial accuracy** of geocoding output is a combination of the accuracy of both the processes applied and the datasets used. The term "accuracy" can and does mean several different things when used in the context of geocoding. In general, **accuracy** typically is a measure of how close to the true value something is. General best practices related to geocoding accuracy are listed in Best Practices 31.

Best Practices 31 – Output data accuracy

Policy Decision	Best Practice
When and how can and should accuracy information be associated with output data?	Any information available about the production of the output data should always be associated with the output data.

12.2 CONTRIBUTORS TO SPATIAL ACCURACY

Refining this definition of accuracy, **spatial accuracy** can be defined as a measure of how true a geographic representation is to the actual physical location it represents. This is a function of several factors (e.g., the resolution it is being modeled at, or the geographic unit used to bind it). For example, a parcel represented as a point would be less spatially accurate than the parcel being represented as a polygon. Additionally, a polygon defined at a local scale with thousands of vertices is more spatially accurate than the same representation at the national scale, where it has been generalized into a dozen vertices. With regard to the geocoding process, spatial accuracy is used to describe the resulting geocode and the limits of the reference dataset. The resulting positional accuracy of a geocode is dependent on every component of the geocoding process. The decisions made by geocoding algorithms at each step can have both positive and negative effects, either potentially increasing or decreasing the resulting accuracy of the spatial output.

12.2.1 Input data specification

Output geocode accuracy can be traced back to the very beginning of the process when the input data initially are specified. There has been some research into associating first-order levels of accuracy with different types of locational descriptions (Davis Jr. et al. 2003, Davis Jr. and Fonseca 2007), but in practice, these distinctions rarely are quantified and returned as accuracy metrics with the resulting data. These different types of data specifications inherently encode different levels of information. As an example, consider the difference between the following two input data examples. The first is a relative locational description that, when considered as an address at the location, could refer to the address closest to the corner on either Vermont Ave or 36th Place (the corner lot problem from Section 9.2), while the second are locational data that describe a specific street address.

"The northeast corner of Vermont Avenue and 36th Place"

"3620 South Vermont Avenue, Los Angeles, CA 90089"

One description clearly inherently encodes the definition of a more precise location than the other. When compared to the specific street address, the relative description implicitly embeds more uncertainty into the location it is describing, which is carried directly into the geocoding process and the resulting geocode. Using the specific street address, a geocoding algorithm can uniquely identify an unambiguous reference feature to match. With the relative location, a geocoder can instead only narrow the result to a set of likely candidate buildings, or the area that encompasses all of them.

Similarly, the amount of information encoded into a description has a fundamental effect on the level of accuracy that can be achieved by a geocoder. Consider the difference in implicit accuracy that can be assumed between the following two input data examples. The first specifies an exact address, while the second specifies a location somewhere on a street, inherently less accurate than the first. In this case the resulting accuracy is a function of the assumed geographic resolution defined by the amount of information encoded in the input data.

"831 Nash Street, El Segundo, CA 90245"

"Nash Street, El Segundo, CA 90245"

Relationships among the implicit spatial accuracies of different types of locational descriptions are shown in Figure 20, with:

a) Depicting accuracy to the building footprint (the outline)
b) Showing how the building footprint (the small dot) is more accurate than the USPS ZIP+4 (United States Postal Service 2008a) "90089-0255" (the polygon)
c) Showing the implicit resolution of combined street segments (the straight line) within a USPS ZIP Code (blue) along "Vermont Ave, 90089"
d) Showing both a relative direction (the large polygon) "northeast corner of Vermont and 36th" and the "3600-3700 block of Vermont Ave, Los Angeles, CA 90089"
e) Showing the relation between the building (the small dot) USPS ZIP+4 (the small inner polygon) to the USPS ZIP "90089" (the larger polygon)
f-g) Showing the relations among the city, county and state.

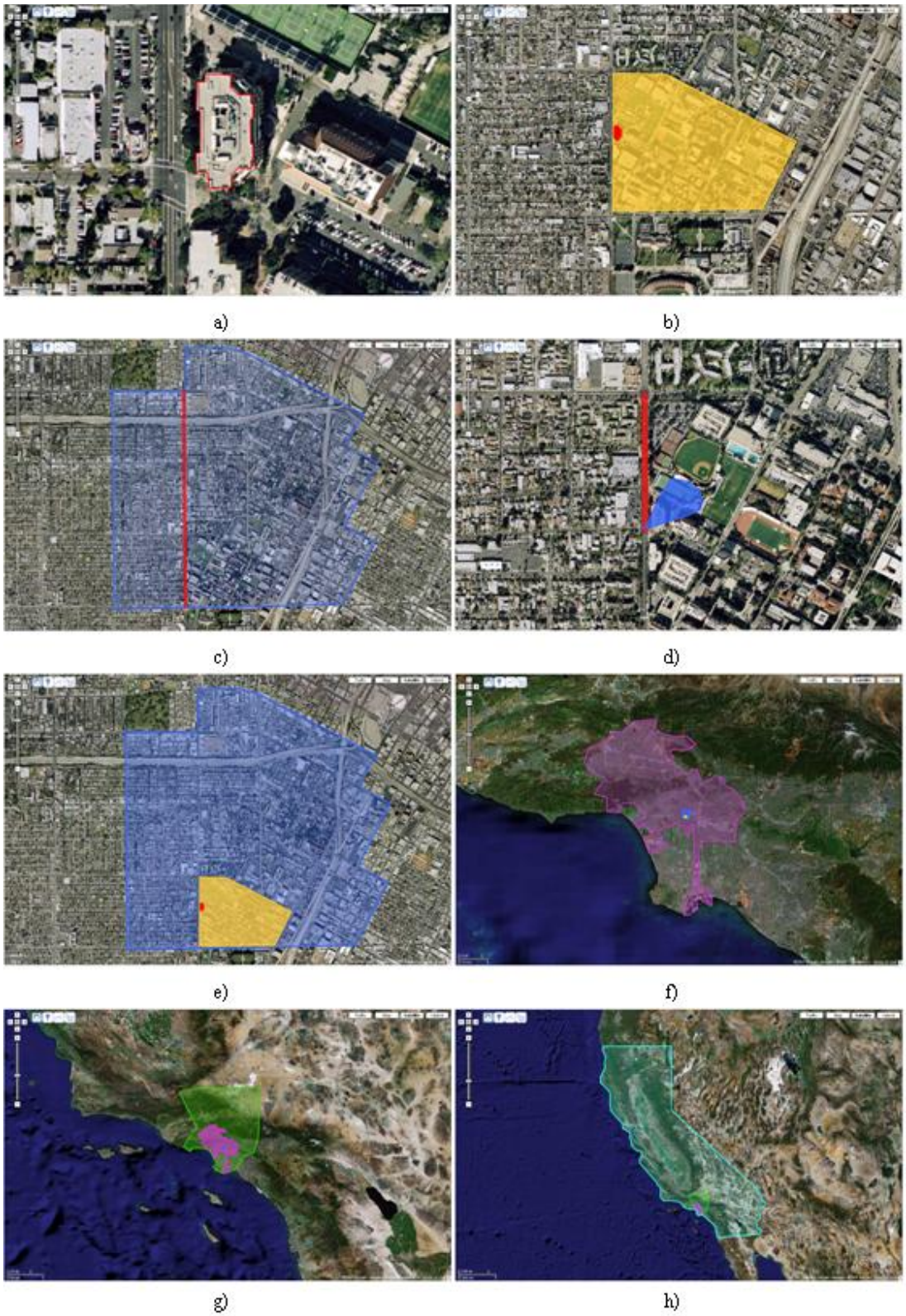


Figure 20 – Certainties within geographic resolutions (Google, Inc. 2008b)

Best practices related to input data are listed in Best Practices 32.

Best Practices 32 – Input data implicit accuracies

Policy Decision	Best Practice
What information quality metrics can and should be associated with input data?	At a minimum, metrics based on the relative amount of information contained in an input address should be associated with a record.
When and how can and should the implicit spatial accuracy of input data be calculated?	Implicit spatial accuracy should always be calculated and associated with any input data that are geocoded. Implicit spatial accuracy should be calculated as the area for which highest resolution reference feature can be unambiguously matched.
When and how can and should the implicit level of information of input data be calculated?	The implicit level of information should always be calculated and associated with any input data that are geocoded. The implicit level of information should be calculated based on the types of features it contains (e.g., point, line, polygon); its overall reported accuracy; and any estimates of atomic feature accuracy within the region that are available.

12.2.2 Normalization and feature matching

The effects on accuracy arising from the specificity of the input data also can be seen clearly in both the address normalization and the feature-matching algorithms. First, recall that substitution-based normalization can be considered an example of deterministic feature matching. It performs the same task at different resolutions (i.e., per attribute instead of per feature).

As the normalization algorithm moves through the input string, if the value of the current input token does not match the rule for the corresponding address attribute, the attribute is skipped and the value tried as a match for the next attribute. In the case of feature matching, the relaxation of the attributes essentially removes them from the input data.

Both of these processes can possibly "throw away" data elements as they process an input address, thus lowering the accuracy of the result by implicitly lowering the amount of information encoded by the input description. For example, consider the following three addresses. The first is the real address, but the input data are presented as the second, and the feature-matching algorithm cannot match the input but can match the third. Throwing away the directional element "E" will have prevented the consideration that "E" might have been correct and the street name "Wall" was wrong.

"14 E Mall St"

"14 E Wall St"

"14 Wall St"

12.2.3 Reference datasets

The reference datasets used by a geocoding process contribute to the spatial accuracy of the output in a similar manner. Different representations of reference features necessarily encode different levels of information. This phenomenon results in the very same accuracy effects seen as a consequence of the different levels of input data specification just described. Also, the spatial accuracy of the reference features may be the most important contributing factor to the overall accuracy of the spatial output. Interpolation algorithms operating on the reference features can only work with what they are given, and will never produce any result more accurate than the original reference feature. Granted, these interpolation algorithms can and do produce spatial outputs of varying degrees of spatial accuracy based on their intrinsic characteristics, but the baseline spatial accuracy of the reference feature is translated directly to the output of the interpolation algorithm. The actual spatial accuracy of these reference features can vary quite dramatically. Sometimes, the larger the geographic coverage of a reference dataset, the worse the spatial accuracy of its features. This has historically been observed when comparing street vectors based on TIGER/Line files to those produced by local governments. Likewise, the differences in the spatial accuracies between free reference datasets (e.g., TIGER/Lines) and commercial counterparts (e.g., NAVTEQ) also can be quite striking as discussed in Sections 4.5 and 7. Best practices relating to reference dataset accuracy are listed in Best Practices 33.

Best Practices 33 – Reference dataset accuracy

Policy Decision	Best Practice
When and how can and should atomic feature accuracy be measured and/or estimated?	Estimates of the atomic feature accuracy within a reference dataset should be made periodically by random selection and manual evaluation of the reference features within the region covered by the dataset.

12.2.4 Feature Interpolation

Any time that feature interpolation is performed, one should ask "how accurate is the result" and "how certain can I be of the result?" In the geocoding literature, however, these questions have largely remained unanswered. Without going to the field and physically measuring the difference between the predicted and actual geocode values corresponding to a particular address, it is difficult to obtain a quantitative value for the accuracy of a geocode due to the issues raised in the sections immediately preceding and following this one. However, it may be possible to derive a **relative predicted certainty,** or a relative quantitative measure of the accuracy of a geocode based on information about how a geocode is produced (i.e., attributes of the reference feature used for interpolation), so long as one assumes that the reference feature was selected correctly (e.g., Davis Jr. and Fonseca 2007, Shi 2007). In other words, the relative predicted certainty is the size of the area within which it can be certain that the actual true value for a geocode falls. For instance, if address range interpolation was used, relative predicted certainty would correspond roughly to an oval-shaped area encompassing the street segment. If area unit interpolation was used, the relative predicted certainty would correspond to the area of the feature. Existing research into identifying, calculating, representing, and utilizing these types of certainty measures for geocodes is in its infancy, but will hopefully provide a much richer description of the quality of a geocode and its suitability for use in research studies once it becomes more fully developed.

12.3 MEASURING POSITIONAL ACCURACY

There are several ways to directly derive quantitative values for the positional accuracy of geocoded data, some more costly than others. The most accurate and expensive way is to go into the field with GPS devices and obtain positional readings for the address data to compare with the geocodes. This option requires a great deal of manpower—especially if the size of one's geographic coverage is large—and therefore may not be a feasible option for most organizations. However, this approach may be feasible with special efforts if only a small subset of data need to be analyzed for a small-area analysis.

Barring the ability to use GPS device readings for any large-scale accuracy measurements, other options do exist. The simplest method is to compare the newly produced geocodes with existing geocodes. New geocoding algorithms typically are evaluated and tested as they are developed in this manner. With a fairly small set of representative gold standard data, the spatial accuracy of new geocoding algorithms can be tested quickly to determine their usefulness. The key is investing resources to acquire appropriate sample data.

Another option that is much more common is to use georeferenced aerial imagery to validate geocodes for addresses. Sophisticated mapping technologies for displaying geocodes on top of imagery are now available for low cost, or are even free online (e.g., Google Earth [Google, Inc. 2008a]). These tools allow one to see visually and measure quantitatively how close a geocode is to the actual feature it is supposed to represent. This approach has successfully been used in developing countries to create geocodes for all addresses in entire cities (Davis Jr. 1993). The time required is modest and the scalability appears feasible (e.g., Zimmerman et al. 2007, Goldberg et al. 2008d, and the references within each), although input verification requires a fair amount of cross-validation among those performing the data input. Also, it should be noted that points created using imagery are still subject to some error because the images themselves may not be perfectly georeferenced, and should not be considered equivalent to ground truthing in every case. However, imagery-derived points generally can be considered more accurate than their feature interpolation-based counterparts. Recent research efforts have begun to provide encouraging results as to the possibility of quantifying positional accuracy (e.g., Strickland et al. 2007), but studies focusing on larger geographic areas and larger sample sizes are still needed.

12.4 GEOCODING PROCESS COMPONENT ERROR INTRODUCTION

In cases for which the cost or required infrastructure/staff are too prohibitive to actually quantitatively assess the positional accuracy of geocoded output, other relative scales can and should be used. For example, a measurement that describes the total accuracy as a function of the resulting accuracies of each component of the process is one way to determine an estimate. There currently are no agreed-upon standards for this practice; each registry may have or be developing their own requirements. One proposed breakdown of benchmarking component accuracy is listed in Table 30, which is a good start at collecting and identifying process errors but may not be at a fine enough granularity to make any real judgments about data quality. Also, it should be noted that reference feature and attributes validation is easier for some reference data types (e.g., address points and parcel centroids) and harder for others (e.g., street centerlines).

Table 30 – Proposed relative positional accuracy metrics

Component	Description
1	Original address quality, benchmarked with address validation
2	Reference attribute quality benchmarked with address validation
3	Geocoding match criteria, benchmarked to baseline data
4	Geocoding algorithm, benchmarked against other algorithms

It is unclear at this point how these benchmarks can/should be quantified in such a manner that they may be combined for a total, overall accuracy measure.

Additionally, emerging research is investigating the links between the geographic characteristics of an area and the resulting accuracy of geocodes (e.g., Zimmerman 2006, Zimmerman et al. 2007). Although it has long been known that geocoding error is reduced in urban areas, where shorter street segments reduce interpolation error, the effects of other characteristics such as street pattern design, slope, etc. are unknown and are coming under increasing scrutiny (Goldberg et al. 2008b). The prediction, calculation, and understanding of the sources of geocoding error presented in this section warrant further investigation.

12.5 USES OF POSITIONAL ACCURACY

First and foremost, these quality metrics associated with the spatial values can be included in the spatial analyses performed by researchers to determine the impact of this uncertainty on their results. They also can be used for quality control and data validation. For example, the geocode and accuracy can be used to validate other parts of the patient abstract such as the dxCounty code, and conversely, a county can reveal inaccuracies in a geocoded result. To attempt this, one simply needs to intersect the geocode with a county layer to determine if the county indication is indeed correct. Also, positional accuracy measures can ensure that problems with boundary cases being misclassified can be identified as potential problems before analyses are performed, allowing first-order estimates of the possible uncertainty levels resulting from these data being grouped into erroneous classifications.

12.5.1 Importance of Positional Accuracy

The importance of the positional accuracy in data produced by a geocoding process cannot be understated. Volumes of literature in many disparate research fields are dedicated to describing the potentially detrimental effects of using inaccurate data. For a health-focused review, see Rushton et al. (2006) and the references within. There are, at present, no set standards as to the minimum levels of accuracy that geocodes must have to be suitable in every circumstance, but in many cases common sense can be used to determine their appropriateness for a particular study's needs. When not up to a desired level of accuracy, the researcher may have no choice other than conducting a case review or manually moving cases to desired locations using some form of manual correction (e.g., as shown in Goldberg et al. [2008d]).

Here, a few illustrative examples are provided to demonstrate only the simplest of the problems that can occur, ranging from misclassification of subject data to misassignment of related data. Even though the majority of studies do not report or know the spatial accuracy of their geocoded data or how they were derived, some number usually is reported anyway. This value for spatial accuracy can range from less than 1 meter to several kilometers. The most common problem that inaccurate data can produce is shown in Figure 21. Here, it can be seen that for a geocode lying close to the boundary of two geographic features, the

potential spatial error is large enough that the geocode could in reality be in either one of the larger features.

These boundary cases represent a serious problem. Although the attributes and/or classifications associated from being inside one polygon might be correct, one cannot be sure if the positional accuracy is larger than the distance to the boundary. The associated data could/would be wrong when a parcel resides in two USPS ZIP Codes (on the border) or when the USPS ZIP Code centroid is in the wrong (inaccurate) location. In either of these cases, the wrong USPS ZIP Code data would be associated with the parcel.

Figure 21 – Example of misclassification due to uncertainty (Google, Inc. 2008b)

This problem was documented frequently in the early automatic geocoding literature (e.g., Gatrell 1989, Collins et al. 1998), yet there still is no clear rule for indicating the certainty of a classification via a point-in-polygon association as a direct function of the spatial accuracy of the geocode as well as its proximity to boundaries. Even if metrics describing this phenomenon became commonplace, the spatial statistical analysis methods in common use are not sufficient to handle these conditional associations of attributes. Certain fuzzy-logic operations are capable of operating under these spatially-based conditional associations of attributes, and their introduction to spatial analysis in cancer-related research could prove useful. Zimmerman (2008) provides an excellent review of current spatial statistical methods that can be used in the presence of incompletely or incorrectly geocoded data.

However, it must be noted that some parcels and/or buildings can and do legitimately fall into two or more administrative units (boundary classifications) such as those right along the boundary of multiple regions. The county assignment for taxation purposes, for example, traditionally has been handled by agreements between county assessor's offices in such cases, meaning that spatial analysis alone cannot distinguish the correct attribute classification in all cases.

Because of these shortcomings, some have argued that an address should be linked directly with the polygon ID (e.g., CT) using lookup tables instead of through point-in-polygon calculations (Rushton et al. 2006). When the required lookup tables exist for the polygon reference data of interest this may prove a better option, but when they do not exist the only choice may be a point-in-polygon approach. Best practices related to positional accuracy are listed in Best Practices 34.

It cannot be stressed enough that in all cases, it is important for a researcher utilizing the geocodes to determine if the reported accuracy suits the needs of the study. An example can be found in the study presented earlier (in Section 2.4) investigating whether or not living near a highway and subsequent exposure to asbestos from brake linings and clutch pads has an effect on the likelihood of developing mesothelioma). The distance decay of the particulate matter is on the order of meters, so a dataset of geocodes accurate to the resolution of the city centroids obviously would not suffice. Essentially, the scale of the phenomenon being studied needs to be determined and the appropriate scale of geocodes used. When the data are not at the desired/required level of accuracy, the researchers may have no other choice but to conduct a case review, or manually move cases to desired (correct) locations (more detail on this is covered in Sections 16 and 19).

Best Practices 34 – Positional accuracy

Policy Decision	Best Practice
When and how can and should GPS be used to measure the positional accuracy of geocoded data?	If possible, GPS measurements should be used to obtain the ground truth accuracy of as much geocoded output as possible. Covering large areas may not be a valid option for policy or budgetary regions, but this approach may be feasible for small areas. Metadata should describe: • Time and date of measurement • Type of GPS • Types of any other devices used – Laser distance meters
When and how can and should imagery be used to measure the positional accuracy of geocoded data?	Imagery should be used to ground truth the accuracy of geocoded data if GPS is not an option. Metadata should describe: • Time, date, and method of measurement • Type and source of imagery
When and how can existing geocodes be used to measure the positional accuracy of geocoded data?	If old values for geocodes exist, they should be compared against the newly produced values every time a new one is created. If geocodes are updated or replaced, metadata should describe: • Justification for change • The old value
When and how can and should georeferenced imagery be used to measure the positional accuracy of geocoded data?	If suitable tools exist, and if time, manpower, and budgetary constraints allow for georeferenced imagery-based geocode accuracy measurements, it should be performed on as much data as possible. Metadata should describe the characteristics of the imagery used: • Source • Vintage • Resolution

When and how can and should the positional accuracy metrics associated with geocodes be used in research or analysis?	At a minimum, the FGDC Content Standards for Digital Spatial Metadata (United States Federal Geographic Data Committee 2008a) should be used to describe the quality of the output geocode. The metrics describing the positional accuracy of geocodes should be used whenever analysis or research is performed using any geocodes. Ideally, confidence metrics should be associated with output geocodes and the entire process used to create it including: • Accuracy • Certainty • Reliability Confidence metrics should be utilized in the analysis of geocoded spatial data.
When and how can and should limits on acceptable levels of positional accuracy of data be placed?	There should be no limits placed on the best possible accuracy that can be produced or kept. There should be limits placed on the worst possible accuracy that can be produced or supported, based on the lowest resolution feature that is considered a valid match (e.g.,. USPS ZIP Code centroid, county centroid), and anything of lower resolution should be considered as a geocoding failure. A USPS ZIP Code centroid should be considered the lowest acceptable match.
When and how can and should a geocoded data consumer (e.g., researcher) ensure the accuracy of their geocoded data?	A data consumer should always ensure the quality of their geocodes by requiring specific levels of accuracy before they use it (e.g., for spatial analyses). If the data to be used cannot achieve the required levels of accuracy, they should not be used.

This page is left blank intentionally.

13. REFERENCE DATA QUALITY

This section discusses the detailed issues involved in the spatial and temporal accuracy of reference datasets, while also introducing the concepts of caching and completeness.

13.1 SPATIAL ACCURACY OF REFERENCE DATA

Geographical bias was introduced earlier to describe how the accuracy of reference features may be dependent on where they are located. This phenomenon can clearly be seen in the accuracy reported in rural areas versus those reported in urban areas, due to two factors. First, the linear-based feature interpolation algorithms used are more accurate when applied to shorter street segments than they are when applied to longer ones, and rural areas have a higher percentage of longer streets than do urban areas.

Second, the spatial accuracy of reference features themselves will differ across the entire reference dataset. Again, in rural areas it has been shown that reference datasets are less accurate then their urban counterparts. For example, the TIGER/Line files (United States Census Bureau 2008d) have been shown to have higher spatial accuracy in urban areas with short street segments. Additionally, as previously discussed, different reference datasets for the same area will have different levels of spatial accuracy (e.g., NAVTEQ [NAVTEQ 2008] may be better than TIGER/Lines).

One aspect of these accuracy differences can be seen in the resolution differences depicted in Figure 8. A registry will need to make tradeoffs between the money and time they wish to invest in reference data and the accuracy of the results they require. There currently is no consensus among registries on this topic. Best practices related to reference dataset spatial accuracy problems are listed in Best Practices 35. Note that a distinction needs to be made between those geocoding with a vendor and those geocoding themselves. In the first case, the registry may not have the authority to apply some of these best practices because it may be up to the particular vendor, while in the second they will have that authority. Also, in some instances it may be beneficial for a registry to partner with some other government organization in performing these tasks (e.g., emergency response organizations or U.S. Department of Health and Human Services), or to utilize their work directly.

13.2 ATTRIBUTE ACCURACY

The accuracy of the non-spatial attributes is as important as the spatial accuracy of the reference features. This can clearly be seen in both the feature-matching and feature interpolation components of the process. If the non-spatial attributes are incorrect in the reference dataset such as an incorrect or reversed address range for a street segment, a match may be impossible or an incorrect feature may be chosen during feature matching. Likewise, if the attributes are incorrect the interpolation algorithm may place the resulting geocode in the wrong location, as in the common case of incorrectly defined address ranges. This is covered in more detail with regard to the input address in Section 17.3.

Best Practices 35 – Reference dataset spatial accuracy problems

Policy Decision	Best Practice
When should the feature spatial accuracy, feature completeness, attribute accuracy, or attribute completeness of a reference dataset be improved or the dataset abandoned?	If the spatial accuracy of a reference dataset is sufficiently poor that it is the main contributor to consistently low accuracy geocoding results, improvements or abandonment of a reference dataset should be considered. Simple examples of how to test some of these metrics can be found in Krieger et al. (2001) and Whitsel et al. (2004)
When and how can and should characteristics of the reference dataset be improved?	If the cost of undertaking a reference dataset improvement is less than the cost of obtaining reference data of quality equivalent to the resulting improvement, improvement should be attempted if the time and money available for the task are available.
When and how can and should the feature spatial accuracy of the reference dataset be improved?	If the spatial accuracy of the reference data is consistently leading to output with sufficiently poor spatial accuracy it should be improved, if the time and money available for the task are available. Improvements can be made by: • Manual or automated conflation techniques (e.g., Chen C.C. et al. 2004) • Using imagery (e.g., O'Grady 1999) • Rubber sheeting (see Ward et al. 2005 for additional details)
When and how can and should the attribute accuracy of the reference dataset be improved?	If the attribute accuracy of the reference data is consistently leading to a high proportion of false positive or false negative feature-matching results it should be improved, if the time and money available for the task are available. Improvements can be made by: • Updating the aspatial attributes through joins with data from other sources (e.g., an updated/changed street name list published by a city) • Appending additional attributes (e.g., actual house numbers along a street instead of a simple range).

13.3 TEMPORAL ACCURACY

Temporal accuracy is a measure of how appropriate the time period the reference dataset represents is to the input data that are to be geocoded. This can have a large effect on the outcome of the geocoding process and affects both the spatial and non-spatial attributes. For example, although it is a common conception that the more recently created reference dataset will be the most accurate, this may not always be the case. The geography of the built environment is changing all the time as land is repurposed for different uses; cities expand

their borders; parcels are combined or split; street names change; streets are renumbered; buildings burn and are destroyed or rebuilt; etc. Input address data collected at one point in time most likely represent where the location existed at that particular instant in time. Although these changes may only affect a small number of features, the work to correct them in temporally inaccurate data versions may be time consuming. This results in reference datasets from different periods of time having different characteristics in terms of the accuracy of both the spatial and aspatial data they contain. This could be seen as one argument for maintaining previous versions of reference datasets, although licensing restrictions may prohibit their retention in some cases.

A **temporal extent** is an attribute associated with a piece of data describing a time period for which it existed, or was valid, and is useful for describing reference datasets. Because most people assume that the most recently produced dataset will be the most accurate, the appropriateness of using a dataset from one time period over another usually is not considered during the geocoding process. However, in some cases it may be more appropriate to use the reference data from the point in time when the data were collected to perform the geocoding process, instead of the most recent versions. Several recent studies have attempted to investigate the question of what is the most appropriate reference dataset to use based on its temporal aspect and time period elapsed since input data collection (e.g., Bonner et al. 2003; Kennedy et al. 2003; McElroy et al. 2003; Han et al. 2004, 2005; Rose et al. 2004).

Although the aspatial attributes of historical reference datasets may be representative of the state of the world when the data were collected, the spatial accuracy of newer datasets is typically more accurate because the tools and equipment used to produce them have improved in terms of precision over time. Barring the possibility that a street was actually physically moved between two time periods, as by natural calamity perhaps, the representation in the recent version will usually be more accurate than the older one. In these cases, the spatial attributes of the newer reference datasets can be linked with the aspatial attributes from the historical data. Most cities as well as the U.S. Census Bureau maintain the lineage of their data for this exact purpose, but some skill is required to temporally link the datasets together. The general practice when considering which reference dataset version to use is to progress from the most recent to the least hierarchically. Best practices relating to the temporal accuracy of reference datasets are listed in Best Practices 36.

Best Practices 36 – Reference dataset temporal accuracy

Policy Decision	Best Practice
When and how can and should historical reference datasets be used instead of temporally current versions?	In most cases, a hierarchical approach should be taken from most recent first to oldest. If the region of interest has undergone marked transformation in terms of the construction or demolition of streets, renumbered buildings or renamed streets, or the merging or division of parcels during the time period between when the address was current and the time the address is to be geocoded, the use of historical data should be considered.

13.4 CACHED DATA

One low-cost approach for producing point-based reference datasets is to perform **geocode caching,** which stores the results of previously derived geocodes produced from an interpolation method. Also, in situations for which the running time of a geocoder is a critical issue, this may be an attractive option. The concept of geocode caching also has been termed **empirical geocoding** by Boscoe (2008, pp. 100). Using cached results instead of recomputing them every time may result in substantial performance gains in cases when a lengthy or complex feature interpolation algorithm is used. There is no consensus about the economy of this approach. Different registries may have different practices, and only a few registries currently make use of geocode caching. The most common case and strongest argument for caching is to store the results of interactive geocoding sessions such that the improvements made to aspects of the geocoding process while working on a particular address can be re-leveraged in the future (e.g., creating better street centerline geometry or better address ranges).

Geocode caching essentially creates a snapshot of the current geocoder configuration (i.e., the state of the reference dataset and the world as it was at the publication date and the feature-matching and interpolation algorithms that produce the geocodes). When the reference data and feature interpolation algorithms do not change frequently, geocode caching can be used. If, however, there is the possibility that the resulting geocode may be different every time the geocoder is run (e.g., the case when any of the components of the geocoding process are dynamic or intentionally changed or updated), using the cached data may produce outdated data.

There are potential dangers to using geocode caches in terms of **temporal staleness,** or the phenomenon whereby previously geocoded results stored in a cache become outdated and no longer valid (i.e., low temporal accuracy), with validity being determined on a per-registry basis because there is no consensus. Also, caching data at all may be moot if there is little chance of ever needing to re-process existing address data that already have been geocoded. As new geocoding algorithms are created the cached results produced by older processes may be proven less and less accurate, and at a certain point it may become apparent that these cached results should be discarded and a new version created and stored. In the data-caching literature, there are generally two choices: (1) associate a time to live (TTL) for each cached value upon creation, after which time it is invalidated and removed; or (2) calculate a value for its freshness each time it is interrogated to determine its suitability and remove once it has passed a certain threshold (Bouzeghoub 2004). There are presently no set standards for determining values for either of these, with numerous criteria to be accounted for in the first and complex decay functions possible for the second resulting from the nature of the geocoding process as well as the nature of the ever-changing landscape. General considerations relating to the assignment of TTL and the calculation of freshness are listed in Table 31.

Table 31 – TTL assignment and freshness calculation considerations for cached data

Consideration	Example
TTL and freshness should depend on the source of the geocode	GPS – indefinite TTL, high freshness Manual correction – indefinite, high freshness Geocoded – time varying, medium freshness
TTL and freshness should be based on the match probability	Higher match score – longer TTL, higher freshness
TTL and freshness should be based on the likelihood of geographic change	High-growth area – shorter TTL, lower freshness
TTL and freshness should depend on the update frequency of the reference data	High frequency – shorter TTL, lower freshness
TTL and freshness should correlate with agreement between sources	High agreement – longer TTL, high freshness
Freshness should correlate with time elapsed since geocode creation	Long elapsed time – lower freshness

In all cases where caching is used, a tradeoff exists between the acceptable levels of accuracy present in the old cached results and the cost of potentially having to recreate them. Best practices relating to geocode caching are listed in Best Practices 37.

13.5 COMPLETENESS

Although the accuracy of a reference dataset can be considered a measure of its precision, or how accurate the reference features it contains are, the completeness of a reference dataset can be considered as a measure of recall. In particular, a more complete reference dataset will contain more of the real-world geographic objects for its area of coverage than would a less complete one. Using a more complete reference dataset, one can achieve better results from the geocoding process. Similar to accuracy, levels of completeness vary both between different reference datasets and within a single one. More in-depth discussions of precision and recall are provided in Section 14.2.1.

A distinction should be made between feature and attribute completeness. As recall measures, both refer to the amount of information maintained out of all possible information that could be maintained. The former case refers to a measurement of the amount of features contained in the reference dataset in comparison to all possible features that exist in reality. The latter refers to a measurement of the amount of information (number of attributes) contained per feature out of all information (possible attributes) that could possibly be used to describe it.

Best Practices 37 – Geocode caching

Policy Decision	Best Practice
Should geocode caching be used?	If addresses are to be geocoded more than once, the use of geocode caching should be considered. If geocoding speed is an issue, interpolation methods are too slow, and addresses are geocoded more than once, geocode caching should be used. Geocode results from interactive geocoding sessions should be cached. Metadata should describe all aspects of the geocoding process: • The feature matched • The interpolation algorithm • The reference dataset
When should a geocode cache be invalidated (e.g., when does temporal staleness take effect)?	If reference datasets or interpolating algorithms are changed, the geocode cache should be cleared. Temporal staleness (freshness and/or TTL evaluation) should be calculated for both an entire geocode cache as well as per geocode every time a geocode is to be created. If a TTL has expired, the cache should be invalidated. If freshness has fallen below an acceptable threshold, the cache should be invalidated. If a cache is cleared (replaced), the original data should be archived.

There is no consensus as to how either of these completeness measures should be calculated or evaluated because any measure would require a gold standard to be compared against, resulting in very few cases for which either of these measures are reported with a reference data source. Instead, completeness measurements usually are expressed as comparisons against other datasets. For instance, it is typical to see one company's product touted as "having the most building footprints per unit area" or "the greatest number of attributes," describing feature completeness and attribute completeness, respectively. Using these metrics as anything other than informative comparisons among datasets should be avoided, because if the vendor actually had metrics describing the completeness in quantitative measures, they would surely be provided. Their absence indicates that these values are not known. Some simple, albeit useful quantitative measures that have been proposed are listed in Table 32. Note that a small conceptual problem exists for the third row of the table. TIGER/Line files (United States Census Bureau 2008d) represent continuous features (address ranges), where USPS ZIP+4 (United States Postal Service 2008a) databases represent discreet features, but the street features themselves can be checked in terms of existence and address

range by making some modifications to the original structures (e.g., grouping all addresses in the USPS ZIP+4 per street to determine the ranges and grouping by street name to determine street existence). Best practices relating to the completeness of the reference datasets are listed in Best Practices 38.

Table 32 – Simple completeness measures

Completeness Measure
True reference feature exist/non-existent in reference dataset
True original address exist/non-existent as attribute of feature in reference dataset
Compare one reference dataset to another (e.g., TIGER/Line files [United States Census Bureau 2008d] vs. USPS ZIP+4 [United States Postal Service 2008a])

Best Practices 38 – Reference dataset completeness problems

Policy Decision	Best Practice
When and how can and should the attribute completeness of the reference dataset be improved?	If the attribute completeness of the reference data is consistently leading to a high proportion of false positive or negative feature-matching results it should be improved, if the time and money available for the task are available. Improvements can be made by: • Filling in the missing aspatial attributes through joins with data from other sources (e.g., a street name file from the USPS) • Appending local scale knowledge of alternate names using alias tables.
When and how can and should the feature completeness of the reference dataset be improved?	If the feature completeness of a reference dataset is consistently leading to a high proportion of false negative feature-matching results it should be improved, if the time and money available for the task are available. Improvements can be made by intersecting with other reference datasets that contain the missing features (e.g., a local road layer being incorporated into a highway layer).

This page is left blank intentionally.

14. FEATURE-MATCHING QUALITY METRICS

This section describes the different types of possible matches and their resulting levels of accuracy, and develops alternative match rates.

14.1 MATCH TYPES

The result of the feature-matching algorithm represents a critical part of the quality of the resulting geocode. Many factors complicate the feature-matching process and result in different match types being achievable. In particular, the normalization and standardization processes are critical for preparing the input data. If these algorithms do a poor job of converting the input to a form and format consistent with that of the reference dataset, it will be very difficult, if not impossible, for the feature-matching algorithm to produce a successful result.

However, even when these processes are applied well and the input data and reference datasets both share a common format, there still are several potential pitfalls. These difficulties are exemplified by investigating the domain of possible outputs from the feature-matching process. These are listed in Table 33, which shows the descriptions and causes of each. An input address can have no corresponding feature in the reference dataset (i.e., the "no match" case), or it can have one or more. These matches can be perfect, meaning that every attribute is exactly the same between the input address and the reference feature, or non-perfect, meaning that some of the attributes do not match between the two. Examples that would result in some of these are depicted in Figure 22. Note that in most cases, an ambiguous perfect match indicates either an error in the reference dataset (E-911 is working toward getting rid of these), or incompletely defined input data matching multiple reference features.

Once a single feature (or multiple features) is successfully retrieved from the reference set by changing the SQL and re-querying if necessary (i.e., attribute relaxation), the feature-matching algorithm must determine the suitability of each of the features selected through the use of some measures. The real power of a feature-matching algorithm therefore is two-fold: (1) it first must be able to realize that no match has been returned, and then (2) subsequently automatically alter and regenerate the SQL to attempt another search for matching features using a different set of criteria. Thus, one defining characteristic distinguishing different feature-matching algorithms is how this task of generating alternate SQL representations to query the reference data is performed. Another is the measures used to determine the suitability of the selected features.

Table 33 – Possible matching outcomes with descriptions and causes

Outcome	Description	Cause	Code
Perfect match	A single feature in the reference dataset could be matched to the input datum, and both share every attribute.	The combination of input attributes exactly matches those of a single reference feature.	P
Non-perfect match	A single feature in the reference dataset could be matched to the input datum, and both share some but not all attributes.	At least one, but not all, of the combinations of input attributes exactly match those of a single reference feature.	Np
Ambiguous perfect match	Multiple features in the reference dataset could be matched to the input datum, and each shares every attribute.	The combination of input attributes exactly matches those of multiple reference features.	Ap
Ambiguous non-perfect match	Multiple features in the reference dataset could be matched to the input datum, and each shares some but not all attributes.	At least one, but not all, of the combinations of input attributes exactly matches those of multiple reference features.	Anp
No match	No features in the reference dataset could be matched to the input datum.	The combination of input attributes is not found in the reference dataset.	N

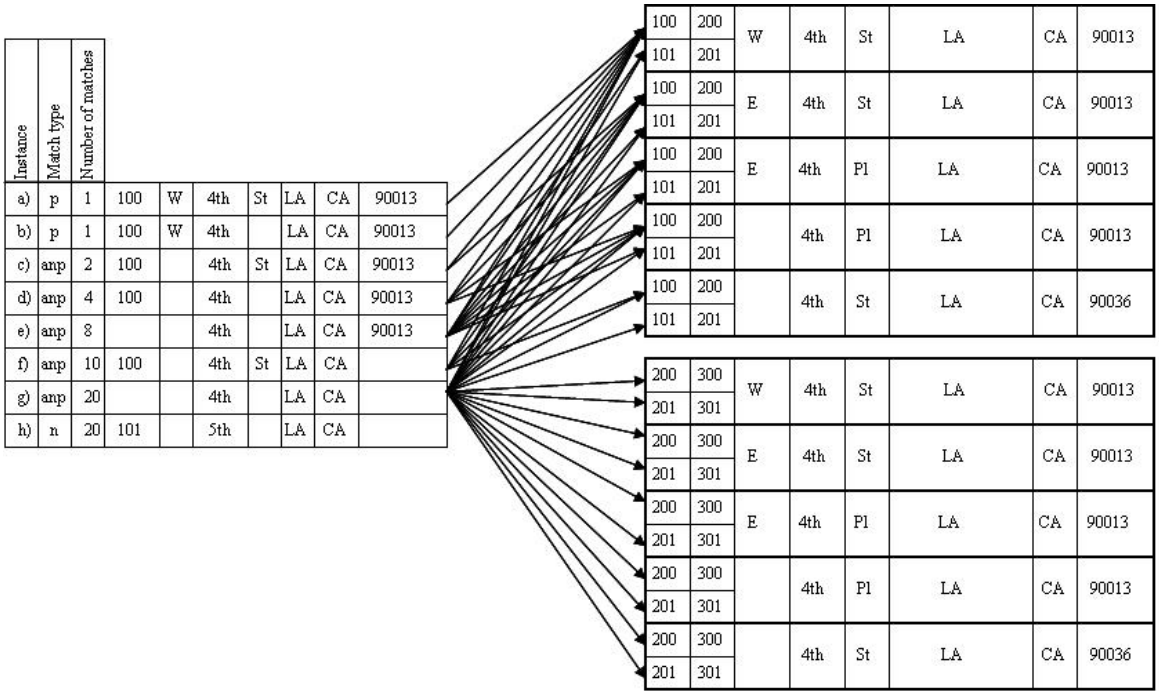


Figure 22 – Examples of different match types

Much like the address normalization process, there are both simplistic and complex ways to achieve this, and each has its particular strengths and weaknesses and is suitable under certain conditions. This process of feature matching is tightly related to the computer science field of record linkage. Many fundamental research questions and concepts developed therein have been applied to this task of feature matching. Best practices related to feature match types are listed in Best Practices 39.

Best Practices 39 – Feature match types

Policy Decision	Best Practice
Which match types should be considered acceptable?	Perfect and non-perfect non-ambiguous matches should be considered acceptable.
Which match types should considered unacceptable?	Ambiguous matches should not be considered acceptable.
What should be done with unacceptable matches?	Non-acceptable matches should be reviewed, corrected, and re-processed.
What metadata should be maintained?	Metadata should describe the reason why an action was taken (e.g., the match type) and what action was taken.
What actions can and should be taken to correct unacceptable matches?	At a minimum, manual review/correction and attribute relaxation should be attempted.

14.2 MEASURING GEOCODING MATCH SUCCESS RATES

Match rates can be used to describe the completeness of the reference data with regard to how much of the input data they contain, assuming that all input data are valid and should rightfully exist within them. Match rates also can be used to test the quality of the input data in the reverse case (i.e., when the reference data are assumed to be complete and unmatchable input data are assumed to be incorrect).

> **Under no circumstances should a high match rate be understood as equivalent to a high accuracy rate; the two terms mean fundamentally different things.**

A geocoder resulting in a 100 percent match rate should not be considered accurate if all of the matches are to the city or county centroid level.

14.2.1 Precision and recall

Precision and recall metrics are often used to determine the quality of an information retrieval (IR) strategy. This measurement strategy breaks the possible results from a retrieval algorithm into two sets of data: (1) one containing the set of data that should have correctly been selected and returned by the algorithm, and (2) another containing a set of data that is actually selected and returned by an algorithm, with the latter one causing the problem (Raghavan et al. 1989). The set of data that was actually returned may contain some data that should not have been returned (i.e., incorrect data), and it may be missing some data that should have been returned. In typical IR parlance, **recall** is a measure that indicates how

much of the data that should have been obtained actually was obtained. **Precision** is a measure of the retrieved data's correctness.

14.2.2 Simplistic match rates

In the geocoding literature, the related term **match rate** often is used to indicate the percentage of input data that were able to be assigned to a reference feature. Although this is related to the recall metric, the two are not exactly equivalent. The match rate, as typically defined in Equation 5, does not capture the notion of the number of records that should have been matched. The match rate usually is defined as the number of matched records (i.e., records from the input data that were successfully linked to a reference feature) divided by the total number of input records:

$$\frac{\#\ \text{Matched Records}}{\#\ \text{All Records}}$$

Equation 5 – Simplistic match rate

This version of match rate calculation corresponds to Figure 23 (a), in which the match rate would be the differences between the areas of records attempted and records matched.

14.2.3 More representative match rates

It may be of more interest to qualify the denominator of this match rate equation in some way to make it closer to a true recall measure, eliminating some of the false negatives. To do this, one needs to determine a more representative number of records with addresses that should have matched. For example, if a geocoding process is limited in using a local-scale reference dataset with limited geographic coverage, input data corresponding to areas that are outside of this coverage will not be matchable. If they are included in the match rate, they are essentially false negatives; they should have been simply excluded from the calculation instead. It might therefore be reasonable to define the match rate by subtracting these records with addresses that are out of the area from the total number of input records:

$$\frac{\#\ \text{Matched Records}}{\#\ \text{All Records} - \#\ \text{Records out of state, county, etc.}}$$

Equation 6 – Advanced match rate

This match rate calculation corresponds to Figure 23(b), in which the match rate would be the differences between the areas of records matched and the records within the coverage area, not simply just records attempted, resulting in a more representative, higher match rate. Registries need to use caution when utilizing this because using these other attributes (e.g., county, USPS ZIP Code) for determining what should have been rightfully included or excluded for geocodability within an area also is subject to error if those attributes themselves are erroneous as well.

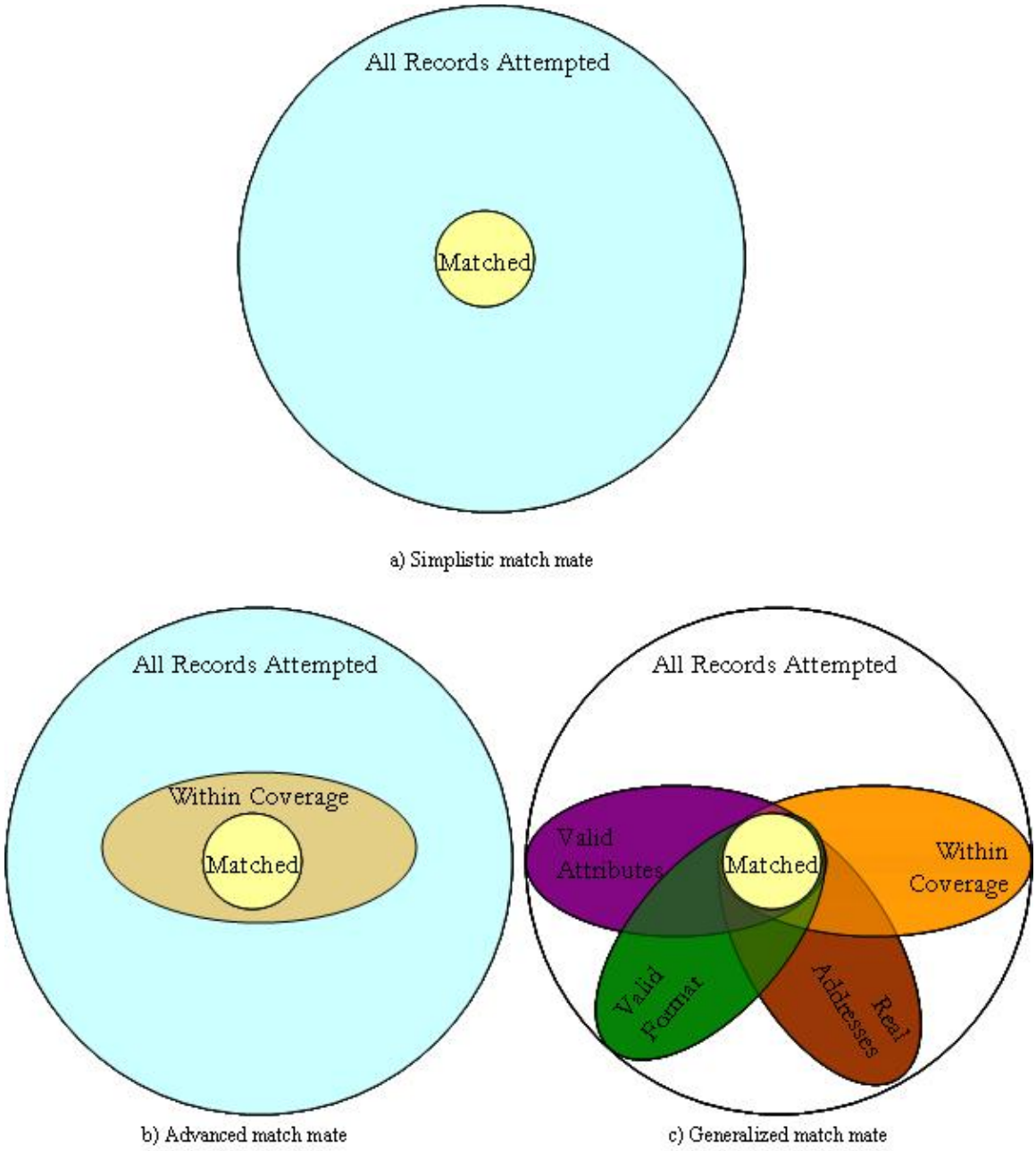


Figure 23 – Match rate diagrams

14.2.4 A generalized match rate

This approach can be generalized even further. There are several categories of data that will not be possible to be matched by the feature-matching algorithm. For instance, data that are outside of the area of coverage of the reference dataset, as in the last example, will posses this property. Input data that are in a format not supported by the geocoding process will as well.

For example, if the geocoding process does not support input in the form of named places, intersections, or relative directions, input in any one of these forms will never be able

to be successfully matched to a reference feature and will be considered "invalid" by this geocoder (i.e., the input data may be fine but the reference data do not contain a match). Finally, a third category comprises data that are simply garbage, and will never be matched to a reference feature simply because they do not describe a real location. This type of data is most typically seen because of data entry errors, when the wrong data have been entered into the wrong field upon entry (e.g., a person's birth date being entered as his or her address). One can take this representative class of invalid data, or data that are impossible to match (or impossible to match without additional research into the patient's usual residence address at the time of diagnosis), into account in the determination of a match rate as follows:

$$\frac{\#\ \text{Matched Records}}{\#\ \text{All Records} - \#\ \text{Records impossible to match}}$$

Equation 7 – Generalized match rate

This match rate calculation corresponds to Figure 23(c), in which the match rate is no longer based on the total number of record attempted; instead, it only includes records that should have been matchable, based on a set of criteria applied. In this case, the set of addresses that should have successfully matched is made up by the area resulting from the union of each of the areas. The match rate then is the difference between this area and the area of records that matched, resulting in an even more representative, higher match rate.

14.2.5 Non-match classification

The difficult part of obtaining a match rate using either of the two latter equations (Equation 6 or Equation 7) is classifying the reason why a match was not obtainable for input data that cannot be matched. If one were processing tens of thousands of records of input data in batch and 10 percent resulted in no matches, it might be too difficult and time-consuming to go through each one and assign an explanation.

Classifying input data into general categories such as valid or invalid input format should be fairly straightforward. This could be accomplished for address input data simply by modifying the address normalization algorithm to return a binary true/false along with its output indicating whether or not it was able to normalize into a valid address. One could also use the lower-resolution attributes (e.g., USPS ZIP Code) to get a general geographic area to compare with the coverage of the reference dataset for classification as inside or outside the coverage area of the reference dataset. Although not exactly precise, these two options could produce first-order estimates for the respective number of non-matches that fall into each category and could be used to derive more representative values for match rates given the reference dataset constraints of a particular geocoding process.

14.3 ACCEPTABLE MATCH RATES

An **acceptable match rate** is a specific match rate value that a geocoding process must meet such that the geocoded data can be considered valid for use in a research study. What constitutes an acceptable match rate is a complex subject and includes many factors such as what type of feature matched to or the particular linkage criteria used at a registry. Further, it needs to be stated that the overall match rate comes from both the input data and the reference data, which together constrain the total value. Other than early reports by Ratcliffe (2004), an exhaustive search at the time of this writing found no other published work investigating this topic. There is a wealth of literature on both the selection bias resulting from

match rates as well as how these rates may effectively change between geographies (c.f., Oliver et al. 2005, and the references within), but a qualitative value for an acceptable match rate for cancer-related research has not been proposed. It is not possible to recommend using the percentages Ratcliffe (2004) defined, as they are derived from and meant to be applied to a different domain (crime instead of health), but it would be interesting to repeat his experiments in the health domain to determine if health and crime have similar cutoffs. Further research into using the more advanced match rates just described would be useful. What can be stated generally is that an acceptable match rate will vary by study, with the primary factor being the level of geographic aggregation that is taking place. Researchers will need to think carefully if the match rates they have achieved allow their geocoded data to safely be used for drawing valid conclusions. Each particular study will need determine if the qualities of their geocodable versus non-geocodable data may indicative of bias in demographic or tumor characteristics, from which they should draw conclusions on the suitability and representativeness of their data (Oliver et al. 2005).

14.4 MATCH RATE RESOLUTION

The discussion thus far has developed a measure of a **holistic-level match rate,** which is a match rate for the entire address as a single component. An alternative to this is to use an **atomic-level match rate,** which is a match rate associated with each individual attribute that together composes the address. This type of measure relates far more information about the overall match rate because it defines it at a higher resolution (i.e., the individual attribute level as opposed to the whole address level). Essentially, this extends the concept of match rate beyond an overall percentage for the dataset as a whole to the level of per-each-geocoded-result.

To achieve this type of match rate resolution, ultimately all that is required is documentation of the process of geocoding. If each process applied, from normalization and standardization to attribute relaxation, recorded or reported the decisions that were made as it processed a particular input datum along with the result it produced, this per-feature match rate could be obtained and an evaluation of the type of address problems in one's input records could be conducted.

For instance, if probabilistic feature matching was performed, what was the uncertainty cutoff used, and what were the weights for each attribute that contributed to the composite weight? If deterministic feature matching was used, which attributes matched and which ones were relaxed and to what extent? This type of per-feature match rate is typically not reported with the output geocode when using commercial geocoding software, as many of the details of the geocoding process used are hidden "under the hood," although it is part of the feature-matching process. However, codes pertaining to the general match process are generally available. Best practices related to success rates (match rates) are listed in Best Practices 40.

Best Practices 40 – Success (match) rates

Policy Decision	Best Practice
Which metrics should be used to describe the success rate of feature-matching algorithms?	At a minimum, feature-matching success should be described in terms of match rates.
How should match rates be calculated?	At a minimum, match rates should be computed using the simplistic match rate formula. If constraints permit, more advanced match rates should be calculated using the other equations (e.g., the advanced and generalized match rate formulas). Metadata should describe the type of match rate calculated and variables used along with how they were calculated.
How should an advanced match rate be calculated?	First-order estimates for the number of input addresses outside the coverage area for the current set of reference datasets should be calculated using lower resolution reference datasets (e.g., USPS ZIP Code reference files). This number should be subtracted from the set of possible matches before doing the match rate calculation. The metadata should describe the lower resolution reference dataset used for the calculation.
How should a generalized match rate be calculated?	If the normalization algorithm can output an indication of why it failed, this should be used for classification, and the resulting classification used to derive counts. This number should be subtracted from the set of possible matches before doing the match rate calculation.
At what resolution and for what components can and should match rates be reported?	Match rates should be reported for all aspects of the geocoding process, at both the holistic and atomic levels.
How can atomic-level match rates be calculated?	If the geocoding process is completely transparent, information about the choices made and output of each component of the geocoding process can be measured and combined to calculate atomic-level match rates.

15. NAACCR GIS COORDINATE QUALITY CODES

This section introduces the NAACCR GIS Coordinate Quality Codes and discusses their strengths and weaknesses.

15.1 NAACCR GIS COORDINATE QUALITY CODES DEFINED

For geocoding output data to be useful to consumers, metadata describing the quality associated with them are needed. To this end, NAACCR has developed a set of **GIS Coordinate Quality Codes** (Hofferkamp and Havener 2008, p. 162) that indicate at a high level the type of data represented by a geocode. It is crucial that these quality codes be associated with every geocode produced at any time by any registry.

> **Without such baseline codes associated with geocodes, researchers will have no idea how good the data they run their studies on are because it will depend on the study size, resolution, etc.—without requiring the need for follow-up contact with the data provider or performing the geocoding themselves—and therefore the researchers will have no clue as to how representative their results are.**

Abbreviated versions of these codes are listed in Table 34, and correspond roughly to the hierarchy presented earlier. For exact codes and definitions, refer to Data Item #366 of *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008, p. 162).

Table 34 – NAACCR recommended GIS Coordinate Quality Codes (paraphrased)

Code	Description
1	GPS
2	Parcel centroid
3	Match to a complete street address
4	Street intersection
5	Mid-point on street segment
6	USPS ZIP+4 centroid
7	USPS ZIP+2 centroid
8	Assigned manually without data linkage
9	5-digit USPS ZIP Code centroid
10	USPS ZIP Code centroid of Post Office Box or Rural Route
11	City centroid
12	County centroid
98	Coordinate quality is unknown
99	Geocoding was attempted but unable or unwilling to assign coordinates

Likewise, researchers should refrain from using any data that do not have accuracy metrics like the codes in the previous table, and they should insist that these be reported in

geocoded data they obtain. It is up to the researcher to decide whether or not to use geocodes with varying degrees of reported quality, but it should be clear that incorporating data without quality metrics can and should lower the confidence that anyone can have in the results produced. Further, the scientific community at large should require that research undergoing peer review for possible scientific publication indicate the lineage and accuracy metrics for the data used as a basis for the studies presented, or at least note its absence as a limitation of the study.

There are three points to note about the present NAACCR GIS Coordinate Quality Codes and other similar schemes for ranking geocodes (e.g., SEER census tract certainty [Goldberg et al. 2008c]). The first is that its code 98--coordinate quality unknown--is effectively the same as having no coordinate quality at all. Therefore, utilization of this code should be avoided as much as possible because it essentially endorses producing geocodes without knowing anything about coordinate quality.

Second, the codes listed in this table are exactly what they indicate that they are, qualitative codes describing characteristics of the geocodes. No quantitative values can be derived from them and no calculations can be based upon them to determine such things as direction or magnitude of the true error associated with a geocode. Thus, they serve little function other than to group geocodes into classes that are (rightfully or wrongfully) used to determine their suitability for a particular purpose or research study.

Finally, the current standard states "Codes are hierarchical, with lower numbers having priority" (Hofferkamp and Havener 2008, p. 162). When taken literally, the standard only discusses the priority that should be given to one geocode over another, not the actual accuracy of geocodes; however, it nonetheless has ramifications on the geocoding process because geocoding developers may use this to guide their work. Without specifically stating it, this table can be seen in one light to imply a hierarchical accuracy scheme, with lower values (e.g., 1) indicating a geocode of higher accuracy and higher values (e.g., 12) indicating a geocode of lower accuracy.

Unfortunately, this may not be correct in all cases and geocoding software developers and users need to be aware that the choice of which is the "best" geocode to choose/output should not be determined from the ranks in this table alone. Currently however, most commercial geocoders do in fact make use of hierarchies such as this in the rules that determine the order of geocodes to attempt, which may not be as good as human intervention, and is definitely incorrect in some cases. For instance, out of approximately 900,000 street segments in California that have both a ZCTA and place designation in the TIGER/Line files (where both the left and right side values are the same for the ZCTA and Place) (United States Census Bureau 2008d), approximately 300,000 street segments have corresponding ZCTA areas that are larger than the corresponding place areas for the same segment. Recall that matching to feature with a smaller area and calculating its centroid is more likely to result in a geocode with greater accuracy. Taken together, it is clear that in the cases for when a postal address fails to match and a matching algorithm relaxes to try the next feature type in the implied hierarchy, one-third of the time, choosing the USPS ZIP Code is the wrong choice (ignoring the fact that ZCTA and USPS ZIP Codes are not the same—see Section 5.1.4 for details). Goldberg et al. (2008c) can be consulted for further discussion of this topic.

It should be clear that although the GIS coordinate quality codes such as those advocated by NAACCR are good first steps toward geocoding accountability, there is still much work to be done before they truly represent qualitative values about the geocodes that they describe. Abe and Stinchcomb (2008, p. 124) clearly articulate the need for "geocoding

software [to] automatically record a quantitative estimate of the positional accuracy of each geocode based on the size and spatial resolution of the matched data source, [which] could be used to provide a positional 'confidence interval' to guide the selection of geocoded records for individual spatial analysis research projects." Best practices related to GIS Coordinate Quality Codes are listed in Best Practices 41.

Best Practices 41 – GIS Coordinate Quality Codes

Policy Decision	Best Practice
When and which GIS coordinate quality codes should be used?	At a minimum, the NAACCR GIS Coordinate Quality Codes specified in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008, p. 162) should always be associated with any geocoded output. Geocode qualities of less than full street address (code 3) should be candidates for manual review.
When and how can and should NAACCR GIS Coordinate Quality Codes be assigned?	NAACCR GIS Coordinate Quality Codes should always be assigned in the same manner, based on the type of reference feature matched and the type of feature interpolation performed.
What other metadata can and should be reported?	If possible, metadata about every decision made by the geocoding process should be reported along with the results (and stored outside of the present NAACCR record layout).
Should any geocodes without NAACCR GIS Coordinate Quality Codes be used for research?	Ideally, any geocodes without NAACCR GIS Coordinate Quality Codes should not be used for research. If geocodes without NAACCR GIS Coordinate Quality Codes must be used, this should be stated as a limitation of the study.

This page is left blank intentionally.

Part 4: Common Geocoding Problems

Throughout this document, potential problems regarding the geocoding process have been discussed as each component has been introduced. This part of the document will list specific problems and pitfalls that are commonly encountered, and provide advice on the best and recommended ways to overcome them. In all cases, the action(s) taken should be documented in metadata that accompany the resulting geocode, and the original data should be maintained for historical lineage.

This page is left blank intentionally.

16. QUALITY ASSURANCE/QUALITY CONTROL

This section provides insight into possible methods for overcoming problems that may encountered in the geocoding process.

16.1 FAILURES AND QUALITIES

As discussed throughout the text of this document, an address may fail to geocode to an acceptable level of accuracy (including not geocoding at all) for any number of reasons including errors within the address itself, errors in the reference dataset, and/or the uncertainty of a particular interpolation algorithm. In Table 35, classes of problems from the previous sections have been listed along with example cases or reasons why they would have occurred for the input address that should be "3620 S. Vermont Ave, Los Angeles, CA 90089." These classifications will be used in the following sections to enumerate the possible options and describe the recommended practice for each type of case. Note that each registry may have its own regulations that determine the protocol of action regarding how certain classes of problems are handled, so some of the recommended solutions may not be applicable universally. In addition to these processing errors, there are also acceptable "quality" levels that may be required at a registry. The current standard of reporting to which vendors are currently held responsible are found within the NAACCR GIS Coordinate Quality Codes (Hofferkamp and Havener 2008, p. 162) as listed in Table 34. Although the shortcomings with these codes have been listed in Section 15.1, these will be used to guide the recommended decisions and practices. The items in these tables are by no means exhaustive; registries may face many more that are not listed. For these cases, the sections in the remainder of this section provide the details as to why a particular option is recommended with the hopes of using similar logic in determine the appropriate action in the appropriate circumstance(s).

Table 35 – Classes of geocoding failures with examples for true address 3620 S. Vermont Ave, Los Angeles CA 90089

Index	Geocoded	Problem	Example
1	No	Failed to geocode because the input data are incorrect.	3620 S Verment St, Los Angeles, CA 90089
2	No	Failed to geocode because the input data are incomplete.	3620 Vermont St, Los Angeles, CA 90089
3	No	Failed to geocode because the reference data are incorrect.	Address range for 3600-3700 segment in reference data is listed as 3650-3700
4	No	Failed to geocode because the reference data are incomplete.	Street segment does not exist in reference data
5	No	Failed to geocode because the reference data are temporally incompatible.	Street segment name has not been updated in the reference data
6	No	Failed to geocode because of combination of one or more of 1-5.	3620 Vermont St, Los Angeles CA 90089, where the reference data has not been updated to a include the 3600-3700 address range for segment
7	Yes	Geocoded to incorrect location because the input data are incorrect.	3620 S Verment St, Los Angeles, CA 90089 was (incorrectly) relaxed and matched to 3620 aaa Ferment St, Los Angeles, CA 90089
8	Yes	Geocoded to incorrect location because the input data are incomplete.	3620 Vermont St, Los Angeles, CA 90089 was arbitrarily (incorrectly) assigned to 3620 N a Vermont St, Los Angeles, CA 90089
9	Yes	Geocoded to incorrect location because the reference data are incorrect.	The address range for 3600-3700 is reversed to 3700-3600
10	Yes	Geocoded to incorrect location because the reference data are incomplete.	Street segment geometry is generalized straight line when the real street is extremely curvy
11	Yes	Geocoded to incorrect location because of interpolation error.	Interpolation (incorrectly) assumes equal distribution of properties along street segment
12	Yes	Geocoded to incorrect location because of dropback error.	Dropback placement (incorrectly) assumes a constant distance and direction
13	Yes	Geocoded to incorrect location because of combination of one or more of 7-12.	The address range for 3600-3700 is reversed to 3700-3600, and dropback of length 0 is used

Table 36 – Quality decisions with examples and rationale

Decision	Practice	Rationale
When only a USPS PO box is available, yet a USPS ZIP+4 is correct; should the geocoded address be based on the USPS ZIP+4 centroid or the USPS ZIP+5 centroid?	The address should be geocoded to the USPS ZIP+4.	The USPS ZIP+5 will be based on the USPS PO box address, which is less accurate than the USPS ZIP+4 based on the address.
When only an intersection is available, should the centroid of the intersection or the centroid of one of the properties on the corners be used?	The centroid of the one of the corner properties should the used.	This increases the likelihood of that the geocode is on the correct location from 0 (the intersection centroid will never be correct), to 1/number of corners.
If the location of the address is known, should the geocode be manually moved to it (e.g., manually dragged using a map interface)?	The geocode should be moved if the location is known.	A known location should be used over a calculated one.
If the location of the building for an address is known, should the geocode be manually moved to its centroid?	The geocode should be moved if the building is known.	A known location should be used over a calculated one.
If only a named place is available as an address, should research be performed to determine an address or should the next lower resolution attribute be used (e.g., city name)	Research for the address of a named place should be attempted before moving to the next lower resolution attribute.	The address information may be trivially available, and it will dramatically improve the resulting geocode.
If the geocode is less accurate than USPS ZIP Code centroid (GIS Coordinate Quality Code 10), should it be reviewed for manual correction?	Geocodes with accuracy less than GIS Coordinate Quality Code 10 should be reviewed for manual correction.	After USPS ZIP Code level certainty, the appropriateness of using a geocode in all but large area aggregation studies diminishes rapidly.
Should manual steps be taken in order to get a geocode for every record?	Manual processing should be attempted to get a geocode for every record.	Patients should not be excluded from research studies because their address was not able to be geocoded.

Decision	Practice	Rationale
If a street segment can be matched, but the address cannot, should the center point of the segment or the centroid of the minimum bounding rectangle (MBR) encompassing the segment be used?	The centroid of the MBR should be used.	In the case where a street is straight, the centroid of the MBR would be the center point of the street. In the case of a curvy street, using the centroid minimizes the possible error from any other point on the street.
If two connected street segments are ambiguously matched, should their intersection point or the centroid of the MBR encompassing them be used?	The centroid of the MBR should be used.	In the case where the two streets are straight, the centroid of their MBR would be the intersection point between them (assuming their lengths are similar and the angle between them is 180 degrees). In the case of two curvy streets, the angle between them being sharp, or the lengths being dramatically different, using the centroid minimizes the possible error from any other point on the two streets.
If two disconnected street segments are ambiguously matched, should the centroid of the MBR encompassing them be used?	The centroid of the MBR should be used.	The centroid of their MBR minimizes the possible error from any other point on the two streets.
If an address geocodes different now than it has in the past, should all records with that geocode be updated?	All records should be updated to the new geocode if it is more accurate.	Research studies should use the most accurate geocode available for a record.

17. ADDRESS DATA PROBLEMS

This section introduces various types of problems at registries that occur with address data (e.g., dxAddress, dxCity, dxZIP, dxState), including lifecycle and formatting problems.

17.1 ADDRESS DATA PROBLEMS DEFINED

Details regarding the exact issues related to a selected set of representative postal addresses are presented next to illustrate the ambiguities that are introduced as one iteratively removes attributes. The best-possible-case scenario is presented first. Best practices relating to the management of common addressing problems are listed in Best Practices 42.

Best Practices 42 – Common address problem management

Policy Decision	Best Practice
What types of lists of common input address problems and solutions should be maintained?	Lists of problems that are both common (occur more than once) and uncommon with recommended solutions should be maintained and consulted when problems occur. Examples of common problems include: • 15% error in dxCounty

17.2 THE GOLD STANDARD OF POSTAL ADDRESSES

The following address example represents the gold standard in postal address data. It contains valid information in each of the possible attribute fields and indicates enough information to produce a geocode down to the sub-parcel unit or the floor level.

"3620 ½ South Vermont Avenue East, Unit 444, Los Angeles, CA, 90089-0255"

In the geographic scale progression used during the feature-matching algorithm, a search for this address is first confined by a state, then by a city, then by a detailed USPS ZIP Code to limit the amount of possible candidate features to within an area. Next, street name ambiguity is removed by the prefix and suffix directionals associated with the name, "South" and "East," respectively, as well as the street type indication, "Avenue." Parcel identification then becomes attainable through the use of the street number, "3620," assuming that a parcel reference dataset exists and is accessible to the feature-matching algorithm. Next, a 3-D geocode can finally be produced from the sub-parcel identification by combining the unit indicators, "½" and "Unit 444" to determine the floor and unit on the floor, assuming that this is an apartment building and that a 3-D building model is available to the feature-matching algorithm. Note that both "½" and "444" can mean different things in different localities (e.g., they can both refer to subdivided parcels, subdivisions within a parcel, or even lots in a trailer park).

This example illustrates the best-possible-case scenario in terms of postal address specification and reference dataset availability, and is for most registries, rarely encountered. This

is because reference datasets of this quality do not exist for many large regions, details such as the floor plan within a building are seldom needed, and input data are hardly ever specified for this completely. It often is assumed that utilization of the USPS ZIP+4 database will provide the gold standard reference dataset, but it actually is only the most up-to-date source for address validation alone and must be used in conjunction with other sources to obtain the spatial aspect of an output geocode, which may be subject to some error. The practice of transforming an incompletely described address into a gold standard address (completely described) is performed by most commercial geocoders, as evidenced by the inclusion of the full attributes of the matched feature generally included with the geocode result. Best practices relating to gold standard addresses are listed in Best Practices 43.

Best Practices 43 – Creating gold standard addresses

Policy Decision	Best Practice
Should non-gold standard addresses have information added or removed to make them "gold standard"?	In the case where legitimate attributes of an address are missing and can be non-ambiguously identified, they should be added to the address. Metadata should include: • Which attributes were added • Which sources were used

17.3 ATTRIBUTE COMPLETENESS

This following depiction of standard address data is far more commonly encountered than the gold standard address:

"3620 Vermont Avenue, Los Angeles, CA, 90089"

Here, the street directional, sub-parcel, and additional USPS ZIP Code components of the address have been removed. A feature-matching algorithm processing this case could again fairly quickly limit its search for matching reference features to within the USPS ZIP Code as in the last example, but from that point, problems may arise due to **address ambiguity,** the case when a single input address can match to more than one reference feature, usually indicative of an incompletely described input address. This can occur at multiple levels of geographic resolution for numerous reasons.

This last address shows the case of **street segment ambiguity,** where multiple street segments all could be chosen as the reference feature for interpolation based on the information available in the input address. First, multiple streets within the same USPS ZIP Code can, and routinely do, have the same name, differing only in the directional information associated with them indicating which side of a city they are on. Further, the address range information commonly associated with street reference features that are used to distinguish them, which will be covered in more detail later, often is repeated for these streets (e.g., 3600-3700 South Vermont, 3600-3700 North Vermont, and 3600-3700 Vermont). Thus, the feature-matching algorithm may be presented with multiple options capable of satisfying the input address.

Moving to a finer scale, **street address ambiguity** is the case when a single input address can match to more than one reference address on a single street segment as in the case

where a correct street segment can unambiguously be determined, but a specific location along the street cannot because the address number is missing:

"South Vermont Avenue East, Los Angeles, CA, 90089"

At a still finer scale, **sub-parcel address ambiguity** is the case when a single input address can match to more than one reference feature that is contained within the same parcel of land. This problem often arises for large complexes of buildings such as Co-op City in Bronx, NY, or as in the following example of the Cardinal Gardens residence buildings on the USC campus, all sharing the same postal street address:

"3131 S. McClintock Avenue, Los Angeles, CA, 90007"

In these ambiguous cases, most feature-matching algorithms alone do not contain enough knowledge to be able to pick the correct one. A detailed analysis of the different methods for dealing with these cases is presented in Section 18.

17.4 ATTRIBUTE CORRECTNESS

"831 North Nash Street East, Los Angeles, CA, 90245"

This case exemplifies the beginning of a "slippery slope," the correctness of address attributes. This example lists the USPS ZIP Code "90245" as being within the city "Los Angeles." In this particular case, this association is incorrect. The city "Los Angeles" does not contain the USPS ZIP Code "90245", which may at first be considered to be a typographical error in the USPS ZIP Code. However, the USPS ZIP Code is in reality correct, but it is part of an independent city, "El Segundo," which is within Los Angeles County. Therefore, one of these attributes is indeed wrong and should be ignored and not considered during the feature selection process, or better yet, corrected and replaced with the appropriate value.

There are many reasons why these types of errors can and do occur. For instance, people sometimes refer to the city or locality in which they live by the name of their neighborhood, instead of the city's official political name or their post office name. As neighborhood names are often only locally known, they are often not included in national-scale reference datasets, and therefore are not applicable and can appear to be incorrect. In Los Angeles, one obvious example is "Korea Town," an area several miles in size slightly southwest of downtown LA that most residents of the city would recognize by name immediately, but would not be found as an official name in the TIGER/Line files. Also, the reverse is possible as in the previous El Segundo address example. People may mistakenly use the name "Los Angeles" instead of the valid city name "El Segundo," because they lack the local knowledge and assume that because the location is part of the "Los Angeles Metropolitan Area," "Los Angeles" is the correct name to use.

> **This disconnect between local-level knowledge possessed by the people creating the data (e.g., the patient describing it or the hospital staff recording it) and the non-local-level knowledge possessed by the sources creating the reference datasets presents a persistent difficulty in the geocoding process.**

Similarly, USPS ZIP Codes and ZCTAs are maintained by separate organizations that do not necessarily share all updates with each other, resulting in the possibility that the data may not be the consistent with each other. As a result, it is often the case that the address data input is referring to the USPS ZIP Code, while the reference data source may be using the ZCTA (e.g., in TIGER/Line files).

Finally, USPS ZIP Code routes have a dynamic nature, changing over time for the purpose of maintaining efficient mail delivery, therefore the temporal accuracy of the reference data may be an issue. USPS ZIP Codes may be added, discontinued, merged, or split, and the boundaries for the geographic regions they are assumed to represent may no longer be valid. Thus, older address data entered as valid in the past may no longer have the correct (i.e., current) USPS ZIP Code. Although these changes generally can be considered rare, they may have a large impact on research studies in particular regions. Best practices relating to input data correctness are listed in Best Practices 44.

Best Practices 44 – Input data correctness

Policy Decision	Best Practice
Should incorrect portions of address data be corrected?	If information is available to deduce the correct attributes, they should be chosen and associated with the input address. Metadata should include: • The information used in the selection • The attributes corrected • The original values

17.5 ADDRESS LIFECYCLE PROBLEMS

The temporal accuracy of address data further depends on what stage in the address lifecycle both the input address and the reference data are at. New addresses take time to get into reference datasets after they are created, resulting in false-negative matches from the feature-matching algorithm. Likewise, they stay longer after they have been destroyed, resulting in false positives. For new construction in many areas, addresses are assigned by county/municipal addressing staff after a developer has received permission to develop the lots. How and when the phone companies and USPS are notified of the new address thereafter depends on the developer, staffing issues, and other circumstances, but this practice does occur. Thus, it may not appear in reference data for some time although it is already being reported at the diagnosing facility. Similarly, upon destruction, an address may still appear to be valid within a reference dataset for some time when it is in fact invalid. Also, just because an address is not in the reference dataset today does not mean that it was invalid in the past (e.g., the time period when the address was reported). These issues need to be considered when dealing with address data whose lifecycle status could be in question. Also, the length of time an individual was at an address (i.e., tenure of address) should be considered in research projects. Best practices related to address lifecycle problems are listed in Best Practices 45.

Best Practices 45 – Address lifecycle problems

Policy Decision	Best Practice
When and how can address lifecycle problems be accommodated in the geocoding process?	Address lifecycle problems can be overcome by obtaining the most recent address reference data for the region as soon as it becomes available, and by maintaining historical versions once new ones are obtained.
When and how should historical reference datasets be used?	The use of historical reference data may provide higher quality geocodes in the cases of: • Historical addresses where changes have been made to the streets or numbering • Diagnosis date may approximate the date the diagnosis address was in existence • If available, tenure of address should be taken into consideration during research projects

17.6 ADDRESS CONTENT PROBLEMS

In many cases, the content of the address used for input data will have errors. These can include addresses with missing, incorrect, or extra information. For all of these cases there are two options and choosing the correct one will depend upon the certainty obtainable for the attributes in question that can be determined from inspecting both the other attributes and the reference dataset. Such errors may be corrected or left incorrect. It should be noted that in some cases, this extra information may be useful. For example, "101 Main Street Apt 5" might be either "N Main St" or "S Main St," but perhaps only one is an apartment building. Best practices related to address content problems are listed in Best Practices 46.

Best Practices 46 – Address content problems

Policy Decision	Best Practice
What can and should be done with addresses that are missing attribute information?	If a correct reference feature can be unambiguously identified in a reference dataset from the amount of information available, the additional missing information from the reference feature should be amended to the original address, and denoted as such in the metadata record to distinguish it as assumed data. If a reference feature cannot be unambiguously identified, the missing data should remain absent.
What can and should be done with addresses that have incorrect attribute information?	If the information that is wrong is obviously the effect of an easily correctable data entry error (e.g., data placed into the wrong field), it should be corrected and indicated in the metadata. This action should only be taken if it can be proven through the identification of an unambiguous reference feature corresponding to the corrected data that this is the only possible explanation for the incorrect data. If it can be proven that there is more than one reference feature that could correspond to the corrected data, or there are multiple equally likely options for correcting the data, it should be left incorrect.
What can and should be done with addresses that have extra attribute information?	If the extra information is clearly not an address attribute and/or is the result of data entry error, it can be removed and this must be indicated in the metadata. It must be proven that this is the only possible reason why this extraneous data should be declared as such, though the use of the reference dataset before this removal can be made. Extraneous information such as unit, floor, building name, etc. should be moved into the Supplemental Field (NAACCR Item #2335) so that it can be retained for possible utilization at a later time. If there are equally probable options as to why this information was included, it should be retained.
What is the best way to correct address errors?	In the ideal case, addresses should be validated as they are entered at the hospital using, at a minimum, the USPS ZIP+4 database

17.7 ADDRESS FORMATTING PROBLEMS

Incorrectly formatted addresses and addresses with non-standard abbreviations should be handled by the address normalization and standardization processes. If not, human intervention may normalize and standardize them. Best practices related to address formatting are listed in Best Practices 47.

Best Practices 47 – Address formatting problems

Policy Decision	Best Practice
What can and should be done with address data that are incorrectly formatted?	If the address is formatted in a known format, the address normalization process could be applied to try to identify the components of the address and subsequently reformat it into a more standard format, which should be noted in the metadata. If the format of the original data is unrecognizable or the address normalization fails, it should be left in its original format.
What can and should be done with address data that include non-standard abbreviations?	The address normalization and standardization components of the geocoding process should be applied to correct the data and the corrections should be noted in the metadata. If these processes fail, the data should be left in its original format.
What should be done with extraneous address data?	Any extra information describing the location or address should be moved into the Supplemental Field (NAACCR Item #2335) for retention in the case that it becomes useful in the future.

17.8 RESIDENCE TYPE AND HISTORY PROBLEMS

Not knowing the type or tenure of address data can introduce uncertainty into the resulting geocode that is not captured merely with a quality code. This shortcoming usually is listed as a limitation of a study and is indicative of a larger public health data issue—these data are not collected during the primary data collection, after which point they generally are difficult to obtain. The missing information relates to items such as the tenure of residence, if it is their home or work address, if this is a seasonal address, and if the address is really representative of their true location if they move frequently or spend a lot of time traveling or on the road. As such, it is recommended that a tenure of residence attribute (i.e., length of time at address) also be associated with an address so that researchers will have a basic understanding of how well this address really represents the location of a patient. This fits with the current trend of opinions in the registry community (e.g., Abe and Stinchcomb 2008). The collection of historical addresses may not be practical for all addresses collected, but could certainly be attempted in small subsets of the total data to be used in small studies.

Currently, the NAACCR record layout does not include fields for these data items, so these would need to be stored outside of the current layout. In the future, a hierarchical and extendable format such as Health Level Seven (HL-7) (Health Level Seven 2007) could be adopted or embedded to capture this additional attributes within the NAACCR layout. Best practices related to conceptual problems are listed in Best Practices 48.

Best Practices 48 – Conceptual problems

Policy Decision	Best Practice
What can and should be done to alleviate address conceptual problems?	As much data as possible should be included about the type of address reported along with a record including: • Tenure of residence • Indication of current or previous address • Indication of seasonal address or not • Indication of residence or work address • Housing type (e.g., single family, apartment building) • Percent of day/week/month/year spent at this address

18. FEATURE-MATCHING PROBLEMS

This section discusses the various types of problems that occur during feature matching, as well as possible processing options that are available for non-matched addresses.

18.1 FEATURE-MATCHING FAILURES

There are two basic reasons why feature matching can fail: (1) ambiguously matching multiple features, and (2) not matching any features. When this occurs, the address can either remain non-matched and be excluded from a study or an attempt can be made to reprocess it in some different form or using another method. Recent research has shown that if a non-matchable address and the patient data it represents are excluded from a study, significant bias can be introduced. In particular, residents in certain types of areas are more likely to report addresses that are non-matchable (e.g., rural areas) and therefore data from such areas will be underrepresented in the study. It follows that simply excluding non-matchable addresses from a study is not recommended (Gregorio et al. 1999, Kwok and Yankaskas 2001, Durr and Froggatt 2002, Bonner et al. 2003, Oliver et al. 2005). For this reason, researchers and registries are advised to re-attempt feature matching by:

- **Hierarchical geocoding,** or using iteratively lower resolution portion of the input address for geocoding
- **Feature disambiguation,** or trying to disambiguate between the ambiguous matches
- **Attribute imputation,** or trying to impute the missing data that caused the ambiguity
- **Pseudocoding**, or determining an approximate geocode from other information
- **Composite feature geocoding,** or deriving and utilizing new reference features based on the ambiguous matches
- **Waiting it out,** simply doing nothing and attempting geocoding after a period of time (e.g., after the reference datasets have been updated).

Best practices relating to feature-matching failures are listed in Best Practices 49. Similar to the warning that match rates may be indicative of bias in one's geocoded data (Section 14.3), researchers need to be aware that using any of the following procedures to obtain a geocode for all of their data may also introduce bias into their datasets. A careful evaluation of the bias introduction from the use of these methods should be undertaken to determine if this may be an issue for one's particular dataset. This is an ongoing area of research and more detailed investigations into this topic are required before specific advice can be given on how to identify and deal with these problems.

Best Practices 49 – Feature-matching failures

Policy Decision	Best Practice
When and how can and should non-matchable addresses be handled?	All non-matchable addresses should be re-attempted using: • Attempt to obtain more information from source • Hierarchical geocoding • Feature disambiguation • Attribute imputation • Composite feature geocoding
When and how can and should ambiguous feature matches be handled?	Any time an ambiguous feature match occurs, only a single feature (which may be a composite feature) should be used for calculating the resulting geocode. If extra information is available that can be used to determine the correct feature, then it should be, and the metadata should record what was used and why that feature was chosen. If extra information is not available and/or the correct feature cannot be identified, a geocode resulting from the interpolation of lower resolution feature, composite feature, or bounding box should be returned.
When should a lower-resolution feature be returned from a feature-matching algorithm?	If the relative predicted certainty produced from feature interpolation using an attribute of lower resolution (e.g., USPS ZIP Code after street address is ambiguous) is less than that resulting from using a composite feature (if the features are topologically connected) or a bounding box (if they are not topologically connected), it should be returned.
When should a derived composite feature be used for feature interpolation?	If the matched features are topologically connected and if the predicted certainty produced from feature interpolation using a composite feature (e.g., street segments joined together) is less than that resulting from using an attribute of lower resolution, it should be used for interpolation.

Policy Decision	Best Practice
When should a derived bounding box be used for feature interpolation?	If the matched features are not topologically connected and if the relative predicted certainty produced from feature interpolation using a bounding box that encompasses all matched features is less than that resulting from using a lower resolution, it should be used for feature interpolation.
How and when can and should missing attributes be imputed?	Whether or not to impute missing attribute information will depend on the subjectivity of the registry or researcher. Metadata should indicate: • Which attributes are imputed • The sources used for imputing them • The original values of any attributes that have been changed
How and when should pseudocoding be used?	Whether or not to pseudocode will depend on the subjectivity of the registry or researcher. Metadata should indicate: • Which attributes were used to determine the pseudocode • The calculation used for approximating the pseudocode
How and when can and should geocoding be re-attempted at a later date after the reference datasets have been updated?	Geocoding should be re-attempted at a later date after the reference datasets have been updated when it is obvious that the geocoding failed because the reference datasets were out-of-date (e.g., geocoding an address in a new development that is not present in current versions of a dataset).

18.1.1 Hierarchical Geocoding

The first approach, hierarchical geocoding, is the one most commonly attempted. The lower resolution attribute chosen depends both on the reason why geocoding failed in the first place as well as the desired level of accuracy and confidence that is required for the research study, and is subject to the warnings regarding implied accuracies within arbitrary feature hierarchies as discussed in Section 15.1. To make the choice of lower resolution feature more accurate, one could use information about the ambiguous features themselves. If the two or more features returned from the feature-matching algorithm are of the same level of geographic resolution, the most probable course of action is to return the next level of geographic resolution to which they both belong. For example, if two streets are returned and both are in the same USPS ZIP Code, then a geocode for that USPS ZIP Code should be returned. If the two streets are in separate USPS ZIP Codes, yet the city is the same, the geocode for the city should be returned. The levels of accuracy for each of these would be the same as the level of accuracy of the level of geographic resolutions presented earlier, in Figure 20.

18.1.2 Feature Disambiguation

In the second approach, **feature disambiguation,** an attempt is made to determine which is the correct choice of the possible options. How this is done depends on why the ambiguity occurred as well as any other information that may be available to help in the choice of the correct option. These cases of ambiguity can result from an error in the reference dataset in the rare case that two separate reference features are described by the same attributes, but this usually indicates an error in the database and will not be discussed here.

Much more likely is the case in which ambiguity results from the input data not being described with enough detail, such as omitting a directional field or the house number. Here, disambiguation typically requires the time and subjectivity of a registry staff member, and is essentially interactive geocoding, but it could be done after the fact. The staff member selects one of the ambiguous matches as correct based on other information associated with the input data, or by reasoning what they have in common and returning the result of what can be deduced from this. The staff member performing the geocoding process can take into account any type of extra information that could be used to indicate and select the correct one. Going back to the source of the data (i.e., the hospital) to obtain some of this information may or may not be an option—if it is, it should be attempted.

For instance, if an input address was simply "Washington Township, NJ" without any form of a street address, USPS ZIP Code, or county (of which there are multiple), but it was known that the person was required to visit a hospital in a certain county due to particular treatment facilities being available, the county of the hospital could be assumed (Fulcomer et al. 1998). If a second hypothetical address, "1200 Main St.," geocoded in the past, but now after E-911 implementation the street has been renamed and renumbered such that the new address is "80 N. Main Street," and the reference data have not yet caught up, the registry could make the link between the old address and the new one based on lists of E-911 changes for their area. A third and more common example may occur when the directional attribute is missing from a street address (e.g., "3620 Vermont Ave," where both "3620 N. Vermont Ave" and "3620 S. Vermont Ave" exist. Solving these cases are the most difficult, unless some other information is available that can disambiguate between the possible options.

18.1.3 Attribute Imputation

Another approach that can be taken is to impute the missing input address attributes that would be required. Unless there is only a single, obvious choice for imputing the missing attributes that have rendered the original input data non-matchable, assigning values will introduce some uncertainty into the resulting spatial output. There currently is no consensus as to why, how, and under what circumstances attribute imputation should be attempted. At the time of this writing, imputing or not imputing is a judgment call that is left up to the registry, person, software, and most importantly, the circumstances of the input address.

A researcher will need to be aware of the greatest possible area of uncertainty that should be associated with the spatial output resulting from imputed data. Also, imputing different attributes will introduce different levels of uncertainty, from one-half the total length of a street in the case of a missing building number and a non-ambiguous street reference feature, to the MBR of possible city boundaries in the case for which ambiguous city names matched and one was imputed as the correct answer.

In all cases, registry staff and researchers need to be aware of the tradeoffs that result from imputing attributes. The confidence/validity one has in the imputed attributes increases if they have been verified from multiple sources. But, as the number of imputed attributes rise, it increases the likelihood of error propagation. Therefore, these imputed values need to be marked as such in the metadata associated with a geocode so that a researcher can choose whether or not to utilize a geocode based on them. The recent works by Boscoe (2008) and Henry and Boscoe (2008) can provide further guidance on many of these issues.

18.1.4 Pseudocoding

Another approach that can be taken is to impute an actual output geocode based on other available information or a predefined formula, known as **pseudocoding.** This has recently been defined by Zimmerman (2008) as the process of determining **pseudocodes,** which are **approximate geocodes.** These pseudocodes can be derived by deterministically reverting to a lower resolution portion of the input address (i.e., following the hierarchies presented in Section 15), or by more complex methods probabilistic/stochastic methods such as assigning approximate geocodes based on a specific mathematic distribution function across a region. Like attribute imputation, there currently is no consensus as to why, how, and under what circumstances pseudocoding should be attempted, but Zimmerman (2008) provides insight on how one should work with these data as well as different techniques for creating them.

18.1.5 Composite Feature Geocoding

If disambiguation through attribute imputation or the subjectivity of a staff member fails, the only option left other than reverting to the next best level of resolution or simply holding off for a period of time may be to create a new feature from the ambiguous matches and use it for interpolation, termed here **composite feature geocoding.** This approach can be seen as an application of the task of delimitating boundaries for imprecise regions (e.g., Reinbacher et al. 2008).

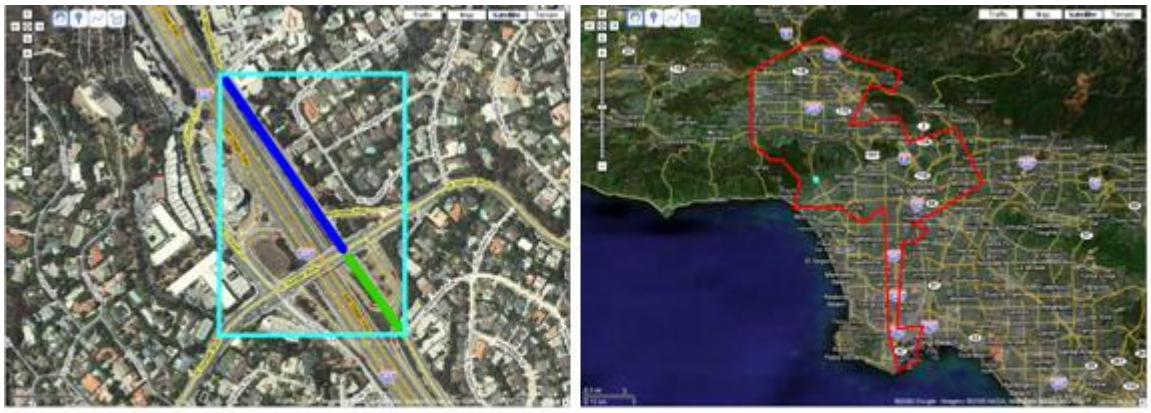
This approach already is essentially taken every time a geocode with the quality "midpoint of street segment" is generated, because the geocoder fundamentally does the same task—derive a centroid for the bounding box of the conjunction of all ambiguous features. Here, "all ambiguous features" consists of only a single street, and the centroid is derived using a more advanced calculation than strictly the "centroid of the bounding box of the ambiguous features." These generated features would be directly applicable to the

quantitative measures based on reference data feature resolution and size called for by Abe and Stinchcomb (2008, p. 124).

If geocoding failed because the street address was missing the directional indicator resulting in ambiguity between reference features that were topologically connected, one could geocode to the centroid of the overall feature created by forming an MBR that encompassed ambiguously matched features, if relevant to the study paying attention to whether or not the entire street is within a single boundary of interest. The relative predicted certainty one can assume from this is, at best, one-half of the total length of the street segment, as depicted in Figure 20(d). This level of accuracy may be more acceptable than simply reverting to the next level of geographic resolution.

However, taking the center point of multiple features in the ambiguous case may not be possible when the input data do not map to ambiguous features that are topologically connected (e.g., when the streets have the same name but different types and are spatially disjoint). Estimating a point from these two non-connected features can be achieved by taking the mid-point between them, but the accuracy of this action essentially increases to the size of the MBR that encompassed both.

This is depicted in Figure 24, in which the left image (a) displays the area of uncertainty for the ambiguously matched streets for the non-existent address "100 Sepulveda Blvd, Los Angeles CA 90049," with the "100 North Sepulveda" block represented by the longer line, the "100 South Sepulveda" block represented by the shorter line, and the MBR of the two (the area of uncertainty) represented by the box. This is in contrast to the size of the area of uncertainty for the whole of the City of LA, as shown in red versus the small turquoise dot representing the same MBR on the image to the right (b).



a) 100 North (longer line) and 100 South Sepulveda (shorter line) with MBR (box)

b) MBR of North and South Sepulveda (small dot) and LA City (outline)

Figure 24 – Example uncertainty areas from MBR or ambiguous streets vs. encompassing city (Google, Inc. 2008b)

Depending on the ambiguous features matched, the size of the resulting dynamically created MBR can vary greatly—from the (small) area of two blocks as in Figure 24 where the street segments are located next to each other, to the (large) area of an entire city where the streets with the same names and ranges appear on opposite sides of the city with only the USPS ZIP Code differing. Thus, it is impossible to indicate that taking the MBR always will be the correct choice in every case because the accuracy of a static feature, such as a single

city polygon, will in contrast always be the same for all features within it, no matter which of its child street segments are ambiguously matched and may represent a smaller area in some cases.

This tradeoff can be both good and bad in that the relationship between the areas of the feature with the static boundary (e.g., the city polygon) can be tested against the feature with the dynamic boundary (i.e., the dynamically created MBR of the ambiguous features) to determine and choose whichever has the smaller area of uncertainty (i.e., the one with the maximum relative predicted certainty). In addition, whether or not the ultimate consumer of the output can handle spatial data with variable-level accuracy, as in the MBR approach, or if they will require the static-level accuracy a uniform areal unit-based approach will produce, needs to be considered. Variations of all of the practices listed in this section may or may not be a cost-effective use of registry resources and will vary by registry (e.g., if accurate data are required for incidence rates). The possible options for dealing with ambiguity through composite feature geocoding as described in this section are listed in Table 37.

Table 37 – Composite feature geocoding options for ambiguous data

Problem	Example	Options
Ambiguity between connected streets	100 Sepulveda: ambiguous between 100 N and 100 S which are connected	• Intersection of streets • Centroid of MBR of streets
Ambiguity between disconnected streets	3620 Vermont: ambiguous between 3620 N Vermont and 3620 S Vermont which are not connected	• Centroid of MBR of streets

18.1.6 Waiting It Out

The final approach is to simply wait for the reference data sources to be updated and try the geocoding process again. This option is suitable if the staff member thinks the address data are indeed correct and that the reference files are out-of-date or contain errors and omissions. This is most often the case in rapidly expanding areas of the country where new construction is underway, or in old areas where significant reorganization of parcels or streets has taken place and street names and parcel delineations have changed between the temporal footprint of the reference data and the current time period. Updating the reference files may assist in these input data that are not represented in the old reference files. In some cases, it may be more suitable to use reference data more representative of the time period the addresses were collected (i.e., the remainder of the input data might suffer from these newly updated reference datasets). Also, this keeps the record in a non-matched state, meaning that it cannot be included in research or analyses, the exact problem pointed out in the opening of this section.

This page is left blank intentionally.

19. MANUAL REVIEW PROBLEMS

In this section, methods for attempting manual review are delineated as are the benefits and drawbacks of each.

19.1 MANUAL REVIEW

It is inevitable that some input will not produce an output geocode when run through the geocoding process. Also, the level of accuracy obtainable for a geocode may not be sufficient for it to be used. In some of these cases, manual review may be the only option. Best practices relating to unmatched addresses are listed in Best Practices 50. Boscoe (2008) and Abe and Stinchcomb (2008) can also be used as a guide in dealing with these circumstances.

Best Practices 50 – Unmatched addresses

Policy Decision	Best Practice
When and how can and should unmatched addresses be handled?	If the geocoding is done per record, an unmatched address should be investigated to determine a corrective action after it is processed, if time and money are available for the task. If the geocoding process is done in-batch, all unmatched addresses should be grouped by failure class (from Table 35) and processed together after the processing has completed, if time and money are available for the task.
When and how should geocoding be re-attempted on the updated input addresses?	The same geocoding process used for the original geocoding attempt should be applied again after the unmatched address has been corrected.

Manual review is both the most accurate and most time-consuming way to handle non-matchable addresses. Depending on the problem that caused the address to be non-matchable, the time it takes to perform a manual review can range from a few seconds to a few hours. An example of the first case would be when one of the components of the address attributes is obviously wrong because of incorrect data entry such as the simple examples listed in Table 38. This can be easily corrected by personal review, but might be difficult for a computer program to recognize and fix, although advances are being made (e.g., employing Hidden Markov Models and other artificial intelligence approaches as in Churches et al. [2002] and Schumacher [2007]). Few studies have quantified the exact effort/time required, but the NJSCR reports being able to achieve processing levels of 150 addresses per hour (Abe and Stinchcomb 2008). Goldberg et al. (2008d) also provide an analysis of the factors involved in these types of processes.

Table 38 – Trivial data entry errors for 3620 South Vermont Ave, Los Angeles, CA

Erroneous Version	Error
3620 Suoth Vermont Avve, Los Angels CA	Misspelling
3620 South Vermont Ave, CA, Los Angeles	Transposition

There are solutions that require research on the part of the staff but fall short of re-contacting the patient, which is usually not an option. Here, a staff member may need to consult other sources of information to remedy the problem when the input data are not trivially and/or unambiguously correctable. The staff member may obtain the more detailed information from other sources if missing data are causing the geocoding process to not match or match ambiguously. This task boils down to querying both the individual address components, combinations, and aliases against different sources (e.g., USPS ZIP+4 database [United States Postal Service 2008a]), other reference sets, local datasets, address points, and/or parcels to either identify an error/alias in the input data or an error/alias in the address or address range in the reference data, as well as the patient's name to find other additional address information (covered in the next section).

If a geocode has too low of a resolution to be useful, the staff member can reason what is the most likely candidate at the most reasonable level of geographic resolution that they can determine. Examples of when this could be done are listed previously in Section 18.1.2. Further options include re-contacting the hospital registry if the record is one that requires annual follow-up, in which case a corrected version of the same address may already have been obtained. As noted earlier, re-contacting the source of the data (e.g., the hospital) may or may not be a viable option.

At the other end of the spectrum are corrections that would require contacting the patient to obtain corrected or more detailed information about their address in the case that it originally was provided incorrectly or with insufficient detail. Typically, this would not be a task performed by a registry during the course of manual review to gain more information to successfully geocode a record. Instead, this would normally only be conducted by individual researchers for research purposes in special studies. Common examples for which this would be the only option include such things as an address consisting of only a USPS ZIP Code, a city/town name, or some other descriptive non-matchable term such as "Overseas" or "Military."

If approach is a valid option, it typically will result in the highest accuracy because the staff member can potentially keep attempting to geocode the address with the patient on the telephone, asking for more information until they obtain a successful geocode of sufficient accuracy. Best practices related to manually reviewing unmatched addresses are listed in Best Practices 51.

Best Practices 51 – Unmatched addresses manual review

Policy Decision	Best Practice
When and how can and should manual review of unmatched addresses be attempted?	If the time and money are available, manual review should be attempted for any and all addresses that are not capable of being processed using automated means.
When and how can and should incorrect data (data entry errors) be corrected?	If the error is obviously a data entry error and the correction is also obvious, it should be corrected and the change noted in the metadata.
When and how can and should a geocode at a higher resolution be attempted to be reasoned?	If the geographic resolution of the output geocode is too low to be useful (e.g., county centroid), a staff member should attempt to reason what better, higher resolution geocode could be assigned based on other information about the patient/tumor (e.g., use city centroid of the diagnosing facility if it is known they visited a facility in their city).

19.2 SOURCES FOR DERIVING ADDRESSES

Often, even though a record's address may not contain enough or correct content to be directly useful, other information associated with the record may lend itself to providing a more accurate address. For example, if a patient-reported address is not useful, the staff member might be able to link with the state DMV and obtain a valid address for the patient. This solution assumes that a working relationship exists between a registry and the DMV in which the former may obtain data from the latter, and in some cases this may not be feasible at the registry level. In these cases, a registry may be able to work with a government agency to set up a relationship of this type at a higher level, instead of the registry obtaining the patient's DMV data directly.

> **It must be stated that when registries utilize additional sources for gathering additional information about individuals, the intention is not to "snoop" on any particular individual. The purpose of gathering this information is entirely altruistic and central to facilitating their role in reducing the burden of cancer.**

In general, linking with large, administrative databases such as the DMV or Medicare can be valuable for augmenting demographic information, such as address, on a cancer record. However, these databases are for administrative purposes and are not intended for surveillance or research. The limitations of these databases for cancer registry objectives need to be understood. For example, although DMV requires a street address in addition to a USPS PO box, the address listed in DMV may have been updated and overwritten since the time of cancer diagnosis. Cancer registry personnel must fully understand the data collection methods to make correct assumptions when attempting to supplement cancer registry data. These situations obviously will be specific for each registry and dependent on local laws.

Other sources of data associated with a patient that have been used in the literature as well as other possible sources are found in Table 39. Some of these sources (e.g., phone books) can be accessed for free as online services; others may require agreements to be made between the registry and private or public institutions. By far, the most common approach is

to look for a patient's name in parcel ownership records and associate the address if the match seems reasonable (e.g., a one-to-one match is found between name and parcel during the correct time period when the person was known to be living in that city). Issues related to querying some of these online sources are covered in Section 26. Best practices related to data sources for manual review are listed in Best Practices 52.

Best Practices 52 – Unmatched address manual review data sources

Policy Decision	Best Practice
When and how can and should alternative sources of information be reviewed to assist in address correction?	If the problem with the input address is not trivially correctable, alternative sources of information should be reviewed to attempt address correction, if time and money are available for the task. If a linkage can be determined with a suitable level of certainty, it should be made as long as privacy and confidentiality concerns in Section 26 are satisfied. Metadata should include: • The source of the supplemental data • The staff member who made the linkage • The method of linkage (i.e., automatic/manual) • The linkage criteria • The date the linkage was made

Table 39 – Common sources of supplemental data with typical cost, formal agreement requirements, and usage type

Supplemental Data Source	Cost	Formal Agreement Required	Usage Type
DMV	free	yes	batch
Phone Books	free	no	per-record
Phone Companies	free	yes	batch
Utility Companies	free	yes	batch
State Bureau of Vital Statistics and Registration	free	yes	batch
Social Security Administration	free	yes	per-record
Military	free	yes	per-record
Educational Institutions	free	yes	per-record
USPS ZIP+4 Database	not free	no	batch
County Deeds/Real Estate Transaction Registries	not free	no	batch
Municipal or County Assessor Databases	varies	no	batch
Municipal Resident Lists	free	no	per-record
State or Municipal Voter Registration Databases	varies	no	batch
Google Earth	free	no	per-record
Social Security Death Index	free	no	per-record
People Finding Web Sites	free	no	per-record
Medicare/Medicaid	free	yes	per-record
Vital Statistics	free	yes	batch
"Googling" a Person's Name	free	no	per-record

This page is left blank intentionally.

20. GEOCODING SOFTWARE PROBLEMS

This section provides insight into required or possible methods of overcoming problems with geocoding software.

20.1 COMMON SOFTWARE PITFALLS

Several common problems and limitations often occur when using commercial geocoding packages. Perhaps the most frustrating is that commercial geocoding processes by their very nature do not reveal much about their inner workings. The actual algorithms implemented in commercial products can be held as trade secrets and not provided in detail (they are, for the most part, sold to make money). As such, a registry or researcher may not know the exact choices made by the components, what its other possible options were, or how the final choice was decided. Some older geocoding platforms simply return the spatial output without any metadata reporting how or why it was derived or providing information on its quality. This can and should prevent a user from having confidence in these results and should be avoided. Registries and consumers of commercial software should push these vendors to be more open about the inner workings of the geocoding processes that they implement, and about reporting metadata.

However, even when commercial software packages expose their algorithms and assumptions, it will take a time commitment from a staff member to read and understand them. Best Practices 53 contains several common limitations that are encountered in working with geocoding software and recommended NAACCR actions for overcoming them. Note that the intent of this section is to remain "software neutral" by refraining from advocating any particular commercial geocoding platform. Also, as the cost of geocoding continues to drop, even becoming free (Goldberg 2008a) some of the issues in this section may no longer apply.

Best Practices 53 – Common geocoding software limitations by component of the geocoding process

Aspect	Limitation	Best Practice
Input Data	Not accepting intersections as input.	The input data will need to be prepared such that one street or the other is chosen and used for input. The street that will produce the lower level of uncertainty should be chosen (e.g., the shorter one).
	Not accepting named places as input.	The input data will need to be prepared such that the next highest level of resolution should be used for input (e.g., move from named building to USPS ZIP Code).
Normalization/Parsing	Cannot change the order of address attributes (tokens).	The input data will need to be prepared such that the address attributes are in the accepted order.
Standardization	An input address standard is not supported.	The input data will need to be prepared such that the input data are in the accepted address standard.
Reference Dataset	Only linear-based reference datasets are supported.	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
	There is no control over which reference dataset is used.	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
Feature Matching	There is no control over which feature-matching algorithm is used.	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
	There is no control over the parameters of the feature-matching algorithm used (e.g., confidence interval).	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
	There is no control over which feature interpolation algorithm is used.	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
	There is no control over the parameters of the feature interpolation algorithm used.	If particular parameters must be settable (e.g., dropback distance and direction, which assumptions are used uniform lot, address range, etc.), a different geocoding process should be identified and obtained.
Output Data	Only capable of producing a single point as output (i.e., no higher complexity geometry types).	This is how most geocoders operate so, at present, in most circumstances this will have to be acceptable.
Metadata	No metadata reported along with the results.	Geocodes returned from a geocoder without any metadata should not be included as the spatial component of the record.
	No coordinate quality reported along with the results.	Geocodes returned from a geocoder without metadata describing, at a minimum, a coordinate quality should not be included as the spatial component of the record.

Part 5: Choosing a Geocoding Process

The choice of a geocoding process and/or components used by a registry will necessarily depend foremost on the restrictions and constraints the registry must adhere to, be they legal, budgetary, technical, or otherwise. The material in this part of the document is presented to help a registry determine the correct solution for them, given their particular restrictions.

This page is left blank intentionally.

21. CHOOSING A HOME-GROWN OR THIRD-PARTY GEOCODING SOLUTION

This section will identify the requirements that must be defined to determine the appropriate geocoding method to be used.

21.1 HOME-GROWN AND THIRD-PARTY GEOCODING OPTIONS

In practice, many different geocoding process options are available to registries and the general public. At the top level, they can either be developed in-house or obtained from a third party. Of the ones that are not developed by the registry, many different varieties are available. They can range in price as well as quality, and are available in many different forms. There are complete software systems that include all of the components of the geocoding process and can be used right out of the box, or each component can be purchased separately. There are freely available online services that require no software other than a Web browser (e.g., Goldberg 2008a, Google, Inc. 2008c, Yahoo!, Inc. 2008), and there are proprietary commercial versions of the same (e.g., Tele Atlas 2008b). Choosing the right option will depend on many factors, including budgetary considerations, accuracy required, level of security provided, technical ability of staff, accountability and metadata reporting capabilities, and flexibility required.

21.2 SETTING PROCESS REQUIREMENTS

Setting the requirements of one's geocoding process should be the first task attempted, even before beginning to consider the different available options. The items listed in Table 40 should serve as a starting point for discussions among registry staff designing the geocoding process when determining the exact requirements of the registry in terms of the geocoding software, the vendor, and the reference data. It also is worthwhile to have a single person designated as the dedicated liaison between the vendor and the registry for asking/answering questions to both build relationships and become an expert on the topics. Keep in mind that geocoding outcomes, at a minimum, must meet the standards set forth in the NAACCR GIS Coordinate Quality Code Item (see Section 15).

Table 40 – Geocoding process component considerations

Process Component	Topic	Issue
Software	Accuracy	What is the minimum acceptable level of spatial accuracy? Are multiple levels of accuracy accepted or required?
	Capacity	What throughput must be achievable? How many concurrent users are expected?
	Reliability	Must the software be failsafe?
	Transparency	What level of overall transparency is required? What level of transparency is required per component?
	Reportability	What information must be reported along with a result?
Vendor	Accuracy	What level of accuracy and completeness do they guarantee?
	Capacity	What capacity can they accommodate?
	Reliability	How reliable are their results?
	Transparency	What information is available about the process they use?
	Reportability	What information do they report along with their results?
Reference datasets	Accuracy	What level of accuracy is associated with the dataset as a whole? What level of accuracy is associated with individual features?
	Completeness	How complete a representation do the reference features provide for the area of interest?
	Reliability	How reliable can the reference features be considered?
	Transparency	How were the reference features created?
	Lineage	Where and when did this data source originate? Is there a date/source specified for each feature type? What processes have been applied to this data source?

21.3 IN-HOUSE VS. EXTERNAL PROCESSING

Whether to perform the geocoding process in-house or by utilizing a third-party contractor or service is perhaps the most important decision that a registry must make regarding their geocoding process. The NAACCR documents (2008a, 2008b) and the work by Abe and Stinchcomb (2008) that present a compressed version report that 28 percent of their registry survey respondents (n=47) report utilizing in-house geocoding, while 70 percent utilize external resources (33% using commercial vendors, 37% using different divisions within their organization).

Performing the process in-house enables a registry to control every aspect of the process, but forces them to become experts on the topic. Contracting a third-party entity to perform the geocoding releases the registry from knowing the intimate details of the

implementation of the geocoding process, but at the same time can keep them in the dark about what is really happening during the process and the choices that are made. The costs associated with using a vendor will vary between registries because the requirements the vendor will have to meet will vary between registries.

Another option is to use a mixture of both in-house and vendor geocoding. The common case of this is sending out all of the data to a vendor, then working on the problem cases that are returned in-house. Best practices related to geocoding in-house or externally are listed in Best Practices 54.

Best Practices 54 – In-house versus external geocoding

Policy Decision	Best Practice
When can and should a registry geocode data in-house or use a third party?	If the cost of geocoding using a third-party provider is higher than obtaining or developing all components of the geocoding process, or if no suitable confidentiality and/or privacy requirements can be met by a third party, it should be performed in-house. If the technical requirements or costs for geocoding in-house cannot be met by a registry and suitable confidentiality and/or privacy requirements can be met by a third party, it should be performed by a third party.

21.4 HOME-GROWN OR COTS

If the choice is made to perform the geocoding in-house, the registry must further decide if they wish to implement their own geocoder, or if they prefer to use a commercial off-the-shelf (COTS) software package. Both of these options have their strengths and weaknesses, and this choice should be considered carefully. On one hand, a COTS package will be, for the most part, ready to start geocoding right out of the box. However, there will be a substantial up-front cost and the details of its inner workings may or may not be available. A home-grown geocoding solution, on the other hand, can take a significant amount of time to develop and test, costing possibly several times more than a commercial counterpart, but its inner workings will be known and can be molded to the particular needs of the registry and modified as needed in the future. As of this writing, only a few registries currently use a composite/home-grown solution (e.g., the NJSCR [Abe and Stinchcomb 2008], but as open-source geocoding platforms become available, such as the one being developed at the University of Southern California (Goldberg et al. 2008a), this may change.

A comprehensive list of both: (1) the commonly encountered costs of using commercial vendors (high end and low end for per-record and batch processing), as well as (2) the costs of performing in-house geocoding including all associated costs for setup (purchasing reference data, purchasing geocoding software, developing custom geocoding software, employee training, etc.) are missing from the registry community. In many cases, the contracts under which the data or services are obtained from vendors are confidential, which may be the major hurdle in assembling this list. However, these data items would be extremely useful for registries just starting to form their geocoding processes and procedures.

21.5 FLEXIBILITY

Commercial software providers tend to create "one-size-fits-all" solutions that appeal to the largest market possible. They usually are not capable of creating individualized software releases for each of their customers geared toward fitting their exact needs. This strategy, while beneficial to the software vendor, almost always results in software that does not *exactly* meet the needs of the consumer. Surely, most of their customers' requirements will be met (otherwise they would have never bought the software), but undoubtedly some specific requirement will not be fulfilled, causing difficulty at the consumer end.

With regard to geocoding software in particular, the lack of flexibility may be the most problematic. The general nature of geocoding, with many different types of input data, reference data sources, and output types, necessitates a certain degree of dynamism in the process. Measuring a particular strategy in terms of how easily it can be adapted to changing conditions of reference datasets, new geocoding algorithms, and varying degrees of required levels of accuracy can be considered the defining metrics of any geocoding process. Therefore, a critical factor that must be taken into consideration when choosing between in-house and commercial geocoding is the amount of flexibility that can be accommodated in the process.

To address this question, a registry will need to determine an anticipated amount of leeway that they will need and investigate if commercially available packages suit their needs. Examples of specific issues that may need to be considered are listed in Table 41. If the needs of a registry are such that no commercial platform can accommodate them, the only option may be to develop their own in-house version.

Table 41 – Commercial geocoding package policy considerations

Policy Considerations
The ability to specify/change reference data sources
The ability to specify/change feature-matching algorithms to suit particular study needs
The ability to specify/change required levels of accuracy
The ability to specify/change feature interpolation algorithms
The ability to control offset/backset parameters

21.6 PROCESS TRANSPARENCY

Process transparency within a geocoding strategy can be critical in determining the confidence one can associate with the results. In the worst possible case, no metadata are returned with a geocode to indicate how it was derived or where it came from. Data in this form should, for the most part, be avoided because they will have no indication of the reliability of the data from which their results are to be obtained. The amount of transparency associated with a geocoding process can be measured in terms of the amount and type of information that is returned with the result. A process that reports every decision made at every step of the process would be completely transparent, while the worst-case scenario just presented would be completely non-transparent.

Commercial geocoding packages vary in terms of transparency. Some do not expose the inner workings of their geocoding process to customers based on "trade secret" concerns, and as such the consumer must take the vendor at their word regarding the operation of the geocoding software. Others, however, report more detailed information. A registry will need to decide what level of transparency it requires and either select a commercial geocoding

package accordingly or develop one in-house if nothing is commercially available that meets these needs. Best practices related to process transparency are listed in Best Practices 55.

Best Practices 55 – Process transparency

Policy Decision	Best Practice
What minimum information needs to be reported by a geocoding process along with the output (e.g., level of transparency)?	Full metadata describing the reference data. The type of feature match. The type of interpolation performed.

21.7 How To Select a Vendor

If the choice has been made to select a vendor to perform geocoding, a registry should ask specific questions to determine both the knowledge they have about the topic as well as the likely capabilities they will be able to provide. Example questions are provided in Table 42.

Table 42 – Topics and issues relevant to selecting a vendor

Topic	Issue
Data sources	What types of data sources are used? What versions? How often are they updated? What level of accuracy do they have? How are those levels guaranteed?
Feature matching	What algorithms are used? How are exceptional cases handled?
Normalization/ Standardization	What algorithms are used?
Input data	What types of input data can be handled? What happens to input data that are not able to be handled?
Output data	What output formats are supported? What information is provided along with the output? What level of spatial accuracy can be achieved? Can they provide the NAACCR certainty codes from Table 34?
Confidentiality	What safeguards and guarantees are in place?
Cost	What is the per-record cost? Do they negotiate separately with each registry and require non-disclosure agreements?
Feedback capability	Do corrections submitted by a cancer registry lead to updates in their data sources?

In addition to receiving answers to these questions, it is advisable for a registry to test the accuracy of their vendor periodically. Published literature exist defining methods for testing both a third-party contractor, as well as commercially purchased software (e.g., Krieger et al. 2001, Whitsel et al. 2004). Essentially, a registry wishing to validate the results from a vendor

can set up lists of input data that include particularly hard cases to successfully geocode, which then can be submitted to the vendor to determine how well they handle the results.

Also, the registry can maintain a small set of ground truth data obtained with GPS devices to measure the spatial accuracy of the output data returned from the vendor. It should be noted that it may be beneficial for registries to coordinate in compiling and maintain a large-area ground truth dataset for this purpose—instead of each registry maintaining a small set just for the area around their geographic location—to leverage the existing work of other registries.

At the time of writing, there are no currently established rules between vendors and registries as to who is responsible for ensuring the correctness of the geocoded results. By default, the registry is ultimately responsible, but it may be worthwhile to initiate discussions on this topic with vendors before entering into any agreements.

21.8 Evaluating and Comparing Geocoding Results

Although geocoding outcomes between two geocoders will be identical for most addresses, there always are subsets of addresses that will generate different outcomes for which there is no clear consensus on which is more correct. Methods of objectively comparing geocoding results that can be applied to both the results returned from vendors and those from other agencies are just emerging (e.g., Lovasi et al. 2007, Mazumdar et al. 2008). The objective when comparing geocoding outcomes is one of placing addresses into the categorization listed in Table 43. The third category represents the instances in which one party believes that all parties involved should be satisfied with the geocodes while another party believes otherwise.

Table 43 – Categorization of geocode results

1) Addresses that can be geocoded to the satisfaction of all parties concerned
2) Addresses that cannot be geocoded to the satisfaction of all parties concerned
3) Addresses for which there is disagreement as to whether they belong in (1) or (2)

The differences are based on assumptions used during the geocoding process, which generally are universally applied for all data processed using a single geocoder, and thus are relatively easy to identify. For example, it may be simple to determine if and which hierarchy of feature matching was used (e.g., matching a USPS ZIP Code centroid every time a street-level match fails). Explicitly defining the required criteria of the geocoding process is the way for registries to make their vendor's geocoding outcomes more similar to those that they may generate in-house.

When the geocoding output of two individuals is compared (as in the between-agency case), the assumptions are less easy to identify because they generally are made on a per-record basis. There will be addresses for which one person thinks an address can geocode based on X number of edits to the address, and another person disagrees and thinks that the record should not be geocoded because the edits required for such a match are based on assumptions not supported by the data at hand. Fortunately, these addresses generally comprise less than 5 percent of all registry records. Best practices related to evaluating third-party geocoding results are listed in Best Practices 56.

Best Practices 56 – Evaluating third-party geocoded results

Policy Decision	Best Practice
When can and should results from a third-party provider be verified?	The results from a vendor should be verified after every submission to ensure some desired level of quality.
How and what can and should be used to verify the quality of a vendor?	A pre-compiled list of problem addresses to check exceptional case handling. Resubmitting data to check for consistent results. A small set of GPS ground truthed data can be used to check the spatial accuracy, with its size based on confidence intervals (although a recommendation as to its calculation needs more research).

This page is left blank intentionally.

22. BUYING VS. BUILDING REFERENCE DATASETS

This section introduces methods for determining which reference datasets to use.

22.1 NO ASSEMBLY REQUIRED

Another decision that registries will be confronted with when they choose to not use a vendor relates to obtaining reference datasets used in the geocoding process. These can either be obtained (possibly purchased) and used directly, created at the registry, or a combination thereof whereby the registry adds value to an acquired dataset.

As noted earlier, the accuracy and completeness of these reference datasets can vary dramatically, as can the costs of obtaining them. The price of free reference datasets such as TIGER/Line files (United States Census Bureau 2008d) makes them attractive. The relatively lower levels of accuracy and completeness included in these (when compared to commercial variants), however, may not be sufficient for a registry's needs. Likewise, although commercial datasets such as Tele Atlas (2008c) will undoubtedly improve the overall accuracy of the geocoding process, their cost may be too prohibitive.

22.2 SOME ASSEMBLY REQUIRED

An alternative approach is for a registry to improve the reference data source, be it a public data source (e.g., TIGER/Line files) or a commercial counterpart. If the technical ability is available, this option may prove worthwhile. For example, one simple improvement that can be made to TIGER/Line files that greatly improves the accuracy of linear-based interpolation is to associate more accurate address ranges with each street reference feature, thus enabling uniform lot interpolation (Bakshi et al. 2004). The required actual number of parcels per street is typically available from local assessors' offices, and they may be willing to share that information with a registry. Usually, these values can easily be associated with each of the street reference features in the TIGER/Line files, unless it is determined they will compromise the synchronicity with census attribute or population data already in use.

Another option is for the registry to directly contact the local government offices that make use of GIS-related data sources to determine if suitable reference data sources exist for use in the geocoding process. Instead of just obtaining the number of parcels along street segments as in the previous example, why not get the actual parcel boundary files from the local assessor's office? If they exist, the assessor's office may be willing to share them with the registry, for free or some fee. In some cases they are becoming available at the state level. Either of these options effectively increases the **matching ability,** or the ability of the reference dataset to match addresses while still maintaining the same linkage criteria.

22.3 DETERMINING COSTS

In general, a registry will need to decide if the increased level of accuracy they can expect from an improved reference dataset is worth the additional cost (e.g., does the level of improved geocode accuracy directly attributable to using Tele Atlas [2008c] instead of TIGER/Line files justify the price of purchasing them?). Further, they need to determine the potential costs of upgrading an existing dataset themselves versus buying one that is already enhanced. The calculation of these costs should include both the initial ascertainment and

acquisition costs, as well as the cost of adding any value to the datasets, if it is to be performed. The cost of maintenance for software and reference data for in-house operations also should be taken into consideration. Best practices relating to choosing a reference dataset are listed in Best Practices 57.

Best Practices 57 – Choosing a reference dataset

Policy Decision	Best Practice
What type of reference data should be chosen?	If a registry requires the ability to investigate and fix addresses that might be incorrect on a street, multi-entity (range) reference data should be used. If a registry needs all addresses to be validated as existing single-entity (discrete) reference features should be used (i.e., point- or areal unit-based). The scale of the reference features should match the scale of the input data element being matched. • National scale city names – National gazetteers (included in TIGER/Lines) • National scale USPS ZIP Codes – National USPS ZIP+4 databases The highest resolution reference dataset should be chosen (given budgetary constraints).
Which point data sources should be used?	If no feature interpolation is possible given the types of data to be geocoded, point-based reference datasets should be considered. All geocoding processes should, at a minimum, include the national-scale gazetteers included with TIGER/Line files.
When and which line data sources should be used?	All geocoding processes should contain at least one linear-based reference dataset. All geocoding processes should, at a minimum, include the TIGER/Line reference files.
When and which polygon-based reference dataset should be used?	If high-resolution data sources are available (e.g., parcel boundaries, building footprints), they should be included in the geocoding process. If 3-D geocoding is required, building models should be used.
When multiple reference dataset types are available, which should be used?	The reference feature type that produces the lowest amount of uncertainty should always be used (e.g., a linear-based reference datasets with high-resolution small features may be more suitable than large parcels in rural areas).

23. ORGANIZATIONAL GEOCODING CAPACITY

This section explores the possible options for developing geocoding capacity at a registry.

23.1 HOW TO MEASURE GEOCODING CAPACITY

The amount of geocoding performed at a registry will necessarily affect the choice of the geocoding option selected. Any decisions made for the determination of an appropriate geocoding process need to first and foremost take into account the research needs and policies of the registry, with the amount of geocoding likely to be performed also considered as a factor. The number of cases geocoded can vary dramatically between registries, and a first order of magnitude estimation of an approximate number of cases will have an effect on which strategy should be undertaken. To determine an estimate for an approximate yearly number of cases, a registry should determine how many cases they have had in previous years. These prior numbers can be good indicators of future needs, but may be biased depending on the particular circumstances contributing to them. For instance, policy changes implemented within a registry as of a particular year can increase or decrease these numbers, and potential future policies will need to be taken into account.

The costs of data and software are effectively "sunk" costs, with the real cost of yearly geocoding performed at a registry depending on the amount of time spent on the task. Therefore, in addition to determining an average number of cases per year, a registry should also determine the average amount of time the interactive geocoding process takes on a per-case basis (because batch match cost is trivial on a per-record basis). The specific geocoding policies in place at a registry will have a substantial effect on this estimate, and likewise the estimated amount of time per case may affect these policies. Time is particularly dependant on the desired geographic level of output. For instance, if a policy were in place that every input address needed to be geocoded to street-level accuracy, the time necessary to achieve this might quickly render the policy infeasible, but requiring a geocode for every address to county level accuracy may be substantially quicker.

The most reliable cost estimates for geocoding at a registry often are obtained when a registry charges for cost recovery because most likely, the client will set the geocoding criteria. Based on the number of geocoding cases that a registry has determined likely and costs for these (as determined by the average amount of time per case), a registry should evaluate the necessity of creating one or more full-time equivalent (FTE) positions dedicated to the task. Example numbers of cases geocoded per year and resulting FTE positions from registries are provided in Table 44 (Abe and Stinchcomb 2008). Best Practices relating to measuring geocoding capacity are listed in Best Practices 58.

Table 44 – Comparison of geocoded cases per year to FTE positions

Registry	Cases Geocoded Per Year	Number of FTE Positions
North Carolina	10,000+	1
New Jersey	80,000+	2
New York	100,000+	4

Best Practices 58 – Measuring geocoding capacity

Policy Decision	Best Practice
When and how should the predicted number of geocoding cases be calculated?	The number of geocoding cases should be calculated before selecting a geocoding process. This can be done by estimating from the number of cases geocoded in previous years.
When and how should average processing time per geocoded case be calculated?	The average per-geocode processing time should be calculated as cases are processed.
How many FTE positions are required for the geocoding process?	This number will depend on the amount of cases that need to be geocoded and the actual work time associated with processing each geocode, which also will depend on the level to which the process is automated.

Part 6: Working With Geocoded Data

After data have been successfully geocoded, they are subsequently utilized in any number of ways. Due to the nature of cancer registries and the types of data that they routinely work with, several peculiarities exist as to how and why particular types of processing can and sometimes must take place. Some of these safeguard the rights of individuals, while others are artifacts of the way in which cancer registries do business. This part of the document will discuss several of the more important issues in play here.

This page is left blank intentionally.

24. TUMOR RECORDS WITH MULTIPLE ADDRESSES

This section discusses the issues that explain why multiple addresses sometimes are generated for a single tumor record and how these should be interpreted and used in subsequent analyses.

24.1 SELECTING FROM MULTIPLE CASE GEOCODES

It is common for a record to have several addresses associated with it, with each one representing the individual's residence. This multi-address situation can occur for several reasons based on when a patient is seen by multiple facilities, each of which records an address for the patient (which can be the same or different), or the patient is seen at the same facility on multiple occasions. Additionally, if multiple abstracts with multiple addresses where received for a single tumor during registry consolidation (in the case that the subject moved their residence, reported different addresses at different times, and/or was treated at multiple facilities), this situation also will arise. Further, one or more facilities may have two different versions of the same address for a single patient, or two different addresses may be maintained because the patient actually moved.

In these cases when multiple addresses exist for a record, one address must be selected as the primary for use in spatial analyses. The others need not be discarded or removed from the abstract, but the primary address should be the only one used for performing analysis. In the ideal case, when confronted with multiple addresses, one should strive to use the address that is based on the more accurate data. The standard in place at registries is to use the patient's usual address at diagnosis (dxAddress), no matter what its quality.

However, if it is unclear which is the primary dxAddress out of several possible addresses associated with a single patient (in the case of multiple tumor abstracts being received), a decision needs to be made as to which should be used. For instance, consider the case when one geocode was produced the first time the patient saw a doctor several decades ago using only the USPS ZIP Code, and another geocode was subsequently created using a full street address obtained from historical documents of unknown accuracy. These both represent the dxAddress, but have different values. The first geocode is most likely more temporally accurate because the data were reported by the patient him or herself, while the second is most likely more spatially accurate, if one assumes that the address obtained is correct. There is no current consensus on how to proceed in these instances, but there has been a recent push to standardize the way that registries deal with records that have multiple addresses. Research is underway to develop best practices for this situation, and the considerations listed in Table 45 have been identified as possible factors that could/should influence these types of decisions.

Table 45 – Possible factors influencing the choice of dxAddress with decision criteria if they have been proposed

Index	Factor	Decision Criteria
1	Abstract submission date	Earliest to latest
2	Address of diagnosing facility	
3	Age at diagnosis	
4	Amount time elapsed between diagnosis and first contact	Least to most
5	Class of case	Class code, this order:{1,0,2,3,4,5,6,7,8,9}
6	Current address of patient	
7	Date of diagnosis on abstract	
8	Date of last contact	
9	External address reference	E.g., motor vehicles database with address near diagnosis time
10	Facility referred from	
11	Facility referred to	
12	Fuzziness of match	E.g., level massaging required standardize address
13	Marital status at diagnosis	
14	Particular reporting facility	Use address from the most trusted one
15	Place of death	
16	Specificity (geographical) of address type	Street address > USPS ZIP Code only > county only
17	Timeliness of reporting	Time elapsed between times of case reportability and submission (less is better)
18	Type of street address-related data submitted	E.g., prison address < known residential address E.g., standard street address < USPS PO box or rural route location E.g., contingent on patient age: old in nursing homes assumed appropriate, young in college assumed appropriate
19	Type of reporting facility	E.g., American College of Surgeons or NCI hospitals < other in-state hospital < in-state clinic < other sources
20	Vital status	Alive < dead

25. HYBRIDIZED DATA

This section introduces the concept of hybridized data and the ways in which it is produced and used.

25.1 HYBRIDIZED DATA DEFINED

Typically, once the geocodes have been created for a set of input data, the next step in the spatial analysis procedure is to associate other relevant aspatial data with them. This process has been termed **hybrid georeferencing,** which describes the association of attributes from other datasets through a spatial join operation based on common spatial attributes (Martin and Higgs 1996). **Hybrid data** are the data created through hybrid georeferencing with attributes from more than one dataset, joined by common spatial attributes.

The **point-in-polygon method** is an approach in which a point that is spatially contained within a polygon has the attributes of the polygon associated with it. This is the most common method of hybrid georeferencing and is suitable when the secondary data that a researcher wishes to associate with a geocode are in a vector-based polygon format. Alternatively, when the data are in a raster-based format, this notion of area within a geographic feature typically does not exist because of their pixel-based nature. In these cases spatial buffers (i.e., catchment areas) around the point normally are used to obtain aggregate values over an area of the raster to associate with the point.

Without the support of a GIS or spatial database capable of performing spatial operations, some type of non-spatial association, either probabilistic or deterministic text linkage, may be the only option. These non-spatially oriented methods of associating supplemental data usually consist of relational database-like procedures in which an aspatial attribute associated with the geocode is used as a key into another dataset containing values for objects with the same key. For example, a city name attribute associated with a geocode can be used as a key into a relational database containing city names as identifiers for demographic information. Note that some vendors and organizations hybrid their reference data (e.g., maintain a layer of municipality names joined with streets).

This association of attributes with geocodes from other datasets using spatially based methods must be undertaken carefully; users need to pay careful attention to the base layers, source, and data used for hybridizing. A researcher must consider how accurate and representative these associations are, given the information that they know about both their geocodes and the supplemental data sources. The literature has shown that point-in-polygon methods are capable of incorrectly associating data from the wrong polygon to a geocode for numerous reasons (e.g., as a result of the data distribution the point-in-polygon methods assume [Sadahiro 2000], simple inaccuracies of the supplemental data, resolution differences between the polygon boundaries and the street networks from which geocodes were produced [Chen W. et al. 2004], or inaccuracy in the point's location). For these reasons, associating higher-level data with geocodes through non-spatial-joins may be an attractive option (e.g., obtaining a CT by indexing off the block face in the TIGER/Line files) rather than relying on a geocoded point.

The accuracy and appropriateness of the geocode in terms of both spatial and temporal characteristics also must be considered. Geocodes close to the boundaries of areal units can

and do get assigned to the wrong polygons, resulting in the incorrect linkage of attributes. Likewise, the polygons used for the hybridization may not be representative of the spatial characteristics of the region they represent at the relevant time. For example, if a case was diagnosed in 1990, it may be more appropriate to associate 1990 Census data with the residence address, rather than Census data from 2000 or 2010. Characteristics of both the geocode and the supplemental datasets used to derive hybrid data need to be recorded along with the hybrid datasets produced so that registries and researchers can weigh these factors when deciding on the appropriateness, correctness, and usefulness of the results derived from them. Best practices relating to the hybridization of data are listed in Best Practices 59. Rushton et al. (2006) and Beyer et al. (2008) and the references within can provide further guidance on the topics in this section.

Best Practices 59 – Hybridizing data

Policy Decision	Best Practice
Which hybridization method should be used?	If spatial operations are supported, the point-in-polygon method can be used to associate an aggregate value to the geocode calculated from the values within the polygon. If spatial operations are not supported, relational joins should be used to match the geocode to values in the secondary data based on shared keys between the two.
When, how, and which secondary data should be spatially associated with geocoded data (i.e., creating hybrid data)	Hybrid georeferenced data should not be considered certain if the uncertainty in the geocoded record is larger than the distance to the nearest boundary.
When and how should the certainty of hybrid data be calculated?	Hybrid georeferenced data should not be considered certain if the uncertainty in the geocoded record is larger than the distance to the nearest boundary.
What metadata should be maintained?	When geocoded locations are associated with other attributes through hybrid georeferencing, the distance to the closest boundary should be included with the record.

25.2 GEOCODING IMPACTS ON INCIDENCE RATES

When determining incidence rates, cancer registries need to be able to "spatially match" the geographic resolution/accuracy of the numerators with the denominators. In this case, the numerators are the case counts and the denominators are the corresponding population counts. Mismatches can and frequently do occur between these two data sources, which can have dramatic effects capable of either erroneously inflating or deflating the observed incidence rates. For instance, it is possible for a cancer registry to have a very accurate case location (e.g., based on geocoding with a satellite map), but then come to an incorrect conclusion in the analysis of incidence rates because the denominator is based on different (less accurate) geography. Care should be taken to ensure that the geographic resolutions/accuracies of the data used to create the cases (numerators), and the geographic resolution/accuracy of the denominator are derived at commensurate scales. For the cases, it should further be noted that resolution/accuracy needs to be considered both in terms of that of the

underlying of the geocoding reference dataset (e.g., TIGER/Lines [United States Census Bureau 2008d] vs. Tele Atlas [2008c]), and that of the resulting geocoded spatial output (e.g., USPS ZIP Code centroid match vs. satellite imagery manual placement). Best practices relating to the calculation of incidence rates are listed in Best Practices 60.

Best Practices 60 – Incidence rate calculation

Policy Decision	Best Practice
At what resolutions should reference locations and geocodes be used to calculate incidence rates?	Incidence rates should only be calculated when the geocode and reference locations are at the same resolution. Incidence rates should be calculated only after considering: • Appropriateness of geographic area of analysis • Issues of confidentiality and data suppression • Tenure of residence

25.3 IMPLICATIONS OF AGGREGATING UP

When the output spatial location from the geocoding process is derived from a reference feature composed of a significant amount of area (e.g., a city), its use in spatial analyses performed with smaller areal units can create problems with the reliability of the research results. Data gathered at one scale and applied at another may be affected by the modifiable areal unit problem (MAUP), which is referred to often in the geography and GIS literatures (e.g., Openshaw 1984, Grubesic and Murray 2004, Gregorio et al. 2005, Zandbergen and Chakraborty 2006). This should be taken into consideration when considering the validity of spatial analyses. Best practices relating to the MAUP are listed in Best Practices 61.

Best Practices 61 – MAUP

Policy Decision	Best Practice
When does the MAUP need to be accounted for in spatial analysis using hybridized geocode data?	The MAUP may need to be accounted for when data gathered at one scale are applied at another.

This page is left blank intentionally.

26. ENSURING PRIVACY AND CONFIDENTIALITY

This section details the reasons for ensuring privacy and confidentiality in the geocoding process and identifies key approaches that help to facilitate these outcomes.

26.1 PRIVACY AND CONFIDENTIALITY

Patient privacy and confidentiality need to be foremost concerns in any health-based data collection or research study. At no point during any of the processes undertaken as part of a geocoding endeavor should a patient ever be identifiable to anyone other than the staff members who have been trained to deal with such data. However, both the input and output data used in the geocoding process are necessarily identifiable information specifically because of what they encode—a location related to a person. The simple act of issuing a query to a geocoding process reveals both these input and output data, so great care needs to be taken to assure that this is done in a secure fashion. An excellent and extremely detailed survey of privacy issues directly related to cancer registries is available in Gittler (2008a) and the extensive references within, and a state-by-state breakdown of applicable statutes and administrative regulations can be found in Gittler (2008b).

Cancer registries already have existing policies in place to securely collect, store, and transmit patient data both within the organization and to outside researchers. These practices also should be applied to every aspect of the geocoding process. The difficulty in actually doing this, however, should be evident from the basic nature of the geocoding process. Geocoding is an extraordinarily complex process, involving many separate components working in concert to produce a single result. Each of these components and the interactions between them need to be subject to these same security constraints. The simplest case is when the geocoding process is performed in a secure environment at the registry, behind a firewall with all components of the process kept locally, under the control of registry staff. In this scenario, because everything needed is at one location and under the control of a single entity, it is possible to ensure that the flow of the data as it moves through the geocoding process never leaves a secure environment. The cost of this level of security equates to the cost of gathering data and bringing it behind the firewall. A differentiation needs to be made between an **information leak,** which is breach of privacy that may be overwritten with public health law and unavoidable in interest of public health, and a **breach of confidentiality,** which is an outright misuse of the data.

If certain aspects of the geocoding process involve third parties, these interactions need to be inspected carefully to ascertain any potential security concerns. For instance, if address validation is performed by querying a county assessor's database outside of the secure environment of the registry, sensitive information may be exposed. Every Web server keeps logs of the incoming requests including the form parameters (which in this case will be a patient address) as well as the source (the Internet protocol [IP] address of the machine requesting the service). These IP addresses are routinely reversed (e.g., using "whois" tools such as those provided by the American Registry of Internet Names and Numbers [2008]) to determine the issuing organization (the registry) for usage and abuse monitoring of their services. It would not take a great leap for someone to conclude that when a registry queries for an

address, they are using it to determine information about people with a disease of some kind. Any time that data are transferred from the registry to another entity, measures need to be taken to ensure security, including at a minimum encrypting the data before transmitting it securely (e.g., hypertext transfer protocol over secure socket layer, or HTTPS). Best practices relating to geocoding process auditing are listed in Best Practices 62.

Best Practices 62 – Geocoding process privacy auditing when behind a firewall

Policy Decision	Best Practice
When and how should geocoding be audited for information leakages and/or breaches of confidentiality?	The geocoding process and all related components should be audited for information leakages and/or breaches of confidentiality any time any component is changed. Information workflow diagrams and organizational charts should be used to determine where, how, and between which organizations information flows as it is processed by a geocoding system from start to finish. At any point where the information leaves the registry, exactly what information is being transmitted and in what manner should be carefully examined to determine if it can be used to identify the patient or tumor.
When should private, confidential, or identifiable information about a patient or tumor be released during the geocoding process?	As little private, confidential, or identifiable information about a patient or tumor as possible should be released only in a secure fashion during the geocoding process.

Upon receiving a query to geocode an address from a cancer registry, the service could reasonably assume that this specific address has something to do with a disease. In this case, the verification service would essentially be receiving a list of the addresses of cancer patients, and this registry would be exposing identifiable health information. Several simple measures can be taken to avoid this, but the specific practices used may depend on the local polices as to what is allowed versus what is not allowed. As a first example, cancer case addresses submitted to address-verification services can be intermixed with other randomly generated, yet perfectly valid addresses (as well as other invalid addresses such that patient records would not be the only ones that were invalid). This is a simple example of ensuring ***k*-anonymity** (Sweeney 2002), where in this case *k* equals the number of real addresses plus the number of fake addresses (which may or may not be an acceptable level). However, this may increase the cost of the service if the registry is charged on a per-address basis. Alternatively, a registry could require the third party to sign a confidentiality agreement to protect the address data transmitted from the registry.

If a registry uses a third party for any portion of the geocoding process, it needs to ensure that this outside entity abides by and enforces the same rigorous standards for security as employed by the registry. These assurances should be guaranteed in legal contracts and

the contractors should be held accountable if breaches in security are discovered (e.g., financial penalties for each disclosure). Example assurance documents are provided in Appendix A. These data also should be transmitted over a secure Internet connection. Free services such as the geocoding application programmer interfaces, or APIs now offered by search engines competing in the **location-based service** market (e.g., Google, Inc. 2008c, Yahoo!, Inc. 2008) cannot, and most likely will not, be able to honor any of these assurances. In these cases, the registry itself must take measures to assure the required level of security such as those previously mentioned (e.g., submitting randomized bundles of erroneous as well as real data). Best practices related to third-party processing are listed in Best Practices 63.

Best Practices 63 – Third-party processing (external processing)

Policy Decision	Best Practice
When and how can geocoding be performed using external sources without revealing private, confidential, or identifiable information about a patient or tumor?	If geocoding, or any portion thereof (e.g., address validation) is performed using external sources, all information except the minimum necessary attributes (e.g., address attributes) should be removed from records before they are sent for processing.
When and how should third-party processing confidentiality requirements be determined?	Any third party that processes records must agree to ensure the same confidentiality standards used at the registry before data are transmitted to them.
How can and should data be transmitted to third-party processors?	Confidential postal mail or secure network connections should be used to transmit data that have been encrypted from registries to third-party processing services.
How can and should data be submitted to third-party processing services that cannot ensure confidentiality?	The data should be submitted as part of a randomized set of data, or some other method that ensures *k*-anonymity.

Registries also should be cautious of the logs maintained by any geocoding process. Typically, for performance and tracking of query based services, (e.g., Web sites and databases), logs of performance-related metrics are generated so that the processes can be optimized. These logs potentially could be used to recreate the queries or portions thereof that created them, essentially reverse engineering the original input. Here, the policies of the registry can and should enforce that after an input query has been submitted and a result returned, no trace of the query should be retained. Although information and statistics on query performance are useful for improving the geocoding process, this information represents a patient's location, and needs to be securely discarded or made de-identified in some fashion. Best practices relating to log files are listed in Best Practices 64.

Best Practices 64 – Geocoding process log files

Policy Decision	Best Practice
When should log files containing records of geocoding transactions be cleared?	Log files containing any private, confidential, or identifiable information about a patient or tumor should be deleted immediately after the geocode is produced.

At the other end of the geocoding process, once spatial locations have been securely produced, confidentiality still must be ensured regarding these spatial locations once they are visualized. Recent research has shown that geocodes can be used to determine the addresses from which they were derived (a process known as **reverse geocoding** [Brownstein et al. 2006, Curtis et al. 2006]). Researchers need to ensure that the methods they use to display their data (e.g., their presentation on a dot map) do not lend themselves to this type of reverse geocoding, from which the input locational data can be derived from the spatial output.

To accomplish this, several methods of geographic masking have been described in the literature (for detailed reviews and technical approaches see Armstrong et al. 1999, Zimmerman et al. 2008, and Chen et al. 2008). One method is to use aggregation to place the spatial location into a larger class of data from which it cannot be individually recognized. Alternatively, one could use a random offset to move the true spatial location a random distance in a random direction. Other methods exist as well, and all serve the same purpose—protecting the privacy of the individual patient. This, of course, comes at the expense of actually introducing spatial error or uncertainty into the data from which the study results will be derived.

Commonly, most researchers limit geographic masking of geocodes to within either county or CBG boundaries to develop meaningful conclusions. Thus, they want to ensure that the geographically masked point and the original point are within the same county/CBG. This presents additional security concerns because in many counties, the universe of address points is quite commonly available and downloadable as GIS data. It is possible to develop geographic masking procedures that meet standards for *k*-anonymity as measured against the universe of all known address points. See Sweeney (2002) for how guidance on this can be accomplished.

Depending on how much of the data actually need to be displayed, a researcher could first use the accurate data to develop their results, and then use geographically masked data in the presentation of the results. It also is possible to use geographically masked data for analyses as well, but this approach requires making the analysis scale larger. Best practices related to masking geocoded data are listed in Best Practices 65.

Best Practices 65 – Geographic masking

Policy Decision	Best Practice
How and when should geocoded data be masked to ensure the required levels of confidentiality?	Any time that geocoded data are visualized (e.g., displayed on a map, Web site, or in a presentation), they should be masked to conform to the confidentiality requirements of the registry.
How can and should geocoded data be masked when visualized?	Any proven and suitable method that accomplishes the required level of confidentiality can be used: • Randomization • Aggregation • Randomized offset

It seems obvious, but is worth mentioning, that once a registry has released data to a researcher, the researcher can/may misuse the data in a fashion that the registry did not intend, even after Institutional Review Board approval. To prevent these types of abuses (and protect themselves), registries should follow the best practices listed in Best Practices 66, as well as make use of research assurance documents such as those in Appendix A.

Best Practices 66 – Post-registry security

Policy Decision	Best Practice
How should registries be involved with researchers to ensure proper use of geocoded data?	Registries should work closely with researchers to be aware of the subsequent results/publications. Registries should provide guidance to researchers on how to securely store, transfer, use, report, and visualize geocoded data.
How can registries protect themselves from the liability of researchers misusing geocoded data?	Registries should have an assurance document that details the appropriate use of the geocoded data and that researchers must initial prior to release of the data.

This page is left blank intentionally.

GLOSSARY OF TERMS

Absolute Geocode: An absolute known geographic (spatial) location or an offset from an absolute known location.

Absolute Input Data: See Absolute Locational Description.

Absolute Locational Description: A description which, by itself, contains enough information to produce an output geographic location (e.g., locations described in terms of linear addressing systems).

Acceptable Match Rate: The match rate value a geocoding process must meet such that the geocoded data can be considered valid for use in a research study.

Accuracy: A measure of how close to a true value something is.

Actual Lot Feature Interpolation: A linear-based feature interpolation algorithm that is not subject to the parcel homogeneity assumption or parcel existence assumption.

Address Ambiguity: With regard to feature matching, the case when a single input address can possibly match to multiple reference features.

Address Matching: A specialized case of feature matching, strictly dealing with matching postal street addresses to features in the reference data source, usually TIGER type street segments or areal unit delineations (Census delineations, USPS ZIP Code delineations, etc.).

Address Normalization: The process of organizing and cleaning address data to increase efficiency for data use and sharing.

Address Parity: An indication of which side of a street an address falls, even or odd (assumes binary address parity for a street segment).

Address Range: Aspatial attributes associated with a linear-based reference data (i.e., street vectors), describing the valid range of addresses on the street.

Address Range Feature Interpolation: A linear-based feature interpolation algorithm that is subject to the parcel existence, parcel homogeneity, and parcel extent assumptions.

Address Standardization: The process of converting an address from a normalized format into a specified format (e.g., United States Postal Service [2008d], United States Federal Geographic Data Committee [2008b]).

Address Validation: The process of determining whether or not an address actually exists.

Administrative Unit: An administratively defined level of geographic aggregation. Typical units include (from largest to smallest): county, state, county, major county subdivision, city, neighborhood, census tract, census block, and parcel.

Advanced Match Rate: A match rate measure that normalizes the number of records attempted by removing those that are outside of the geographic boundaries of the entity performing the geocoding.

Approximate Geocodes: See Pseudocodes.

Arc: See Edge.

Areal Unit: See Polygon.

Areal Unit-Based Data: See Polygon-Based Data.

Areal Unit-Based Feature Interpolation: A feature interpolation algorithm that uses a computational process (e.g., geographic centroid) to determine a suitable output from the spatial geometry of polygon-based reference features (e.g., parcel).

Areal Unit-Based Reference Dataset: See Polygon-Based Reference Dataset.

Aspatial: Data or attributes describing data that do not refer to spatial properties.

Atomic-Level Match Rate: A match rate associated with an individual attribute of an input address.

Atomic Metrics: Metrics that describe the characteristics of individual members of a set.

Attribute Constraints: With regard to SQL-like feature matching, one or more predicates that limit the reference features returned from a reference dataset in response to a query.

Attribute Imputation: With regard to feature matching, the process of imputing missing attribute for an input address values using other known variables about the address, region, and/or individual to which it belongs.

Attribute Relaxation: Part of the feature-matching process. The procedure of easing the requirement that all street address attributes (street number, name, pre-directional, post-directional, suffix, etc.) must exactly match a feature in the reference data source to obtain a matching street feature, thereby increasing the probability of finding a match, while also increasing the probability or error. This is commonly performed by removing or altering street address attributes in an iterative manner using a predefined order.

Attribute Weighting: A form of probabilistic feature matching in which probability-based values are associated with each attribute and either subtract from or add to the composite score for the feature as a whole.

Backus-Naur Form (BNF): A notation used to construct a grammar describing the valid components of an object (e.g., an address).

Batch-Geocoding: A geocoding process that operates in an automated fashion and processes more than a single input datum at a time.

Best Practice: A policy or technical decision that is recommended but not required.

Binary Address Parity: The case when all addresses on one side of a street segment are even and all addresses on the other side of the street segment are even.

Blocking Scheme: Method used in record linkage to narrow the set of possible candidate values that can match an input datum.

Breach of Confidentiality: An outright misuse of confidential data, including its unauthorized release, or its use for an unintended purpose.

Cancer Registry: A disease registry for cancer.

Case Sensitivity: Whether or not a computational system differentiates between the case of alphabetic characters (i.e., upper-case and lower-case).

Character-Level Equivalence: A string comparison algorithm that enforces character-by-character equivalence between two or more strings.

City-Style Postal Address: A postal address that describes a location in terms of a numbered building along a street with more detailed attributes defining higher resolution within structures or large area geographic features possible.

Composite Feature Geocoding: A geocoding process capable of creating and utilizing composite features in response to an ambiguous feature-matching scenario.

Composite Feature: A geographic feature created through the union of two or more disparate geographic features (e.g., a bounding box encompassing all of the geographic features).

Composite Score: The overall score of a reference feature being a match to an input datum resulting from the summation of the individually weighted attributes.

Conditional Probability: The probability of something occurring, given that other information is known.

Confidence Interval: The percentage of the data that are within a given range of values.

Confidence Threshold: The level of certainty above which results will be accepted and below which they will be rejected.

Context-Based Normalization: An address normalization method that makes use of syntactic and lexical analysis.

Continuous feature: With regard to geocoding reference features, a geographic feature that corresponds to more than one real-world entity (e.g., a street segment).

Coordinate System: A system for delineating where objects exist in a space.

Corner Lot Assumption: See Corner Lot Problem.

Corner Lot Problem: The issue arising during linear-based feature interpolation that when using a measure of the length of the segment for interpolation it is unknown how much real estate may be taken up along a street segment by parcels from other intersecting street segments (around the corner), and the actual street length may be shorter than expected.

Data Source: With regard to SQL-like feature matching, the relational table or tables of the reference dataset that should be searched.

Deterministic Matching Algorithm: A matching algorithm based on a series of predefined rules that are processed in a specific order.

Discrete feature: With regard to geocoding reference features, a geographic feature that corresponds to a single real-world entity (e.g., a single address point).

Disease Registry: Organizations that gather, store, and analyze information about a disease for their area.

Dropback: The offset used in linear-based interpolation to move from the centerline of the street vector closer to the face of the parcel.

Edge: The topological connection between nodes in a graph.

Empirical Geocoding: See Geocode Caching.

Error Rate: With regard to probabilistic feature matching, denotes a measure of the instances in which two address attributes do not match, even though two records do.

Essence-Level Equivalence: A string comparison algorithm that determines if two or more strings are "essentially" the same (e.g., a phonetic algorithm).

False Negative: With regard to feature matching, the result of a true match being returned from the feature-matching algorithm as a non-match.

False Positive: With regard to feature matching, the result of a true non-match being returned from the feature-matching algorithm as a match.

Feature: An abstraction of a real-world phenomenon.

Feature Disambiguation: With regard to feature matching, the process of determining the correct feature that should be matched out of a set of possible candidates using additional information and/or human intuition.

Feature Interpolation: The process of deriving an output geographic feature from a geographic reference features.

Feature Interpolation Algorithm: An algorithm that implements a particular form of feature interpolation.

Feature Matching: The process of identifying a corresponding geographic feature in the reference data source to be used to derive the final geocode output for an input.

Feature-Matching Algorithm: An algorithm that implements a particular form of feature matching.

Footprint: See Geographic Footprint.

Gazetteer: A digital data structure that is composed of a set of geographic features and maintains information about their geographic names, geographic types, and geographic footprints.

Generalized Match Rate: A match rate measure that normalizes the number of records attempted by removing those that could never be successfully geocoded.

Geocode (n.): A spatial representation of locational descriptive locational text.

Geocode (v.): To perform the process of geocoding.

Geocode Caching: Storing and re-using the geocodes produced for input data from previous geocoding attempts.

Geocoder: A set of inter-related components in the form of operations, algorithms, and data sources that collaboratively work together to produce a spatial representation for aspatial locationally descriptive text.

Geocoding: The act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process.

Geocoding Algorithm: The computational component of the geocoder that determines the correct reference dataset feature based on the input data and derives a spatial output.

Geocoding Data Consumer-Group: The group in the geocoding process that utilizes geocoded data (e.g., researchers).

Geocoding General Interest-Group: The group in the geocoding process that has a general interest in the geocoding process (e.g., the general public).

Geocoding Practitioner-Group: The group in the geocoding process that performs the task of the actual geocoding of input data.

Geocoding Process Designer-Group: The group in the geocoding process that is in charge of making policy decisions regarding how geocoding will be performed.

Geocoding Requirements: The set of limitations, constraints, or concerns that influence the choice of a particular geocoding option. These may be technical, budgetary, legal, and/or policy.

Geographical Bias: The observation that the accuracy of geocoding strategy may be a function of the area in which the geocode resides.

Geographic Coordinate System: A coordinate system that is a representation of the Earth as an ellipsoid.

Geographic Feature: A feature associated with a location relative to the Earth.

Geographic Footprint: A spatial description representing the location of a geographic feature on Earth.

Geographic Information System: A digital system that stores, displays, and allows for the manipulation of digital geographic data.

Geographic Location: See Geographic Feature.

Geographic Name: A name that refers to a geographic feature.

Geographic Type: A classification that describes a geographic feature taken from an organized hierarchy of terms.

Gold Standard: Data that represent the true state of the world.

Georeference: To transform non-geographic information (information that has no geographically valid reference that can be used for spatial analyses) into geographic information (information that has a valid geographic reference that can be used for spatial analyses).

Georeferenced: Geographic information that was originally non-geographic and has been transformed into it.

Georeferencing: The process of transforming non-geographic information into geographic information.

GIS Coordinate Quality Codes: A hierarchical scheme of qualitative codes that indicate the quality of a geocode in terms of the data it represents.

Global Positioning System: The system of satellites, calibrated ground stations, and temporally based calculations used to obtain geographic coordinates using digital receiver devices.

Grammar: An organized set of rules that describe a language.

Graph: A topologically connected set of nodes and edges.

Highway Contract Address: See Rural Route Address.

Hierarchical Geocoding: A feature-matching strategy that uses geographic reference features of progressively lower and lower accuracy.

Holistic-Level Match Rate: A match rate associated with an overall set of input address data.

Holistic Metrics: Metrics that describe the overall characteristics of a set.

Hybrid Data: Data that are created by the intersection of two datasets along a shared key; this key can be spatial, as in a geographic intersection, or aspatial as in a relational linkage in a relational database.

Hybrid Georeferencing: The procedure of associating relevant data (spatial or aspatial) with geocoded data.

Information Leak: The release of information to individuals outside of the owning institution; may be a breach of privacy that is overwritten by public health laws and unavoidable in the interest of public health.

In-House Geocoding: When geocoding is performed locally, and not sent to a third party.

Input Data: The non-spatial locationally descriptive texts that is to be turned into spatial data by the process of geocoding.

Interactive Geocoding: When the geocoding process allows for intervention when problems or issues arise.

Interactive Matching Algorithm: See Interactive Geocoding.

Interactive-Mode Geocoding: See Interactive Geocoding.

K-Anonymity: The case when a single individual cannot be uniquely identified out of at least *k* other individuals.

Latitude: The north-south axis describing a location in the Geographic Coordinate System.

Lexical Analysis: The process by which an input address is broken up into tokens.

Line: A 1-D geographic object having a length and is composed of two or more 0-D point objects.

Linear-Based Data: Geographic data that is based upon lines.

Linear-Based Feature Interpolation: A feature interpolation algorithm that operates on lines (e.g., street vectors) and produces an estimate of an output feature using a computational process on the geometry of the line (e.g., estimation between the endpoints).

Line-Based Data: See Linear-Based Reference Dataset.

Linear-Based Reference Dataset: A reference dataset that is composed of linear-based data.

Location-Based Service: A service that is provided based upon the geography of the client.

Longitude: The east-west axis describing a location in the Geographic Coordinate System.

Lookup Tables: With regard to substitution-based address normalization, is a set of known normalized values for common address tokens, (e.g., those from the United States Postal Service [2008d]).

Machine Learning: The subfield of computer science dealing with algorithms that induce knowledge from data.

Mapping Function: The portion of the address standardization algorithm that translates between an input normalized form and the target output standard.

Match: With regard to substitution-based address normalization, is the process of identifying alias for an input address token within a lookup table of normalized values.

Matching Ability: With regard to a reference dataset, a measure of its ability to match an input address while maintaining the same consistent matching criteria as applied to other reference datasets.

Match Probability: With regard to probabilistic feature matching, the degree of belief, ranging from 0 to 1, that a feature matches.

Matched Probability: With regard to probabilistic feature matching, the probability that the attribute values of an input datum and a reference feature matching when the records themselves match.

Match Rate: A measure of the amount of input data that were able to be successfully geocoded (i.e., assigned to an output geocode).

Matching Rules: With regard to substitution-based address normalization, are the set of rules that determine the valid associations between an input address token and the normalized values in a lookup table.

Metadata: Descriptions associated with data that provide insight into attributes about it (e.g., lineage).

NAACCR Data Standard: A mandatory (required) data formatting and/or content scheme for a particular data item.

Network: A topologically connected graph of nodes and edges.

Node: The endpoints in a graph that are topologically connected together by edges.

North American Association of Central Cancer Registries (NAACCR): A professional organization that develops and promotes uniform data standards for cancer registration; provides education and training; certifies population-based registries; aggregates and publishes data from central cancer registries; and promotes the use of cancer surveillance data and systems for cancer control and epidemiologic research, public health programs, and patient care to reduce the burden of cancer in North America.

Non-Binary Address Parity: The case when addresses along one side of a street segment can be even, odd, or both.

Non-Interactive Matching Algorithm: See Batch Geocoding.

Non-Spatial: See aspatial

Output Data: The valid spatial representations returned from the geocoder derived from features in the reference dataset.

Parcel Existence Assumption: The assumption used in linear-based feature interpolation that all parcels associated with the address range of a reference features exist.

Parcel Extent Assumption: The assumption used in linear-based feature interpolation that parcels associated with the address range of a reference features start immediately at the beginning of the segment and fill all space to the other end.

Parcel Homogeneity Assumption: The assumption used in linear-based feature interpolation that all parcels associated with the address range of a reference features have the same dimensions.

Parse Tree: A data structure representing the decomposition of an input string into its component parts.

Pass: With regard to attribute relaxation, the relaxation of a single address attribute within a step that does not result in a change of geographic resolution.

Per-Record-Geocoding: A geocoding process that processes a geocode for single input datum at a time; it may either be automated or not.

Place Name: See Geographic Name.

Phonetic Algorithm: A string comparison algorithm that is based on the way that a string is pronounced.

Point: A 0-dimensional (0-D) object that has a position in space but no length.

Point-Based Data: Geographic data that are based upon point features.

Point-Based Reference Dataset: A reference dataset that is composed of point-based data.

Point-Based Feature Interpolation: A feature interpolation algorithm that operates on points (e.g., a database of address points) and simply returns the reference feature point as output.

Point-In-Polygon Association: The process of spatially intersecting a geographic feature that is a point with another geographic feature that is areal unit-based such that the attributes of the areal unit can be associated with the point or vice versa.

Polygon: A geographic object bounded by at least three 1-D line objects or segments with the requirement that they must start and end in the same location (i.e., node).

Polygon-Based Data: Geographic data that is based upon polygon features.

Polygon-Based Reference Dataset: A reference dataset that is composed of polygon features.

Polyline: A geographic object that is composed of a series of lines.

Porter Stemmer: A word-stemming algorithm that works by removing common suffixes and applying substitution rules.

Postal Address: A form of input data describing a location in terms of a postal addressing system.

Postal Code: A portion of a postal address designating a region.

Postal Street Address: A form of input data containing attributes that describe a location in terms of a postal street addressing system (e.g., USPS street address).

Postal ZIP Code: The USPS address portion denoting the route a delivery address is on.

Post Office Box Address: A postal address designating a storage location at a post office or other mail-handling facility.

Precision: Regarding information retrieval, a measure of how correct the data retrieved are.

Predicate: See Query Predicate.

Probability-Based Normalization: An address normalization method that makes use probabilistic methods (e.g., support vector machines or Hidden Markov Models).

Prior Probability: See Unconditional Probability.

Projection: A mathematical function to transfer positions on the surface of the Earth to their approximate positions on a flat surface.

Pseudocoding: The process deriving pseudocodes using a deterministic or probabilistic method.

Pseudocodes: An approximation of a true geocode.

Query Predicate: In an SQL query, the list of attribute-value pairs indicating which attributes of a reference features must contain what values for it to be returned.

Raster-Based Data: Data that divide the area of interest into a regular grid of cells in some specific sequence, usually row-by-row from the top left corner; each cell is assigned a single value describing the phenomenon of interest.

Real-Time Geocoding: With regard to patient/tumor record geocoding upon intake, is the process of geocoding a patient/tumor record while the patient is available to provide more detailed or correct information using an iterative refinement approach.

Recall: Regarding information retrieval, a measure of how complete the data retrieved are.

Record Linkage: A sub-field of computer science relating to finding features in two or more datasets which are essentially referring to the same feature.

Reference Dataset: The underlying geographic database containing geographic features the geocoder uses to derive a geographic output.

Reference Data Source: See Reference Dataset.

Reference Set: With regard to record linkage, refers to the set of candidate features that may possibly be matched to an input feature.

Registry: See Disease Registry.

Relative Geocode: Geographic (spatial) location that is relative to some other reference geographic feature.

Relative Input data: See Relative Locational Description.

Relative Locational Description: A description which, by itself, does not contain enough information to produce an output geographic location (e.g., locations described in terms directions from some other features).

Relative Predicted Certainty: A relative quantitative measure of the area of the accuracy of a geocode based on information about how a geocode is produced.

Reverse Geocoding: The process of determining the address used to create a geocode from the output geographic location.

Rural Route Address: A postal address identifying a stop on a postal delivery route.

Scrubbing: The component of address normalization that removes illegal characters and white space from an input datum.

Selection Attributes: With regard to SQL-like feature matching, the attributes of the reference feature that should be returned from the reference dataset in response to a query.

Simplistic Match Rate: A match rate measure computed as the number of input data successfully assigned a geocode divided by the number of input data attempted.

Situs Address: The actual physical address associated with the parcel.

Socioeconomic Status: Descriptive attributes associated with individuals or groups referring to social and economic variables.

Software-Based Geocoding: A geocoding process in which a significant portion of the components are software systems.

Soundex: A phonetic algorithm that encodes the sound of a string using a series of character removal and substitution rules from a known table of values.

Spatial Accuracy: A measure of the correctness of a geographic location based on some metric; can be qualitative or quantitative.

Spatial Resolution: A measure describing the scale of geographic data; can be qualitative or quantitative.

Stemming: See Word Stemming.

Step: With regard to attribute relaxation, the relaxation of a multiple address attributes at once that results in a change of geographic resolution.

Street Address Ambiguity: With regard to feature matching, the case when a single input address can possibly match to multiple addresses along a street reference feature.

Street Network: A linear-data graph structure with edges representing streets and nodes representing their intersections.

Street Segment Ambiguity: With regard to feature matching, the case when a single input address can possibly match to multiple street reference features.

String Comparison Algorithms: An algorithm that calculates a similarity measure between two or more input strings.

Sub-Parcel Address Ambiguity: With regard to feature matching, the case when a single input address can possibly match to multiple addresses (e.g., buildings) within a single parcel reference feature.

Substitution-Based Normalization: An address normalization method that makes use of lookup tables for identifying commonly encountered terms based on their string values.

Syntactic Analysis: The process by which tokens representing an input address are placed into a parse tree based on the grammar which defines possible valid combinations.

Temporal Accuracy: A measure of how appropriate the time period the reference dataset represents is to the input data that is to be geocoded; can be qualitative or quantitative.

Temporal Extent: An attribute associated with a datum describing a time period for which it existed, or was valid.

Temporal Staleness: The phenomenon that occurs when data become out-of-date and less accurate after the passage of time (e.g., a geocode cache becoming outdated after a newer more accurate reference dataset becomes available).

Topologically Integrated Geographic Encoding and Referencing (TIGER) Line Files: Series of vector data products distributed by and created to support of the mission the U.S. Census Bureau (United States Census Bureau 2008d).

Token: An attribute of an input address after it has been split into its component parts.

Tokenization: The process used to convert the single complete string representing the whole address into a series of separate tokens.

Topological: Describes the connection between nodes and edges in a graph.

Toponym: See Geographic Name.

True Negative: With regard to feature matching, the result of a true non-match being returned from the feature-matching algorithm as a non-match.

True Positive: With regard to feature matching, the result of a true match being returned from the feature-matching algorithm as a match.

Unconditional Probability: The probability of something occurring, given that no other information is known.

Uniform Lot Feature Interpolation: A linear-based feature interpolation algorithm that is subject to the parcel homogeneity assumption and parcel existence assumption.

United States Bureau of the Census: United States federal agency responsible for performing the Census.

United States Census Bureau: See United States Bureau of the Census.

United States Postal Service (USPS): United States federal agency responsible for mail delivery.

Unmatched Probability: With regard to probabilistic feature matching, the probability that the attribute values of an input datum and a reference feature matching when the records themselves do not match.

Urban and Regional Information Systems Association (URISA): A non-profit professional and educational association that promotes the effective and ethical use of spatial information and information technologies for the understanding and management of urban and regional systems.

Vector: An object with a direction and magnitude, commonly a line.

Vector-Based Data: Geographic data that consist of vector features.

Vector Feature: Phenomena or things of interest in the world around us (i.e., a specific street like Main Street) that cannot be subdivided into phenomena of the same kind (i.e., more streets with new names).

Vector Object: See Vector Feature.

Vertex: See Node.

Waiting It Out: The process of holding off re-attempting geocoding for a period of time until something happens to increase the probability of a successful attempt (e.g., new reference datasets are released).

Weights: With regard to probabilistic feature matching, are numeric values calculated as a combination of matched and unmatched probabilities and assigned to each attribute of an address to denote its level of importance.

Word Stemming: To reduce a word to its fundamental stem.

ZIP Code Tabulation Area: A geographical areal unit defined by the United States Bureau of the Census.

REFERENCES

Abe T and Stinchcomb D 2008 Geocoding Best Practices in Cancer Registries. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 111-126.

Agarwal P 2004 Contested nature of "place": Knowledge mapping for resolving ontological distinctions between geographical concepts. In Egenhofer et al. (eds) *Proceedings of 3rd International Conference on Geographic Information Science (GIScience)*. Berlin, Springer Lecture Notes in Computer Science No 3234: 1-21.

Agouris P, Beard K, Mountrakis G, and Stefanidis A 2000 Capturing and Modeling Geographic Object Change: A SpatioTemporal Gazetteer Framework. *Photogrammetric Engineering and Remote Sensing* 66(10): 1224-1250.

Agovino PK, Niu X, Henry KA, Roche LM, Kohler BA, and Van Loon S 2005 Cancer Incidence Rates in New Jersey's Ten Most Populated Municipalities, 1998-2002. NJ State Cancer Registry Report. Available online at: http://www.state.nj.us/health/ces/documents/cancer_municipalities.pdf. Last accessed April 23rd, 2008.

Alani H 2001 Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science* 15(4): 287-306.

Alani H, Kim S, Millard DE, Weal MJ, Hall W, Lewis PH, and Shadbolt N 2003 Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation. In *Proceedings of Knowledge Capture (K-Cap'03), Workshop on Knowledge Markup and Semantic Annotation.*

Alexandria Digital Library 2008 Alexandria Digital Library Gazetteer. Available online at: http://alexandria.ucsb.edu/clients/gazetteer. Last accessed April 23rd, 2008.

American Registry for Internet Names 2008 WHOIS Lookup. Available online at: http://www.arin.net. Last accessed April 23rd, 2008.

Amitay E, Har'El N, Sivan R, and Soffer A 2004 Web-A-Where: Geotagging Web Content. In Sanderson et al. (eds) *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR '04)*. New York, ACM Press: 273-280.

Arampatzis A, van Kreveld M, Reinbacher I, Jones CB, Vaid S, Clough P, Joho H, and Sanderson M 2006 Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems.*

Arbia G, Griffith D, and Haining R 1998 Error Propagation Modeling in Raster GIS: Overlay Operations. *International Journal of Geographical Information Science* 12(2): 145-167.

Arikawa M and Noaki K 2005 Geocoding Natural Route Descriptions using Sidewalk Network Databases. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration, 2005 (WIRI '05)*: 136-144.

Arikawa M, Sagara T, Noaki K, and Fujita H 2004 Preliminary Workshop on Evaluation of Geographic Information Retrieval Systems for Web Documents. In Kando and Ishikawa (eds) *Proceedings of the NTCIR Workshop 4 Meeting: Working Notes of the Fourth NTCIR Workshop Meeting*. Available online at: http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/WEB/NTCIR4WN-WEB-ArikawaM.pdf. Last accessed April 23rd, 2008.

Armstrong MP, Rushton G, and Zimmerman DL 1999 Geographically Masking Health Data to Preserve Confidentiality. *Statistics in Medicine* 18(5): 497–525.

Armstrong MP, Greene BR, and Rushton G 2008 Using Geocodes to Estimate Distances and Geographic Accessibility for Cancer Prevention and Control. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 181-194.

Armstrong MP and Tiwari C 2008 Geocoding Methods, Materials, and First Steps toward a Geocoding Error Budget. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 11-36.

Axelrod A 2003 On Building a High Performance Gazetteer Database. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 63-68.

Bakshi R, Knoblock CA, and Thakkar S 2004 Exploiting Online Sources to Accurately Geocode Addresses. In Pfoser et al. (eds) *Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS '04)*: 194-203.

Beal JR 2003 Contextual Geolocation: A Specialized Application for Improving Indoor Location Awareness in Wireless Local Area Networks. In Gibbons (ed) *The 36th Annual Midwest Instruction and Computing Symposium (MICS2003)*, Duluth, Minnesota.

Beaman R, Wieczorek J, and Blum S 2004 Determining Space from Place for Natural History Collections: In a Distributed Digital Library Environment. *D-Lib Magazine* 10(5).

Bell EM, Hertz-Picciotto I, and Beaumont JJ 2001 A case-control study of pesticides and fetal death due to congenital anomalies. *Epidemiology* 12(2): 148-56.

Berney LR and Blane DB 1997 Collecting Retrospective Data: Accuracy of Recall After 50 Years Judged Against Historical Records. *Social Science & Medicine* 45(10): 1519-1525.

Beyer KMM, Schultz AF, and Rushton G 2008 Using ZIP Codes as Geocodes in Cancer Research. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 37-68.

Bichler G and Balchak S 2007 Address matching bias: ignorance is not bliss. *Policing: An International Journal of Police Strategies & Management* 30(1): 32-60.

Bilhaut F, Charnois T, Enjalbert P, and Mathet Y 2003 Geographic reference analysis for geographic document querying. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 55-62.

Blakely T and Salmond C 2002 Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 31(6): 1246-1252.

Block R 1995 Geocoding of Crime Incidents Using the 1990 TIGER File: The Chicago Example. In Block et al. (eds) *Crime Analysis Through Computer Mapping*. Washington, DC, Police Executive Research Forum: 15.

Bonner MR, Han D, Nie J, Rogerson P, Vena JE, and Freudenheim JL 2003 Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology* 14(4): 408-411.

Boscoe FP, Kielb CL, Schymura MJ, and Bolani TM 2002 Assessing and Improving Census Track Completeness. *Journal of Registry Management* 29(4): 117–120.

Boscoe FP, Ward MH, and Reynolds P 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).

Boscoe FP 2008 The Science and Art of Geocoding: Tips for Improving Match Rates and Handling Unmatched cases in Analysis. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 95-110.

Boulos MNK 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).

Bouzeghoub M 2004 A framework for analysis of data freshness. In *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*: 59-67.

Bow CJD, Waters NM, Faris PD, Seidel JE, Galbraith PD, Knudtson ML, and Ghali WA 2004 Accuracy of city postal code coordinates as a proxy for location of residence. *International Journal of Health Geographics* 3(5).

Brody JG, Vorhees DJ, Melly SJ, Swedis SR, Drivas PJ, and Rudel RA 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 J 2003 Building and Maintaining a Reliable Enterprise Street Database. In *Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference.*

Brownstein JS, Cassa CA, and Mandl KD 2006 No Place to Hide - Reverse Identification of Patients from Published Maps. *New England Journal of Medicine* 355(16): 1741-1742.

Can A 1993 TIGER/Line Files in Teaching GIS. *International Journal of Geographical Information Science* 7(8): 561-572.

Canada Post Corporation 2008 Postal Standards: Lettermail and Incentive Lettermail. Available online at: http://www.canadapost.ca/tools/pg/standards/pslm-e.pdf. Last accessed April 23rd, 2008.

Casady T 1999 Privacy Issues in the Presentation of Geocoded Data. *Crime Mapping News* 1(3).

Cayo MR and Talbot TO 2003 Positional Error in Automated Geocoding of Residential Addresses. *International Journal of Health Geographics* 2(10).

Chalasani VS, Engebretsen O, Denstadli JM, and Axhausen KW 2005 Precision of Geocoded Locations and Network Distance Estimates. *Journal of Transportation and Statistics* 8(2).

Chavez RF 2000 Generating and Reintegrating Geospatial Data. In *Proceedings of the 5th ACM Conference on Digital Libraries (DL '00)*: 250-251.

Chen CC, Knoblock CA, Shahabi C, and Thakkar S 2003 Building Finder: A System to Automatically Annotate Buildings in Satellite Imagery. In Agouris (ed) *Proceedings of the International Workshop on Next Generation Geospatial Information (NG2I '03)*, Cambridge, Mass.

Chen CC, Knoblock CA, Shahabi C, Thakkar S, and Chiang YY 2004 Automatically and Accurately Conflating Orthoimagery and Street Maps. In Pfoser et al. (eds) *Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems (ACMGIS '04)*, Washington DC: 47-56.

Chen MF, Breiman RF, Farley M, Plikaytis B, Deaver K, and Cetron MS 1998 Geocoding and Linking Data from Population-Based Surveillance and the US Census to Evaluate the Impact of Median Household Income on the Epidemiology of Invasive Streptococcus Pneumoniae Infections. *American Journal of Epidemiology* 148(12): 1212-1218.

Chen W, Petitti DB, and Enger S 2004 Limitations and potential uses of census-based data on ethnicity in a diverse community. *Annals of Epidemiology* 14(5): 339-345.

Chen Z, Rushton G, and Smith G 2008 Preserving Privacy: Deidentifying Data by Applying a Random Perturbation Spatial Mask. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 139-146.

Chiang YY and Knoblock CA 2006 Classification of Line and Character Pixels on Raster Maps Using Discrete Cosine Transformation Coefficients and Support Vector Machines. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*.

Chiang YY, Knoblock CA, and Chen CC 2005 Automatic extraction of road intersections from raster maps. In *Proceedings of the 13th annual ACM International Workshop on Geographic Information Systems*: 267-276.

Chou YH 1995 Automatic Bus Routing and Passenger Geocoding with a Geographic Information System. In *Proceedings of the 1995 IEEE Vehicle Navigation and Information Systems Conference*: 352-359.

Christen P and Churches T 2005 A Probabilistic Deduplication, Record Linkage and Geocoding System. In *Proceedings of the Australian Research Council Health Data Mining Workshop (HDM05)*, Canberra, AU. Available online at: http://acrc.unisa.edu.au/groups/health/hdw2005/Christen.pdf. Last accessed April 23rd, 2008.

Christen P, Churches T, and Willmore A 2004 A Probabilistic Geocoding System Based on a National Address File. In *Proceeding of the Australasian Data Mining Conference*, Cairns, AU. Available online at: http://datamining.anu.edu.au/publications/2004/aus-dm2004.pdf. Last accessed April 23rd, 2008.

Chua C 2001 An Approach in Pre-processing Philippine Address for Geocoding. In *Proceedings of the Second Philippine Computing Science Congress (PCSC 2001)*.

Chung K, Yang DH, and Bell R 2004 Health and GIS: Toward Spatial Statistical Analyses. *Journal of Medical Systems* 28(4): 349-360.

Churches T, Christen P, Lim K, and Zhu JX 2002 Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models. *Medical Informatics and Decision Making* 2(9).

Clough P 2005 Extracting Metadata for Spatially-Aware information retrieval on the internet. In Jones and Purves (eds) *Proceedings of the 2005 ACM Workshop of Geographic Information Retrieval (GIR'05)*: 17-24.

Cockburn M, Wilson J, Cozen W, Wang F, and Mack T 2008 Residential proximity to major roadways and the risk of mesothelioma, lung cancer and leukemia. Under review, personal communication.

Collins SE, Haining RP, Bowns IR, Crofts DJ, Williams TS, Rigby AS, and Hall DM 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.

County of Sonoma 2008 Vector Data - GIS Data Portal - County of Sonoma. Available online at: https://gis.sonoma-county.org/catalog.asp. Last accessed April 23rd, 2008.

Cressie N and Kornak J 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 CM 2003 Public Health GIS and the Internet. *Annual Review of Public Health* 24: 57-82.

Curtis AJ, Mills JW, and Leitner M 2006 Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* 5(44).

Dao D, Rizos C, and Wang J 2002 Location-based services: technical and business issues. *GPS Solutions* 6(3): 169-178.

Davis Jr. CA 1993 Address Base Creation Using Raster/Vector Integration. In *Proceedings of the URISA 1993 Annual Conference*, Atlanta, GA: 45-54.

Davis Jr. CA and Fonseca FT 2007 Assessing the Certainty of Locations Produced by an Address Geocoding System. *GeoInformatica* 11(1): 103-129.

Davis Jr. CA, Fonseca FT, and De Vasconcelos Borges, KA 2003 A Flexible Addressing System for Approximate Geocoding. In *Proceedings of the Fifth Brazilian Symposium on GeoInformatics (GeoInfo 2003)*, Campos do Jordão, São Paulo, Brazil.

Dawes SS, Cook ME, and Helbig N 2006 Challenges of Treating Information as a Public Resource: The Case of Parcel Data. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*: 1-10.

Dearwent SM, Jacobs RR, and Halbert JB 2001 Locational Uncertainty in Georeferencing Public Health Datasets. *Journal of Exposure Analysis Environmental Epidemiology* 11(4): 329-334.

Densham I and Reid J 2003 A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 80-81.

Diez-Roux AV, Merkin SS, Arnett D, Chambless L, Massing M, Nieto FJ, Sorlie P, Szklo M, Tyroler HA, and Watson RL 2001 Neighborhood of Residence and Incidence of Coronary Heart Disease. *New England Journal of Medicine* 345(2): 99-106.

Dijkstra EW 1959 A note on two problems in connexion with graphs. *Numerische Mathematik* 1: 269-271.

Dru MA and Saada S 2001 Location-based mobile services: the essentials. *Alcatel Telecommunications Review* 1: 71-76.

Drummond WJ 1995 Address Matching: GIS Technology for Mapping Human Activity Patterns. *Journal of the American Planning Association* 61(2): 240-251.

Dueker KJ 1974 Urban Geocoding. *Annals of the Association of American Geographers* 64(2): 318-325.

Durbin E, Stewart J, Huang B 2008 Improving the Completeness and Accuracy of Address at Diagnosis with Real-Time Geocoding. Presentation at *The North American Association of Central Cancer Registries 2008 Annual Meeting*, Denver, CO.

Durr PA and Froggatt AEA 2002 How Best to Georeference Farms? A Case Study From Cornwall, England. *Preventive Veterinary Medicine* 56: 51-62.

Efstathopoulos P, Mammarella M, and Warth A 2005 The Meta Google Maps "Hack". Unpublished Report. Available online at: http://www.cs.ucla.edu/~pefstath/papers/google-meta-paper.pdf. Last accessed April 23rd, 2008.

Eichelberger P 1993 The Importance Of Addresses: The Locus Of GIS. In *Proceedings of the URISA 1993 Annual Conference*, Atlanta, GA: 212-213.

El-Yacoubi MA, Gilloux M, and Bertille JM 2002 A Statistical Approach for Phrase Location and Recognition within a Text Line: An Application to Street Name Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(2): 172-188.

Environmental Sciences Research Institute 1999 GIS in Health Care Terms. *ArcUser Magazine* April-June. Available online at: http://www.esri.com/news/arcuser/0499/terms.html. Last accessed April 23rd, 2008.

Feudtner C, Silveira MJ, Shabbout M, and Hoskins RE 2006 Distance From Home When Death Occurs: A Population-Based Study of Washington State, 1989–2002. *Pediatrics* 117(5): 932-939.

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

Foody GM 2003 Uncertainty, Knowledge Discovery and Data Mining in GIS. *Progress in Physical Geography* 27(1): 113-121.

Fortney J, Rost K, and Warren J 2000 Comparing Alternative Methods of Measuring Geographic Access to Health Services. *Health Services and Outcomes Research Methodology* 1(2): 173-184.

Frank AU, Grum E, and Vasseur B 2004 Procedure to Select the Best Dataset for a Task. In Egenhofer et al. (eds) *Proceedings of 3rd International Conference on Geographic Information Science (GIScience)*. Berlin, Springer Lecture Notes in Computer Science No 3234: 81-93.

Fremont AM, Bierman A, Wickstrom SL, Bird CE, Shah M, Escarce JJ, Horstman T, and Rector T 2005 Use Of Geocoding In Managed Care Settings To Identify Quality Disparities. *Health Affairs* 24(2): 516-526.

Frew J, Freeston M, Freitas N, Hill LL, Janee G, Lovette K, Nideffer R, Smith TR, and Zheng Q 1998 The Alexandria Digital Library Architecture. In Nikalaou and Stephanidis (eds) *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98)*. Berlin, Springer Lecture Notes in Computer Science No 1513: 61-73.

Fu G, Jones CB, and Abdelmoty AI 2005a Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proceedings of the IASTED International Conference on Databases and Applications (DBA-2005)*.

Fu G, Jones CB, and Abdelmoty AI 2005b Ontology-based Spatial Query Expansion in Information Retrieval. Berlin, Springer Lecture Notes in Computer Science No 3761: 1466-1482.

Fulcomer MC, Bastardi MM, Raza H, Duffy M, Dufficy E, and Sass MM 1998 Assessing the Accuracy of Geocoding Using Address Data from Birth Certificates: New Jersey, 1989 to 1996. In Williams et al. (eds) *Proceedings of the 1998 Geographic Information Systems in Public Health Conference, San Diego,* CA: 547-560. Available online at: http://www.atsdr.cdc.gov/GIS/conference98/proceedings/pdf/gisbook.pdf. Last accessed April 23rd, 2008.

Gabrosek J and Cressie N 2002 The Effect on Attribute Prediction on Location Uncertainty in Spatial Data. *Geographical Analysis* 34: 262-285.

Gaffney SH, Curriero FC, Strickland PT, Glass GE, Helzlsouer KJ, and Breysse PN 2005 Influence of Geographic Location in Modeling Blood Pesticide Levels in a Community Surrounding a US Environmental Protection Agency Superfund Site. *Environmental Health Perspectives* 113(12): 1712-1716.

Gatrell AC 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.

Geronimus AT and Bound J 1998 Use of Census-based Aggregate Variables to Proxy for Socioeconomic Group: Evidence from National Samples. *American Journal of Epidemiology* 148(5): 475-486.

Geronimus AT and Bound J 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.

Geronimus AT and Bound J 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 AT, Bound J, and Neidert LJ 1995 On the Validity of Using Census Geocode Characteristics to Proxy Individual Socioeconomic Characteristics. Technical Working Paper 189. Cambridge, MA, National Bureau of Economic Research.

Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, Savitz DA, and Herring AH 2006 Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research* 101(2): 256-262.

Gittler J 2008a Cancer Registry Data and Geocoding: Privacy, Confidentiality, and Security Issues. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 195-224.

Gittler J 2008b Cancer Reporting and Registry Statutes and Regulations. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 227-234.

Goldberg DW 2008a Free Online Geocoding. Available online at: https://webgis.usc.edu/Services/Geocode/Default.aspx. Last accessed October 20th, 2008.

Goldberg DW 2008b Free Online Static Address Validation. Available online at: https://webgis.usc.edu/Services/AddressValidation/StaticValidator.aspx. Last accessed October 20th, 2008.

Goldberg DW, Wilson JP, and Knoblock CA 2007a From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal* 19(1): 33-46.

Goldberg DW, Zhang X, Marusek JC, Wilson JP, Ritz B, and Cockburn MG 2007b Development of an Automated Pesticide Exposure Analyst for the California's Central Valley. In the *Proceedings of the URISA GIS in Public Health Conference*, New Orleans, LA: 136-156.

Goldberg DW, Wilson JP, Knoblock CA, and Cockburn MG 2008a The Development of an Open-Source, Scalable and Flexible Geocoding Platform. In preparation.

Goldberg DW, Shahabi K, Wilson JP, Knoblock CA, and Cockburn MG 2008b The Geographic Characteristics of Geocoding Error. In preparation.

Goldberg DW, Wilson JP, Knoblock CA, and Cockburn MG 2008c Geocoding Quality Metrics: One Size Does Not Fit All. In preparation.

Goldberg DW, Wilson JP, Knoblock CA, and Cockburn MG 2008d An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*, Submitted September 23rd, 2008.

Goodchild MF and Hunter GJ 1997 A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science* 11(3): 299-306.

Google, Inc. 2008a Google Earth. Available online at: http://earth.google.com. Last accessed April 23rd, 2008.

Google, Inc. 2008b Google Maps. Available online at: http://www.maps.google.com. Last accessed April 23rd, 2008.

Google, Inc. 2008c Google Maps API Documentation. Available online at: http://www.google.com/apis/maps/documentation. Last accessed April 23rd, 2008.

Grand Valley Metropolitan Council 2008 REGIS: Purchase Digital Data. Available online at: http://www.gvmc-regis.org/data/ordering.html. Last accessed April 23rd, 2008.

Gregorio DI, Cromley E, Mrozinski R, and Walsh SJ 1999 Subject Loss in Spatial Analysis of Breast Cancer. *Health & Place* 5(2): 173-177.

Gregorio DI, DeChello LM, Samociuk H, and Kulldorff M 2005 Lumping or Splitting: Seeking the Preferred Areal Unit for Health Geography Studies. *International Journal of Health Geographics* 4(6).

Griffin DH, Pausche JM, Rivers EB, Tillman AL, and Treat JB 1990 Improving the Coverage of Addresses in the 1990 Census: Preliminary Results. In *Proceedings of the American Statistical Association Survey Research Methods Section*, Anaheim, CA: 541–546. Available online at: http://www.amstat.org/sections/srms/Proceedings/papers/1990_091.pdf. Last accessed April 23rd, 2008.

Grubesic TH and Matisziw TC 2006 On the use of ZIP Codes and ZIP Code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics* 5(58).

Grubesic TH and Murray AT 2004 Assessing the Locational Uncertainties of Geocoded Data. In *Proceedings from the 24th Urban Data Management Symposium*. Available online at: http://www.tonygrubesic.net/geocode.pdf. Last accessed April 23rd, 2008.

Han D, Rogerson PA, Nie J, Bonner MR, Vena JE, Vito D, Muti P, Trevisan M, Edge SB, and Freudenheim JL 2004 Geographic Clustering of Residence in Early Life and Subsequent Risk of Breast Cancer (United States). *Cancer Causes and Controls* 15(9):921-929.

Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, and Freudenheim JL 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).

Hariharan R and Toyama K 2004 Project Lachesis: Parsing and Modeling Location Histories. In Egenhofer et al. (eds) *Proceedings of 3rd International Conference on Geographic Information Science (GIScience)*. Berlin, Springer Lecture Notes in Computer Science No 3234: 106-124.

Harvard University 2008 The Public Health Disparities Geocoding Project Monograph Glossary. Available online at: http://www.hsph.harvard.edu/thegeocodingproject/webpage/monograph/glossary.htm. Last accessed April 23rd, 2008.

Haspel M and Knotts HG 2005 Location, Location, Location: Precinct Placement and the Costs of Voting. *The Journal of Politics* 67(2): 560-573.

Health Level Seven, Inc. 2007 Application Protocol for Electronic Data Exchange in Healthcare Environments, Version 2.6. Available online at: http://www.hl7.org/Library/standards.cfm. Last accessed April 23rd, 2008.

Henry KA and Boscoe FP 2008 Estimating the accuracy of geographical imputation. *International Journal of Health Geographics* 7(3).

Henshaw SL, Curriero FC, Shields TM, Glass GE, Strickland PT, and Breysse PN 2004 Geostatistics and GIS: Tools for Characterizing Environmental Contamination. *Journal of Medical Systems* 28(4): 335-348.

Higgs G and Martin DJ 1995a The Address Data Dilemma Part 1: Is the Introduction of Address-Point the Key to Every Door in Britain? *Mapping Awareness* 8: 26-28.

Higgs G and Martin DJ 1995b The Address Data Dilemma Part 2: The Local Authority Experience, and Implications for National Standards. *Mapping Awareness* 9: 26-39.

Higgs G and Richards W 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.

Hild H and Fritsch D 1998 Integration of vector data and satellite imagery for geocoding. *International Archives of Photogrammetry and Remote Sensing* 32(4): 246-251.

Hill LL 2000 Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In Borbinha and Baker (eds) *Research and Advanced Technology for Digital Libraries, 4th European Conference (ECDL '00)*. Berlin, Springer Lecture Notes in Computer Science No 1923: 280-290.

Hill LL 2006 *Georeferencing: The Geographic Associations of Information*. Cambridge, Mass MIT Press.

Hill LL and Zheng Q 1999 Indirect Geospatial Referencing Through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers. In *Proceedings if the 62nd Annual Meeting of the American Society for Information Science*, Washington, DC: 57-69.

Hill LL, Frew J, and Zheng Q 1999 Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib Magazine* 5(1).

Himmelstein M 2005 Local Search: The Internet Is the Yellow Pages. *Computer* 38(2): 26-34.

Hofferkamp J and Havener L (eds) 2008 *Standards for Cancer Registries: Data Standards and Data Dictionary, Volume II* (12th Edition). Springfield, IL North American Association of Central Cancer Registries.

Howe HL 1986 Geocoding NY State Cancer Registry. *American Journal of Public Health* 76(12): 1459-1460.

Hurley SE, Saunders TM, Nivas R, Hertz A, and Reynolds P 2003 Post Office Box Addresses: A Challenge for Geographic Information System-Based Studies. *Epidemiology* 14(4): 386-391.

Hutchinson M and Veenendall B 2005a Towards a Framework for Intelligent Geocoding. In *Spatial Intelligence, Innovation and Praxis: The National Biennial Conference of the Spatial Sciences Institute (SSC 2005)*, Melbourne, AU.

Hutchinson M and Veenendall B 2005b Towards Using Intelligence To Move From Geocoding To Geolocating. In *Proceedings of the 7th Annual URISA GIS in Addressing Conference*, Austin, TX.

Jaro M 1984 Record Linkage Research and the Calibration of Record Linkage Algorithms. Statistical Research Division Report Series SRD Report No. Census/ SRD/RR-84/27. Washington, DC United States Census Bureau. Available online at: http://www.census.gov/srd/papers/pdf/rr84-27.pdf. Last accessed April 23rd, 2008.

Jaro M 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 SD 1998a Address Matching with Stand-Alone Geocoding Engines: Part 1. *Business Geographics*: 24-32.

Johnson SD 1998b Address Matching with Stand-Alone Geocoding Engines: Part 2. *Business Geographics*: 30-36.

Jones CB, Alani H, and Tudhope D 2001 Geographical Information Retrieval with Ontologies of Place. In Montello (ed) *Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science*. Berlin, Springer Lecture Notes In Computer Science No 2205: 322-335.

Karimi HA, Durcik M, and Rasdorf W 2004 Evaluation of Uncertainties Associated with Geocoding Techniques. *Journal of Computer-Aided Civil and Infrastructure Engineering* 19(3): 170-185.

Kennedy TC, Brody JG, and Gardner JN 2003 Modeling Historical Environmental Exposures Using GIS: Implications for Disease Surveillance. In *Proceedings of the 2003 ESRI Health GIS Conference*, Arlington, VA. Available online at: http://gis.esri.com/library/userconf/health03/papers/pap3020/p3020.htm. Last accessed April 23rd, 2008.

Kim U 2001 Historical Study on the Parcel Number and Numbering System in Korea. In *Proceedings of the International Federation of Surveyors Working Week 2004*.

Kimler M 2004 Geo-Coding: Recognition of geographical references in unstructured text, and their visualization. PhD Thesis. University of Applied Sciences in Hof, Germany. Available online at: http://langtech.jrc.it/Documents/0408_Kimler_Thesis-GeoCoding.pdf. Last accessed April 23rd, 2008.

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.

Krieger N 2003 Place, Space, and Health: GIS and Epidemiology. *Epidemiology* 14(4): 384-385.

Krieger N and Gordon D 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, Williams DR, and Moss NE 1997 Measuring Social Class in US Public Health Research: Concepts, Methodologies, and Guidelines. *Annual Review of Public Health* 18(1): 341-378.

Krieger N, Waterman P, Lemieux K, Zierler S, and Hogan JW 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, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, and Carson R 2002a 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, Waterman PD, Chen JT, Soobader MJ, Subramanian SV, and Carson R 2002b 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.

Krieger N, Waterman PD, Chen JT, Soobader M, and Subramanian SV 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, Chen JT, Waterman PD, Rehkopf DH, and Subramanian SV 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, Chen JT, Waterman PD, Rehkopf DH, Yin R, and Coull BA 2006 Race/Ethnicity and Changing US Socioeconomic Gradients in Breast Cancer Incidence: California and Massachusetts, 1978–2002 (United States). *Cancer Causes and Control* 17(2): 217-226.

Krishnamurthy S, Sanders WH, and Cukier M 2002 An adaptive framework for tunable consistency and timeliness using replication. In *Proceedings of the International Conference on Dependable Systems and Networks*: 17-26.

Kwok RK and Yankaskas BC 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 AHF, Borges KAV, Carvalho JCP, Medeiros CB, da Silva AS, and Davis Jr. CA 2005 Integrating Web Data and Geographic Knowledge into Spatial Databases. In Manalopoulos and Papadapoulos (eds) *Spatial Databases: Technologies, Techniques and Trends*. Hershey, PA Idea Group Publishing: 23-47.

Lam CS, Wilson JP, and Holmes-Wong DA 2002 Building a Neighborhood-Specific Gazetteer for a Digital Archive. In *Proceedings of the Twenty-second International ESRI User Conference*: 7-11.

Lee J 2004 3D GIS for Geo-coding Human Activity in Micro-scale Urban Environments. In Egenhofer et al. (eds) *Geographic Information Science: Third International Conference (GIScience 2004)*, College Park, MD: 162-178.

Lee MS and McNally MG 1998 Incorporating Yellow-Page Databases in GIS-Based Transportation Models. In Easa (ed) *Proceedings the American Society of Civil Engineers Conference on Transportation Land Use, and Air Quality*: 652-661. Available online at: repositories.cdlib.org/itsirvine/casa/UCI-ITS-AS-WP-98-3.

Leidner JL 2004 Toponym Resolution in Text: "Which Sheffield is it"? In Sanderson et al. (eds) *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR '04)*: 602.

Levesque M 2003 West Virginia Statewide Addressing and Mapping Project. In *Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference*, Providence, RI.

Levine N and Kim KE 1998 The Spatial Location of Motor Vehicle Accidents: A Methodology for Geocoding Intersections. *Computers, Environment, and Urban Systems* 22(6): 557-576.

Li H, Srihari RK, Niu C, and Li W 2002 Location Normalization for Information Extraction. In *Proceedings of the 19th international conference on Computational linguistics*: 1-7.

Lianga S, Banerjeea S, Bushhouseb S, Finleyc AO, and Carlin BP 2007 Hierarchical multiresolution approaches for dense point-level breast cancer treatment data. *Computational Statistics & Data Analysis* 52(5): 2650-2668.

Lind M 2001 Developing a System of Public Addresses as a Language for Location Dependent Information. In *Proceedings of the 2001 URISA Annual Conference*. Available online at: http://www.adresseprojekt.dk/files/Develop_PublicAddress_urisa2001e.pdf. Last accessed April 23rd, 2008.

Lind M 2005 Addresses and Address Data Play a Key Role in Spatial Infrastructure. In *Proceedings of GIS Planet 2005 International Conference and Exhibition on Geographic Information, Workshop on Address Referencing*. Available online at: http://www.adresseprojekt.dk/files/ECGI_Addr.pdf. Last accessed April 23rd, 2008.

Locative Technologies 2006 Geocoder.us: A Free US Geocoder. Available online at: http://geocoder.us. Last accessed April 23rd, 2008.

Lockyer B 2005 Office of the Attorney General of the State of California Legal Opinion 04-1105. Available online at: http://ag.ca.gov/opinions/pdfs/04-1105.pdf. Last accessed April 23rd, 2008.

Los Angeles County Assessor 2008 LA Assessor - Parcel Viewer. Available online at: http://assessormap.co.la.ca.us/mapping/viewer.asp. Last accessed April 23rd, 2008.

Lovasi GS, Weiss JC, Hoskins R, Whitsel EA, Rice K, Erickson CF, and Psaty BM 2007 Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree? *International Journal of Health Geographics* 6(12).

MacDorman MF and Gay GA 1999 State Initiatives in Geocoding Vital Data. *Journal of Public Health Management and Practice* 5(2): 91-93.

Maizlish NA and Herrera L 2005 A Record Linkage Protocol for a Diabetes Registry at Ethnically Diverse Community Health Centers. *Journal of the American Medical Informatics Association* 12(3): 331-337.

Markowetz A 2004 Geographic Information Retrieval. Diploma Thesis, Philipps University. Available online at: http://www.cs.ust.hk/~alexmar/papers/DA.pdf. Last accessed April 23rd, 2008.

Markowetz A, Chen YY, Suel T, Long X, and Seeger B 2005 Design and implementation of a geographic search engine. In *Proceedings of the 8th International Workshop on the Web and Databases (WebDB)*. Available online at: http://cis.poly.edu/suel/papers/geo.pdf. Last accessed April 23rd, 2008.

Martin DJ 1998 Optimizing Census Geography: The Separation of Collection and Output Geographies. *International Journal of Geographical Information Science* 12(7): 673-685.

Martin DJ and Higgs G 1996 Georeferencing People and Places: A Comparison of Detailed Datasets. In Parker (ed) *Innovations in GIS 3: Selected Papers from the Third National Conference on GIS Research UK*. London, Taylor & Francis: 37-47.

Martins B and Silva MJ 2005 A Graph-Ranking Algorithm for Geo-Referencing Documents. In *Proceedings of the Fifth IEEE International Conference on Data Mining*: 741-744.

Martins B, Chaves M, and Silva MJ 2005a Assigning Geographical Scopes To Web Pages. In Losada and Fernández-Luna (eds) *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*. Berlin, Springer Lecture Notes in Computer Science No 3408: 564-567.

Martins B, Silva MJ, and Chaves MS 2005b Challenges and Resources for Evaluating Geographical IR. In Jones and Purves (eds) *Proceedings of the 2005 ACM Workshop of Geographic Information Retrieval (GIR'05)*: 17-24.

Marusek JC, Cockburn MG, Mills PK, and Ritz BR 2006 Control Selection and Pesticide Exposure Assessment via GIS in Prostate Cancer Studies. *American Journal of Preventive Medicine* 30(2S): 109-116.

Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, and Donham KJ 2008 Geocoding Accuracy and the Recovery of Relationships Between Environmental Exposures and Health. *International Journal of Health Geographics* 7(13).

McCurley KS 2001 Geospatial Mapping and Navigation of the Web. In Shen and Saito (eds) *Proceedings of the 10th International World Wide Web Conference*: 221-229.

McEathron SR, McGlamery P, and Shin DG 2002 Naming the Landscape: Building the Connecticut Digital Gazetteer. *International Journal of Special Libraries* 36(1): 83-93.

McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, and Newcomb PA 2003 Geocoding Addresses from a Large Population-Based Study: Lessons Learned. *Epidemiology* 14(4): 399-407.

Mechanda M and Puderer H., 2007 How Postal Codes Map to Geographic Areas. Geography Working Paper Series #92F0138MWE. Ottawa, Statistics Canada.

Meyer M, Radespiel-Tröger M, and Vogel C 2005 Probabilistic Record Linkage of Anonymous Cancer Registry Records. In *Studies in Classification, Data Analysis, and Knowledge Organization: Innovations in Classification, Data Science, and Information Systems*. Berlin, Springer: 599-604.

Michelson M and Knoblock CA 2005 Semantic Annotation of Unstructured and Ungrammatical Text. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland.

Miner JW, White A, Palmer S, and Lubenow AE 2005 Geocoding and Social Marketing in Alabama's Cancer Prevention Programs. *Preventing Chronic Disease* 2(A17).

Ming D, Luo J, Li J, and Shen Z 2005 Features Based Parcel Unit Extraction From High Resolution Image. In Moon (ed) *Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'05)*: 1875-1878.

Murphy J and Armitage R 2005 Merging the Modeled and Working Address Database: A Question of Dynamics and Data Quality. In *Proceedings of GIS Ireland 2005*, Dublin.

National Institute of Standards and Technology 2008 *Federal Information Processing Standards Publications*. Available online at: http://www.itl.nist.gov/fipspubs. Last accessed April 23rd, 2008.

Nattinger AB, Kneusel RT, Hoffmann RG, and Gilligan MA 2001 Relationship of Distance From a Radiotherapy Facility and Initial Breast Cancer Treatment. *Journal of the National Cancer Institute* 93(17): 1344-1346.

NAVTEQ 2008 NAVSTREETS. Available online at: http://developer.navteq.com/site/global/dev_resources/170_navteqproducts/navdataformats/navstreets/p_navstreets.jsp. Last accessed April 23rd, 2008.

Nicoara G 2005 Exploring the Geocoding Process: A Municipal Case Study using Crime Data. Masters Thesis, The University of Texas at Dallas, Dallas, TX.

Noaki K and Arikawa M 2005a A Method for Parsing Route Descriptions using Sidewalk Network Databases. In *Proceedings of the 2005 International Cartographic Conference (ICC 2005)*.

Noaki K and Arikawa M 2005b A geocoding method for natural route descriptions using sidewalk network databases. In Kwon et al. (eds) *Proceedings of the 4th International Workshop on Web and Wireless Geographical Information Systems (W2GIS 2004) Revised Selected Papers*. Berlin, Springer Lecture Notes in Computer Science No 3428: 38-50.

Nuckols JR, Ward MH, and Jarup L 2004 Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. *Environmental Health Perspectives* 112(9): 1007-1115.

Nuckols JR, Gunier RB, Riggs P, Miller R, Reynolds P, and Ward MH 2007 Linkage of the California Pesticide Use Reporting Database with Spatial Land Use Data for Exposure Assessment. *Environmental Health Perspectives* 115(1): 684-689.

NAACCR 2008a NAACCR GIS Committee Geographic Information Systems Survey. Available online at: http://www.naaccr.org/filesystem/word/GIS%20survey_Final.doc. Last accessed April 23rd, 2008.

NAACCR 2008b NAACCR Results of GIS Survey. Spring 2006 NAACCR Newsletter. Available online at: http://www.naaccr.org/index.asp?Col_SectionKey=6&Col_ContentID=497. Last accessed April 23rd, 2008.

O'Grady KM 1999 A DOQ Test Project: Collecting Data to Improve TIGER. In *Proceedings of the 1999 ESRI User's Conference*, San Diego, CA. Available online at: http://gis.esri.com/library/userconf/proc99/proceed/papers/pap635/p635.htm. Last accessed April 23rd, 2008.

O'Reagan RT and Saalfeld A 1987 Geocoding Theory and Practice at the Bureau of the Census. Statistical Research Report Census/SRD/RR-87/29. Washington, DC United States Bureau of Census.

Oliver MN, Matthews KA, Siadaty M, Hauck FR, and Pickle LW 2005 Geographic Bias Related to Geocoding in Epidemiologic Studies. *International Journal of Health Geographics* 4(29).

Olligschlaeger AM 1998 Artificial Neural Networks and Crime Mapping. In Weisburd and McEwen (eds) *Crime Mapping and Crime Prevention*. Monsey, NY Criminal Justice Press: 313–347.

Olston C and Widom J 2005 Efficient Monitoring and Querying of Distributed, Dynamic Data via Approximate Replication. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*: 1-8.

Openshaw S 1984 The modifiable areal unit problem. *Concepts and Techniques in Modern Geography* 38. Norwich, GeoBooks.

Openshaw S 1989 Learning to Live with Errors in Spatial Databases. In Goodchild and Gopal (eds) *Accuracy of Spatial Databases*. Bristol, PA Taylor & Francis: 263-276.

Oppong JR 1999 Data Problems in GIS and Health. In *Proceedings of Health and Environment Workshop 4: Health Research Methods and Data*, Turku, Finland. Available online at: geog.queensu.ca/h_and_e/healthandenvir/Finland%20Workshop%20Papers/OPPONG.DOC. Last accessed April 23rd, 2008.

Ordnance Survey 2008 ADDRESS-POINT: Ordnance Survey's Map Dataset of All Postal Addresses in Great Britain. Available online at: http://www.ordnancesurvey.co.uk/oswebsite/products/addresspoint. Last accessed April 23rd, 2008.

Organization for the Advancement of Structured Information Standards 2008 OASIS xAL Standard v2.0. Available online at: http://www.oasis-open.org/committees/ciq/download.html. Last accessed April 23rd, 2008.

Paull D 2003 A Geocoded National Address File for Australia: The G-NAF What, Why, Who and When? Available online at: http://www.psma.com.au/file_download/24. Last accessed April 23rd, 2008.

Porter MF 1980 An algorithm for suffix stripping, *Program* 14(3): 130-137.

Purves R, Clough P, and Joho H 2005 Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of GISRUK*: 313-318.

Raghavan VV, Bollmann P, and Jung GS 1989 Retrieval system evaluation using recall and precision: problems and answers. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '89)*: 59-68.

Ratcliffe JH 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.

Ratcliffe JH 2004 Geocoding Crime and a First Estimate of a Minimum Acceptable Hit Rate. *International Journal of Geographical Information Science* 18(1): 61-72.

Rauch E, Bukatin M, and Baker K 2003 A confidence-based framework for disambiguating geographic terms. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 50-54.

Reid J 2003 GeoXwalk: A Gazetteer Server and Service for UK Academia. In Koch and Sølvberg (eds) *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '03)*. Berlin, Springer Lecture Notes in Computer Science No 2769: 387-392.

Reinbacher I 2006 Geometric Algorithms for Delineating Geographic Regions. Ph.D. Thesis, Utrecht University, NL. Available online at: http://igitur-archive.library.uu.nl/dissertations/2006-0620-200747/full.pdf. Last accessed April 23rd, 2008.

Reinbacher I, Benkert M, van Kreveld M, Mitchell JSB, and Wolff A 2008 Delineating Boundaries for Imprecise Regions. *Algorithmica* 50(3): 386-414.

Revie P and Kerfoot H 1997 The Canadian Geographical Names Data Base. Available online at: http://geonames.nrcan.gc.ca/info/cgndb_e.php. Last accessed April 23rd, 2008.

Reynolds P, Hurley SE, Gunier RB, Yerabati S, Quach T, and Hertz A 2005 Residential Proximity to Agricultural Pesticide Use and Incidence of Breast Cancer in California, 1988-1997. *Environmental Health Perspectives* 113(8): 993-1000.

Riekert WF 2002 Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources. *Journal of Universal Computer Science* 8(6): 581-590.

Rose KM, Wood JL, Knowles S, Pollitt RA, Whitsel EA, Diez-Roux AV, Yoon D, and Heiss G 2004 Historical Measures of Social Context in Life Course Studies: Retrospective Linkage of Addresses to Decennial Censuses. *International Journal of Health Geographics* 3(27).

Rull RP and Ritz B 2003 Historical pesticide exposure in California using pesticide use reports and land-use surveys: an assessment of misclassification error and bias. *Environmental Health Perspectives* 111(13): 1582-1589.

Rull RP, Ritz B, Krishnadasan A, and Maglinte G 2001 Modeling Historical Exposures from Residential Proximity to Pesticide Applications. In *Proceedings of the Twenty-First Annual ESRI User Conference*, San Diego, CA.

Rull RP, Ritz B, and Shaw GM 2006 Neural Tube Defects and Maternal Residential Proximity to Agricultural Pesticide Applications. *American Journal of Epidemiology* 163(8): 743-753.

Rull RP, Ritz B, and Shaw GM 2006 Validation of Self-Reported Proximity to Agricultural Crops in a Case-Control Study of Neural Tube Defects. *Journal of Exposure Science Environmental Epidemiology* 16(2): 147-155.

Rushton G, Peleg I, Banerjee A, Smith G, and West M 2004 Analyzing Geographic Patterns of Disease Incidence: Rates of Late-Stage Colorectal Cancer in Iowa. *Journal of Medical Systems* 28(3): 223-236.

Rushton G, Armstrong M, Gittler J, Greene B, Pavlik C, West M, and Zimmerman D 2006 Geocoding in Cancer Research - A Review. *American Journal of Preventive Medicine* 30(2): S16–S24.

Rushton G, Armstrong, MP, Gittler J, Greene BR, Pavlik CE, West MW, and Zimmerman DL (eds) 2008a *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, Boca Raton, Fl CRC Press.

Rushton G, Cai Q, and Chen Z 2008b Producing Spatially Continuous Prostate Cancer Maps with Different Geocodes and Spatial Filter Methods. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 69-94.

Sadahiro Y 2000 Accuracy of count data estimated by the point-in-polygon method. *Geographical Analysis* 32(1): 64-89.

Schlieder C, Vögele TJ, and Visser U 2001 Qualitative Spatial Representation for Information Retrieval by Gazetteers. In Montello (ed) *Proceedings of the 5th International Conference on Spatial Information Theory*. Berlin, Springer Lecture Notes in Computer Science No 2205: 336-351.

Schockaert S, De Cock M, and Kerre EE 2005 Automatic Acquisition of Fuzzy Footprints. In Meersman et al. (eds) *On the Move to Meaningful Internet Systems 2005 (OTM 2005)*. Berlin, Springer Lecture Notes in Computer Science No 3762: 1077-1086.

Schootman M, Jeffe D, Kinman E, Higgs G, and Jackson-Thompson J 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.

Schumacher S 2007 Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision. *DM Direct* Special Report (January 18, 2007 Issue). Available online at: http://www.dmreview.com/article_sub.cfm?articleId=1071712. Last accessed April 23rd, 2008.

Sheehan TJ, Gershman ST, MacDougal L, Danley RA, Mroszczyk M, Sorensen AM, and Kulldorff M 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.

Skelly C, Black W, Hearnden M, Eyles R, and Weinstein P 2002 Disease surveillance in rural communities is compromised by address geocoding uncertainty: A case study of campylobacteriosis. *Australian Journal of Rural Health* 10(2): 87-93.

Smith DA and Crane G 2001 Disambiguating Geographic Names in a Historical Digital Library. In Constantopoulos and Sølvberg (eds) *Research and Advanced Technology for Digital Libraries, 5th European Conference*. Berlin, Springer Lecture Notes in Computer Science No 2163: 127-136.

Smith DA and Mann GS 2003 Bootstrapping toponym classifiers. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 45-49.

Smith GD, Ben-Shlomo Y, and Hart C 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, LeClere FB, Hadden W, and Maury B 2001 Using Aggregate Geographic Data to Proxy Individual Socioeconomic Status: Does Size Matter? *American Journal of Public Health* 91(4): 632-636.

Southall H 2003 Defining and identifying the roles of geographic references within text: Examples from the Great Britain Historical GIS project. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 69-78.

Stage D and von Meyer N 2005 An Assessment of Parcel Data in the United States 2005 Survey Results. Federal Geographic Data Committee Subcommittee on Cadastral Data. Available online at: http://www.nationalcad.org/showdocs.asp?docid=170. Last accessed April 23rd, 2008.

Statistics Canada 2008 Census geography – Illustrated Glossary: Geocoding – plain definition. Available online at: http://geodepot.statcan.ca/Diss2006/Reference/COGG/Short_RSE_e.jsp?REFCODE=10&FILENAME=Geocoding&TYPE=L. Last accessed April 23rd, 2008.

Stefoski Mikeljevic J, Haward R, Johnston C, Crellin A, Dodwell D, Jones A, Pisani P, and Forman D 2004 Trends in postoperative radiotherapy delay and the effect on survival in breast cancer patients treated with conservation surgery. *British Journal of Cancer* 90: 1343-1348.

Stevenson MA, Wilesmith J, Ryan J, Morris R, Lawson A, Pfeiffer D, and Lin D 2000 Descriptive Spatial Analysis of the Epidemic of Bovine Spongiform Encephalopathy in Great Britain to June 1997. *The Veterinary Record* 147(14): 379-384.

Strickland MJ, Siffel C, Gardner BR, Berzen AK, and Correa A 2007 Quantifying geocode location error using GIS methods. *Environmental Health* 6(10).

Stitzenberg KB, Thomas NE, Dalton K, Brier SE, Ollila DW, Berwick M, Mattingly D, and Millikan RC 2007 Distance to Diagnosing Provider as a Measure of Access for Patients With Melanoma. *Archives of Dermatology* 143(8): 991-998.

Sweeney L 2002 *k*-Anonymity: A Model For Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 557-570.

Tele Atlas Inc. 2008a Dynamap Map Database. Available online at: http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm. Last accessed April 23rd, 2008.

Tele Atlas Inc. 2008b Geocode.com // Tele Atlas Geocoding Services. Available online at: http://www.geocode.com. Last accessed April 23rd, 2008

Tele Atlas Inc. 2008c MultiNet Map Database. Available online at: http://www.teleatlas.com/OurProducts/MapData/Multinet/index.htm. Last accessed April 23rd, 2008.

Temple C, Ponas G, Regan R, and Sochats K 2005 The Pittsburgh Street Addressing Project. In *Proceedings of the 25th Annual ESRI International User Conference*. Available online at: http://gis2.esri.com/library/userconf/proc05/papers/pap1525.pdf. Last accessed April 23rd, 2008.

Tezuka T and Tanaka K 2005 Landmark Extraction: A Web Mining Approach. In Cohn and Mark (eds) *Proceedings of the 7th International Conference on Spatial Information Theory*. Berlin, Springer Lecture Notes in Computer Science No 3693: 379-396.

Thrall GI 2006 Geocoding Made Easy. *Geospatial Solutions*. Available online at: http://ba.geospatial-online.com/gssba/article/articleDetail.jsp?id=339221. Last accessed April 23rd, 2008.

Tobler W 1972 Geocoding Theory. In *Proceedings of the National Geocoding Conference*, Washington, DC Department of Transportation: IV.1.

Toral A and Muñoz R 2006 A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the EACL-2006 Workshop on NEW TEXT: Wikis and blogs and other dynamic text sources*: 56-61. Available online at: acl.ldc.upenn.edu/W/W06/W06-2800.pdf. Last accessed April 23rd, 2008.

United States Board on Geographic Names 2008 *Geographic Names Information System*. Reston, VA United States Board on Geographic Names. Available online at: http://geonames.usgs.gov/pls/gnispublic. Last accessed April 23rd, 2008.

United States Census Bureau 2008a *American Community Survey*, Washington, DC United States Census Bureau. Available online at: http://www.census.gov/acs. Last accessed April 23rd, 2008.

United States Census Bureau 2008b *MAF/TIGER Accuracy Improvement Project*, Washington, DC United States Census Bureau. Available online at: http://www.census.gov/geo/mod/maftiger.html. Last accessed April 23rd, 2008.

United States Census Bureau 2008c *Topologically Integrated Geographic Encoding and Referencing System*, Washington, DC United States Census Bureau. Available online at: http://www.census.gov/geo/www/tiger. Last accessed April 23rd, 2008.

United States Department of Health and Human Services 2000 *Healthy People 2010: Understanding and Improving Health* (Second Edition), Washington, DC United States Government Printing Office. Available online at: http:/www.healthypeople.gov/Document/pdf/uih/2010uih.pdf. Last accessed April 23rd, 2008.

United States Federal Geographic Data Committee 2008a *Content Standard for Digital Geospatial Metadata*. Available online at: http://www.fgdc.gov/metadata/csdgm. Last accessed April 23rd, 2008.

United States Federal Geographic Data Committee 2008b *Street Address Data Standard*. Reston, VA United States Federal Geographic Data Committee. Available online at: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/ streetaddress/index_html. Last accessed April 23rd, 2008.

United States National Geospatial-Intelligence Agency 2008 *NGA GNS Search*. Bethesda, MD United States National Geospatial-Intelligence Agency. Available online at: http://geonames.nga.mil/ggmagaz/geonames4.asp. Last accessed April 23rd, 2008

United States Postal Service 2008a *Address Information System Products Technical Guide*. Washington, DC United States Postal Service. Available online at: http://ribbs.usps.gov/files/Addressing/PUBS/AIS.pdf. Last accessed April 23rd, 2008.

United States Postal Service 2008b *CASS Mailer's Guide*. Washington, DC United States Postal Service. Available online at: http://ribbs.usps.gov/doc/cmg.html. Last accessed April 23rd, 2008.

United States Postal Service 2008c *Locatable Address Conversion System*. Washington, DC United States Postal Service. Available online at: http://www.usps.com/ncsc/addressservices/addressqualityservices/lacsystem.htm. Last accessed April 23rd, 2008.

United States Postal Service 2008d *Publication 28 – Postal Addressing Standards*. Washington, DC United States Postal Service. Available online at: http://pe.usps.com/text/pub28/welcome.htm. Last accessed April 23rd, 2008.

University of California, Los Angeles 2008 *Interactive UCLA Campus Map*. Los Angeles, CA, University of California, Los Angeles. Available online at: http://www.fm.ucla.edu/CampusMap/Campus.htm. Last accessed April 23rd, 2008.

University of Southern California 2008 *UPC Color Map*. Los Angeles, CA University of Southern California. Available online at: http://www.usc.edu/private/about/visit_usc/USC_UPC_map_color.pdf. Last accessed April 23rd, 2008.

Vaid S, Jones CB, Joho H, and Sanderson M 2005 Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th Symposium on Spatial and Temporal Databases (SSTD-05)*.

Van Kreveld M and Reinbacher I 2004 Good NEWS: Partitioning a Simple Polygon by Compass Directions. *International Journal of Computational Geometry and Applications* 14(4): 233-259.

Veregin H 1999 Data Quality Parameters. In Longley et al. (eds) *Geographical Information Systems, Volume 1* (Second Edition). New York, Wiley: 177-189.

Vestavik Ø 2004 Geographic Information Retrieval: An Overview. Unpublished, presented at Internal Doctoral Conference, Department of Computer and Information Science, Norwegian University of Science and Technology. Available online at: http://www.idi.ntnu.no/~oyvindve/article.pdf. Last accessed April 23rd, 2008.

Vine MF, Degnan D, and Hanchette C 1998 Geographic Information Systems: Their Use in Environmental Epidemiologic Research. *Journal of Environmental Health* 61: 7-16.

Vögele TJ and Schlieder C 2003 Spatially-Aware Information Retrieval with Graph-Based Qualitative Reference Models. In Russell and Haller (eds) *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*: 470-474.

Vögele TJ and Stuckenschmidt H 2001 Enhancing Gazetteers with Qualitative Spatial Concepts. In Tochtermann and Arndt (eds) *Proceedings of the Workshop on Hypermedia in Environmental Protection.*

Vögele TJ, Schlieder C, and Visser U 2003 Intuitive modeling of place name regions for spatial information retrieval. In Kuhn et al. (eds) *Proceedings of the 6th International Conference on Foundations of Geographic Information Science (COSIT 2003)*. Berlin, Springer Lecture Notes in Computer Science No 2825: 239-252.

Voti L, Richardson LC, Reis IM, Fleming LE, MacKinnon J, Coebergh JWW 2005 Treatment of local breast carcinoma in Florida. *Cancer* 106(1): 201-207.

Waldinger R, Jarvis P, and Dungan J 2003 Pointing to places in a deductive geospatial theory. In Kornai and Sundheim (eds) *Proceedings of Workshop on the Analysis of Geographic References held at Joint Conference for Human Language Technology and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '03)*: 10-17.

Waller LA 2008 Spatial Statistical Analysis of Point- and Area-Referenced Public Health Data. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 147-164.

Walls MD 2003 Is Consistency in Address Assignment Still Needed?. In *Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference*, Providence, RI.

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

Werner PA 1974 National Geocoding. *Annals of the Association of American Geographers* 64(2): 310-317.

Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, and Heiss G 2004 Accuracy and Repeatability of Commercial Geocoding. *American Journal of Epidemiology* 160(10): 1023-1029.

Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, and Heiss G 2006 Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives & Innovations* 3(8).

Wieczorek J 2008 MaNIS/HerpNet/ORNIS Georeferencing Guidelines. Available online at: http://manisnet.org/GeorefGuide.html. Last accessed April 23rd, 2008.

Wieczorek J, Guo Q, and Hijmans RJ 2004 The Point-Radius Method for Georeferencing Locality Descriptions and Calculating Associated Uncertainty. *International Journal of Geographical Information Science* 18(8): 745-767.

Wilson JP, Lam CS, and Holmes-Wong DA 2004 A New Method for the Specification of Geographic Footprints in Digital Gazetteers. *Cartography and Geographic Information Science* 31(4): 195-203.

Winkler WE 1995 Matching and Record Linkage. In Cox et al. (eds) *Business Survey Methods*. New York, Wiley: 355-384.

Wong WS and Chuah MC 1994 A Hybrid Approach to Address Normalization. *IEEE Expert: Intelligent Systems and Their Applications* 9(6): 38-45.

Woodruff AG and Plaunt C 1994 GIPSY: Georeferenced Information Processing System. *Journal of the American Society for Information Science*: 645-655.

Wu J, Funk TH, Lurmann FW, and Winer AM 2005 Improving Spatial Accuracy of Roadway Networks and Geocoded Addresses. *Transactions in GIS* 9(4): 585-601

Yahoo!, Inc. 2008 Yahoo! Maps Web Services - Geocoding API. Available online at: http://developer.yahoo.com/maps/rest/V1/geocode.html. Last accessed April 23rd, 2008.

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

Yildirim V and Yomralioglu T 2004 An Address-based Geospatial Application. In *Proceedings of the International Federation of Surveyors Working Week 2004*.

Yu L 1996 Development and Evaluation of a Framework for Assessing the Efficiency and Accuracy of Street Address Geocoding Strategies. Ph.D. Thesis, University at Albany, State University of New York - Rockefeller College of Public Affairs and Policy, New York.

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

Zandbergen PA 2008 A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*.

Zandbergen PA and Chakraborty J 2006 Improving environmental exposure analysis using cumulative distribution functions and individual geocoding. *International Journal of Health Geographics* 5(23).

Zillow.com 2008 Zillow - Real Estate Search Results. Available online at: http://www.zillow.com/search/Search.htm?mode=browse. Last accessed April 23rd, 2008.

Zimmerman DL 2006 Estimating spatial intensity and variation in risk from locations coarsened by incomplete geocoding. Technical report #362, Department of Statistics and Actuarial Science, University of Iowa: 1-28. Available online at: http://www.stat.uiowa.edu/techrep/tr362.pdf. Last accessed April 23rd, 2008.

Zimmerman DL 2008 Statistical Methods for Incompletely and Incorrectly Geocoded Cancer Data. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 165-180.

Zimmerman DL, Armstrong MP, and Rushton G 2008 Alternative Techniques for Masking Geographic Detail to Protect Privacy. In Rushton et al. (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 127-138.

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

Zong W, Wu D, Sun A, Lim EP, and Goh DHL 2005On Assigning Place Names to Geography Related Web Pages. In Marlino et al. (eds) *Proceeding of the 2005 ACM/IEEE Joint Conference on Digital Libraries*: 354-362.

This page is left blank intentionally.

APPENDIX A: EXAMPLE RESEARCHER ASSURANCE DOCUMENTS

Cancer registries should already have formalized policies regarding the distribution of registry data in terms of who can access the data for what purposes. The following pages contain an example researcher data release request document that registries can use as a starting point to standardize these procedures, if this is needed. Also included is an example researcher assurances agreement which can be used to partially protect the registry by specifying the acceptable usage of registry data and outlining the responsibilities of the researcher.

This page is left blank intentionally.

Research Review Procedure

Before the release of any data, all research proposals requesting the use of confidential cancer registry data must be reviewed by the *Name_of_Registry* for compliance with the following criteria:

1. the proposed research will be used to determine the sources of cancer among the residents of *Name_of_Locality* or to reduce the burden of cancer in *Name_of_Locality*;

2. the data requested are necessary for the efficient conduct of the study;

3. adequate protections are in place to provide secure conditions to use and store the data;

4. assurances are given that the data will only be used for the purposes of the study, and assurances that confidential data will be destroyed at the conclusion of the study (see Assurances Form);

5. the researcher has adequate resources to carry out the proposed research;

6. the proposal has been reviewed and approved by the *Name_of_Committee_for_the Protection_of_Human_Subjects* or is exempt from such review;

7. any additional safeguards needed to protect the data from inadvertent disclosure due to unique or special characteristics of the proposed research have been required of the researcher; and

8. the research methodology has been reviewed for scientific excellence by a nationally recognized peer group, or if such a review has not taken place, that an ad hoc peer review subcommittee of the *Name_of_Advisory_Committee* containing appropriately qualified scientists has performed a peer review of the research.

Additionally, all relevant research fees have been paid prior to data release.

Name_of_Health_Division
Name_of_Registry

Research Proposal Review

Please complete each section of this form and return with all attachments to: *Address_of_Registry*.

Principal Investigator_____ Date_____

Organization_____

Address_____

City_____ State_____ ZIP_____

Tel _____ Fax _____ Email _____

Title of Research Project _____

List other institutions or agencies that will collaborate in conducting the project:

Note: please attach a copy of the proposed protocol or methods section of your project.

1. *Section* of *Legislative_Code* states *" the purpose of the registry shall be to purpose_of_registry."* In the section below, please describe how your proposed research will be used to determine the sources of cancer among the residents of *Name_of_Locality* or to reduce the burden of cancer in *Name_of_Locality*. If additional space is needed, please attach a separate sheet.

2. Details of data necessary for conduct of the study elements.

3. Describe procedures for identifying patients (patient population).

4. All protocols including a request for confidential data require peer review for scientific merit. *Name_of_Registry* accepts review by nationally recognized peer review groups. Please indicate below whether or not such a review has been performed.

☐ No

☐ Yes, if your proposal has been reviewed for scientific merit, please attach a copy of that review.

If your proposal has **not** been reviewed for scientific merit by a nationally recognized peer review group, the Division shall convene an ad hoc peer review subcommittee of the Cancer Registry Advisory Committee. The data shall not be released unless and until the proposed research is judged to be scientifically meritorious by the peer group. Review for scientific merit must be completed prior to Committee for Protection for Human Research Subjects Institutional Review Board (IRB) review if one has not already been performed.

5. All requests for confidential data must be approved by an IRB established in accordance with *Section* of *Legislative_Code*. Please indicate whether or not this proposal has already been approved by an IRB.

☐ No Please indicate the approximate review date: _____

☐ Yes Date: _____

If your proposal has been approved by an IRB, please attach a copy of the approval. *Name_of_Health_Division* may require approval by the *Name_of_Health_Division* IRB. Please contact *Name_of_Contact_Person* at *Phone_Number_of_Contact_Person* for instructions on obtaining *Name_of_Health_Division* IRB approval.

6. The data must be protected against inadvertent disclosure of confidential data. In the section below, please address the following issues: (If additional space is needed, please attach a separate sheet.)

a) How you will provide secure conditions to use and store the data:

b) Assurances that the data will be used only for the purposes of the study:

c) Assurances that confidential data will be destroyed at the conclusion of the research:

The review committee may require additional safeguards if it is determined that these are necessary due to unique or special characteristics of your proposed research.

7. Prior to the release of confidential data, assurances must be given that you have adequate financial resources to carry out the proposed research. Please document adequate project funding and attach supporting documentation. If additional space is needed, please attach a separate sheet.

8. Please complete the following Researcher Assurances Form on page 5.

Attachments (please check applicable boxes):

Research protocol attached	☐
IRB approval	☐
Project funding	☐
Peer review approval	☐
Researcher Assurances Form	☐

Date reviewed by *Name_of_Health_Division* administration: _____

Approved ☐ Denied ☐

Comments:

Researcher Assurances Form

The undersigned agrees to (initial each statement, sign and date):

___ accept responsibility for the ethical conduct of the study and the protection of the rights, privacy and welfare of the individuals whose private health information is retained in the *Name_of_Regsitry*;

___ conduct this study in compliance with the protocol as reviewed and approved by *Name_of_Regsitry* and/or the Advisory Committee;

___ submit all proposed study changes, including those accepted by an IRB, to *Name_of_Regsitry* to seek approval prior to implementing changes. This includes but is not limited to change in venue, change in PI or other investigators, change in study focus, and any change requiring IRB approval;

___ report upon discovery all unanticipated problems, protocol violations, and breaches of confidentiality to *Name_of_Regsitry*;

___ submit copies of literature and formal presentations generated using *Name_of_Regsitry* data;

___ pay all relevant fees prior to receiving *Name_of_Regsitry* data (see Schedule of Research Fees); and

___ complete dataset received from *Name_of_Regsitry* will be destroyed upon conclusion of the study and *Name_of_Regsitry* will be informed.

I agree to comply with the above requirements. I attest that information in this Research Proposal Review Form and attachments are true and complete. I also attest that I have no conflicts of interest to disclose regarding this study.

Non-compliance to this agreement may result in termination of the study approval. This means approval for use of *Name_of_Regsitry* study data may be revoked. If this occurs, proof is required that all data obtained from *Name_of_Regsitry* for the purposes of this study are destroyed. If this occurs, no investigator on this study may benefit from the use of *Name_of_Regsitry* data either monetarily, including grant funding, nor through publications, presentations, or any other means.

_____ Date _____ (PI signature)

This page is left blank intentionally.

APPENDIX B: ANNOTATED BIBLIOGRAPHY

The tables appearing on the following pages provide an annotated bibliography of the majority of previous work related to the field of geocoding to date. The manuscripts listed include those that are explicitly about the geocoding process itself, as well as those that make use of it as a part of the research presented and address an aspect of the geocoding process (both explicitly and implicitly). Each table focuses on a specific theme—the works listed within are further classified into which aspect of the theme they are relevant to. These tables should be used to guide the reader to the relevant works for further background reading on a topic and/or to see how other research studies have addressed and/or dealt with an issue related to the geocoding process.

Table 46 – Previous geocoding studies classified by topics of input data utilized

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Abe and Stinchcomb 2008	●		●	●	●	●	●	●	●	●	●	●	●	●
Agarwal 2004	●	●	●		●						●	●		
Agouris et al. 2000	●		●									●	●	
Agovino et al. 2005			●	●	●							●	●	
Alani 2001	●										●	●		●
Alani et al. 2003	●										●	●	●	
Amitay et al. 2004	●										●	●	●	●
Arampatzis et al. 2006	●	●	●		●		●				●	●		
Arbia et al. 1998														
Arikawa and Noaki 2005	●	●	●				●	●	●	●				
Arikawa et al. 2004	●	●	●											
Armstrong et al. 2008			●		●							●		
Armstrong and Tiwari 2008	●		●	●	●	●	●	●	●		●	●	●	●
Axelrod 2003	●										●	●	●	●
Bakshi et al. 2004			●									●		
Beal 2003		●	●									●		
Beaman et al. 2004	●	●	●		●		●	●	●		●	●	●	●
Berney and Blane 1997			●								●		●	●
Beyer et al. 2008			●	●	●							●	●	●
Bichler and Balchak 2007			●				●	●	●	●	●	●		●
Bilhaut et al. 2003	●	●	●	●	●		●	●	●		●	●		●
Blakely and Salmond 2002			●				●				●	●		
Block 1995			●											
Bonner et al. 2003			●										●	
Boscoe et al. 2002	●		●	●	●	●	●	●	●	●	●	●		●
Boscoe et al. 2004		●	●	●	●									
Boscoe 2008	●		●	●	●	●	●	●	●	●	●	●		●
Bow et al. 2004			●	●	●	●						●		●
Brody et al. 2002			●		●								●	
Cayo and Talbot 2003			●		●		●					●		
Chalasani et al. 2005	●	●	●				●	●	●		●	●		●
Chavez 2000	●	●										●	●	
Chen, C.C. et al. 2003			●											
Chen, C.C. et al. 2004			●											
Chen, M.F. et al. 1998			●				●	●	●					
Chen, W et al. 2004			●		●							●		
Chen et al. 2008			●								●	●		
Chou 1995	●	●	●											
Christen and Churches 2005			●				●	●	●	●	●	●		●
Christen et al. 2004			●				●	●	●	●	●	●		●
Chua 2001	●		●				●	●	●	●	●	●	●	●

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Chung et al. 2004			●											
Churches et al. 2002			●				●	●	●	●	●			●
Clough 2005	●	●	●		●		●	●	●		●	●		●
Collins et al. 1998			●		●							●		
Croner 2003			●	●	●							●	●	
Curtis et al. 2006			●									●		
Davis Jr. 1993			●											
Davis Jr. and Fonseca 2007	●	●	●	●	●	●	●	●	●	●	●	●		●
Davis Jr. et al. 2003	●	●	●	●	●	●	●	●	●	●	●	●		●
Dawes et al. 2006			●											
Dearwent et al. 2001			●		●		●	●	●		●	●		●
Densham and Reid 2003	●						●	●	●		●			
Diez-Roux et al. 2001			●											
Drummond 1995			●				●	●	●		●	●	●	●
Dueker 1974	●	●	●	●	●									
Durr and Froggatt 2002			●		●							●		
Efstathopoulos et al. 2005	●		●				●	●	●					
Eichelberger 1993			●	●	●		●	●	●	●	●	●		●
El-Yacoubi et al. 2002			●				●	●	●					
Fonda-Bonardi 1994			●					●	●		●			●
Foody 2003											●	●		
Fortney et al. 2000			●		●	●								
Fremont et al. 2005					●									
Frew et al. 1998	●						●	●	●		●	●	●	
Fu et al. 2005a	●	●					●	●	●		●	●		
Fu et al. 2005b	●	●					●	●	●		●	●		
Fulcomer et al. 1998			●	●	●	●	●	●	●	●	●	●		●
Gaffney et al. 2005			●		●								●	
Gatrell 1989			●		●									
Geronimus and Bound 1998					●									
Geronimus and Bound 1999a					●									
Geronimus and Bound 1999b					●									
Geronimus et al. 1995			●		●									
Gilboa et al. 2006			●				●							
Gilboa et al. 2006			●				●							
Goldberg et al. 2007	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Gregorio et al. 1999			●	●	●						●			●
Gregorio et al. 2005			●	●	●						●			●
Griffin et al. 1990	●		●		●	●								
Grubesic and Matisziw 2006			●		●						●	●		
Grubesic and Murray 2004			●											
Han et al. 2004			●	●	●		●	●	●	●	●		●	●
Han et al. 2005			●										●	

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Hariharan and Toyama 2004	●		●				●					●	●	
Haspel and Knotts 2005			●		●									
Henshaw et al. 2004			●											
Higgs and Martin 1995a			●		●							●	●	
Higgs and Martin 1995b	●		●		●		●				●	●		●
Higgs and Richards 2002					●									●
Hill 2000	●		●				●	●	●	●	●	●	●	●
Hill and Zheng 1999	●		●				●	●	●	●	●	●	●	●
Hill et al. 1999	●		●				●	●	●	●	●	●	●	●
Himmelstein 2005			●				●	●	●		●			●
Hurley et al. 2003			●	●							●	●		
Hutchinson and Veenendall 2005a	●	●	●	●	●		●	●	●		●	●		
Hutchinson and Veenendall 2005b	●	●	●	●	●		●	●	●		●	●		
Jaro 1984			●								●			●
Jaro 1989			●								●			●
Johnson 1998a			●	●	●		●	●	●		●	●	●	●
Johnson 1998b			●	●	●		●	●	●		●	●	●	●
Jones et al. 2001	●	●	●				●	●	●		●	●		
Karimi et al. 2004			●				●	●	●		●	●		●
Kennedy et al. 2003			●								●		●	●
Kim 2001			●				●	●	●	●	●	●	●	●
Kimler 2004	●	●					●	●	●	●	●	●		●
Krieger 1992			●									●		●
Krieger 2003			●	●	●							●	●	●
Krieger and Gordon 1999					●									
Krieger et al. 1997			●		●							●		
Krieger et al. 2001			●	●	●	●	●	●	●	●	●	●		●
Krieger et al. 2002a	●		●	●	●	●						●		
Krieger et al. 2002b			●	●	●						●	●		●
Krieger et al. 2003			●		●							●		
Krieger et al. 2005			●		●							●		
Krieger et al. 2006			●	●	●							●		
Kwok and Yankaskas 2001			●		●							●		
Laender et al. 2005	●	●	●	●	●		●	●	●		●	●		●
Lam et al. 2002	●	●									●	●		
Lee 2004		●	●				●	●	●			●	●	
Lee and McNally 1998	●	●	●	●	●						●	●	●	●
Leidner 2004	●						●	●	●		●	●		
Levesque 2003			●		●	●		●	●	●	●	●	●	●
Levine and Kim 1998	●	●	●				●	●	●	●	●	●	●	●
Li et al. 2002	●	●					●	●	●		●	●		
Lind 2001	●		●		●		●		●		●	●		●
Lind 2005	●		●		●		●		●		●	●		●

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Lovasi et al. 2007			●	●	●	●	●	●	●	●		●	●	●
Maizlish and Herrera 2005			●											
Markowetz 2004	●	●	●		●		●				●	●		
Markowetz et al. 2005	●	●	●		●		●				●	●		
Martin 1998			●		●			●	●			●	●	●
Martin and Higgs 1996			●		●									
Martins and Silva 2005	●						●				●	●		
Martins et al. 2005a	●						●				●	●		
Martins et al. 2005b	●						●				●	●		
Mazumdar et al. 2008			●								●	●		●
McCurley 2001	●	●	●	●	●		●	●	●		●	●		
McEathron et al. 2002	●						●	●	●		●	●		
McElroy et al. 2003			●	●	●	●	●	●	●	●	●	●	●	●
Mechanda and Puderer 2007			●	●	●		●	●	●	●	●	●	●	●
Michelson and Knoblock 2005			●											
Miner et al. 2005			●										●	
Ming et al. 2005														
Murphy and Armitage 2005	●		●		●		●	●	●	●	●	●		●
Nicoara 2005	●		●	●	●	●	●	●	●	●	●	●		●
Noaki and Arikawa 2005a	●	●	●				●	●	●		●	●		
Noaki and Arikawa 2005b	●	●	●				●	●	●		●	●		
Nuckols et al. 2004			●		●							●	●	●
O'Reagan and Saalfeld 1987			●		●									
Oliver et al. 2005	●	●	●	●	●	●					●	●		●
Olligschlaeger 1998			●								●			●
Oppong 1999			●								●	●	●	●
Paull 2003			●						●		●	●	●	●
Purves et al. 2005	●	●	●		●		●				●	●		
Ratcliffe 2001	●	●	●	●	●						●	●		●
Ratcliffe 2004	●	●	●	●	●		●	●	●	●	●	●		●
Rauch et al. 2003	●	●					●				●	●	●	
Reid 2003	●		●		●		●					●		
Reinbacher 2006	●	●					●	●	●		●	●		
Reinbacher et al. 2008	●	●					●	●	●		●	●		
Revie and Kerfoot 1997	●													
Riekert 2002	●						●							
Rose et al. 2004			●	●	●	●					●	●	●	●
Rull et al. 2006			●										●	
Rushton et al. 2006			●	●	●	●					●	●	●	●
Rushton et al. 2008b			●		●							●		
Schlieder et al. 2001	●		●		●		●				●	●		
Schockaert et al. 2005	●	●									●	●		

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Schootman et al. 2004	●	●	●	●	●	●	●	●	●	●	●	●		●
Sheehan et al. 2000			●		●							●		●
Shi 2007			●	●	●						●	●		●
Smith and Crane 2001	●						●				●			
Smith and Mann 2003	●						●				●			
Smith et al. 1999					●							●		
Soobader et al. 2001			●		●							●		
Southall 2003	●	●					●	●	●	●	●	●		●
Stevenson et al. 2000					●							●	●	
Strickland et al. 2007			●								●	●		
Temple et al. 2005			●							●				●
Tezuka and Tanaka 2005	●	●	●				●							
Thrall 2006			●				●	●	●	●	●	●	●	●
Tobler 1972	●		●		●						●	●		
Toral and Muñoz 2006	●										●	●	●	●
UN Economic Commission 2005	●	●	●	●	●						●	●	●	●
United States Department of Health and Human Services 2000			●											
Vaid et al. 2005	●	●	●		●									
Van Kreveld and Reinbacher 2004	●	●												
Vestavik 2004	●	●	●		●						●	●	●	
Vine et al. 1998			●	●	●	●					●	●	●	●
Vögele and Schlieder 2003	●		●		●		●				●	●		
Vögele and Stuckenschmidt 2001	●		●		●		●				●	●		
Vögele et al. 2003	●		●		●		●				●	●		
Waldinger et al. 2003	●	●	●		●						●	●		
Waller 2008			●									●		
Walls 2003	●	●	●		●		●	●	●	●	●	●	●	●
Ward et al. 2005			●	●	●	●						●		●
Werner 1974	●		●		●				●			●		
Whitsel et al. 2004		●	●					●	●	●	●		●	●
Whitsel et al. 2006		●	●	●	●	●		●	●	●		●		●
Wieczorek 2008	●	●	●								●	●	●	●
Wieczorek et al. 2004	●	●	●								●	●	●	●
Wilson, et al. 2004	●	●									●	●		
Winkler 1995			●								●			
Wong and Chuah 1994			●		●		●	●	●	●	●	●		●
Woodruff and Plaunt 1994	●	●					●	●			●	●	●	●
Wu et al. 2005			●		●							●		●
Yang et al. 2004		●	●	●	●	●	●	●	●	●		●		●
Yildirim and Yomralioglu 2004	●		●											
Yu 1996			●		●		●	●	●		●	●		●
Zandbergen 2007			●									●		

	Input Data													
	Types						Process				Accuracy			
	Named Places	Relative Descriptions	Postal Addresses	USPS PO Boxes	Postal Codes	Rural Routes	Parsing	Normalization	Standardization	Validation	Ambiguity	Resolution	Temporality	Correctness
Zandbergen 2008			●				●	●	●	●	●	●	●	●
Zandbergen and Chakraborty 2006			●											
Zimmerman 2006			●								●	●		●
Zimmerman 2008			●		●						●	●	●	●
Zimmerman et al. 2007			●								●	●		●
Zimmerman et al. 2008			●								●	●		●
Zong et al. 2005	●	●					●	●	●		●	●		●

Table 47 – Previous geocoding studies classified by topics of reference data source

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Abe and Stinchcomb 2008	●	●	●	●	●		●	●	●	●	●
Agouris et al. 2000	●	●		●			●	●	●	●	
Agovino et al. 2005			●	●					●	●	
Alani 2001	●	●		●			●	●	●	●	
Alani et al. 2003	●	●						●	●	●	
Amitay et al. 2004	●							●			
Arampatzis et al. 2006	●							●	●		
Arbia et al. 1998			●	●	●			●	●		
Arikawa and Noaki 2005	●	●	●								
Arikawa et al. 2004		●	●	●							
Armstrong et al. 1999		●	●	●				●	●		●
Armstrong et al. 2008			●	●				●	●		
Armstrong and Tiwari 2008	●		●	●				●	●		
Axelrod 2003	●	●	●	●				●	●	●	●
Bakshi et al. 2004			●					●	●		
Beal 2003		●	●								
Beaman et al. 2004	●	●									
Beyer et al. 2008				●				●	●	●	●
Bichler and Balchak 2007			●		●			●	●	●	●
Bilhaut et al. 2003	●	●		●				●	●		
Block 1995			●				●	●	●		●
Bonner et al. 2003			●				●		●	●	
Boscoe et al. 2002			●	●			●	●	●		●
Boscoe et al. 2004			●	●	●			●	●		●
Boscoe 2008			●					●	●		
Boulos 2004	●	●	●	●	●	●		●	●	●	●
Bow et al. 2004			●	●				●	●		
Brody et al. 2002		●	●	●	●			●	●	●	
Broome 2003			●				●	●	●	●	●
Can 1993		●	●	●				●	●	●	●
Cayo and Talbot 2003			●	●					●		
Chalasani et al. 2005	●	●	●	●	●						
Chavez 2000	●	●		●					●	●	
Chen, C.C. et al. 2003				●	●	●	●	●	●		
Chen, C.C. et al. 2004		●	●	●	●	●	●	●	●		
Chen, M.F. et al. 1998				●							
Chen, W et al. 2004		●	●	●	●			●	●		●
Chiang and Knoblock 2006		●	●		●		●		●		
Chiang et al. 2005		●	●		●		●		●		
Chou 1995			●								

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Christen and Churches 2005		●	●	●				●	●		
Christen et al. 2004		●		●				●	●		
Chua 2001		●	●	●				●	●		
Chung et al. 2004			●	●				●	●		
Churches et al. 2002		●		●							
Clough 2005	●	●		●				●	●		
Collins et al. 1998		●	●	●				●	●	●	
Cressie and Kornak 2003											
Croner 2003		●	●	●		●	●	●	●	●	
Curtis et al. 2006			●	●				●	●		
Davis Jr. 1993		●	●	●	●	●	●	●	●		
Davis Jr. and Fonseca 2007	●	●	●	●				●	●		
Davis Jr. et al. 2003	●	●	●	●				●	●		
Dawes et al. 2006				●		●	●	●	●	●	●
Dearwent et al. 2001			●	●				●	●		
Densham and Reid 2003	●										
Diez-Roux et al. 2001				●						●	
Drummond 1995			●	●				●	●	●	
Dueker 1974	●	●	●	●				●	●		●
Durr and Froggatt 2002		●	●	●				●	●		
Efstathopoulos et al. 2005			●								
Eichelberger 1993			●	●				●	●		
Fonda-Bonardi 1994			●	●			●	●			●
Foody 2003								●	●		
Fortney et al. 2000			●	●					●		
Frank et al. 2004		●	●	●	●			●	●	●	●
Fremont et al. 2005				●							
Frew et al. 1998	●	●	●	●				●	●	●	●
Fu et al. 2005a	●							●	●		
Fu et al. 2005b	●							●	●		
Fulcomer et al. 1998			●	●				●	●	●	
Gabrosek and Cressie 2002			●								
Gatrell 1989		●	●	●				●	●		
Geronimus and Bound 1998				●				●			
Geronimus and Bound 1999a				●				●			
Geronimus and Bound 1999b				●				●			
Geronimus et al. 1995				●				●			
Gilboa et al. 2006			●						●		
Goldberg et al. 2007	●	●	●	●	●	●	●	●	●	●	●
Goodchild and Hunter 1997			●	●				●	●		
Gregorio et al. 1999			●	●				●			
Gregorio et al. 2005			●	●				●			

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Griffin et al. 1990		●	●	●			●	●	●	●	●
Grubesic and Matisziw 2006			●	●				●	●		
Grubesic and Murray 2004		●	●					●	●	●	
Han et al. 2004			●						●	●	
Han et al. 2005			●							●	
Hariharan and Toyama 2004			●						●	●	
Haspel and Knotts 2005			●					●	●		
Henshaw et al. 2004		●								●	
Higgs and Martin 1995a		●		●				●	●	●	●
Higgs and Martin 1995b	●	●	●	●				●	●		●
Higgs and Richards 2002		●		●				●			
Hild and Fritsch 1998				●	●	●	●	●	●		
Hill 2000	●	●	●	●				●	●	●	
Hill and Zheng 1999	●	●	●	●				●	●	●	
Hill et al. 1999	●	●	●	●				●	●	●	
Hurley et al. 2003		●	●	●				●	●		
Hutchinson and Veenendall 2005a	●	●	●	●							
Hutchinson and Veenendall 2005b	●	●	●	●							
Johnson 1998a			●	●				●	●	●	
Johnson 1998b			●	●				●	●	●	
Jones et al. 2001	●	●		●				●	●		
Karimi et al. 2004			●	●				●	●		
Kennedy et al. 2003				●	●				●	●	
Kim 2001				●							
Kimler 2004	●	●						●	●		
Krieger 1992				●				●			
Krieger 2003								●	●	●	
Krieger and Gordon 1999				●				●			
Krieger et al. 1997				●				●			
Krieger et al. 2001				●				●	●		
Krieger et al. 2002a				●				●	●		
Krieger et al. 2002b				●				●	●		
Krieger et al. 2003				●				●			
Krieger et al. 2005				●				●			
Krieger et al. 2006			●	●				●	●		
Kwok and Yankaskas 2001			●	●				●			
Laender et al. 2005	●	●	●	●				●	●		●
Lam et al. 2002	●	●		●				●	●		
Lee 2004			●	●				●	●	●	
Lee and McNally 1998			●	●				●	●	●	
Levesque 2003	●	●	●	●	●	●	●	●	●	●	●
Levine and Kim 1998	●	●	●	●		●	●	●	●		

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Li et al. 2002	●							●	●		
Lind 2001	●	●	●	●			●	●	●		
Lind 2005	●	●	●	●			●	●	●		
Lovasi et al. 2007			●	●	●			●	●	●	
Markowetz 2004	●	●		●							
Markowetz et al. 2005	●	●		●							
Martin 1998		●	●	●			●	●	●	●	
Martin and Higgs 1996		●		●			●	●	●		
Martins and Silva 2005	●										
Martins et al. 2005a	●										
Martins et al. 2005b	●										
Mazumdar et al. 2008	●	●	●	●	●			●	●		●
McCurley 2001	●	●	●	●				●	●		
McEathron et al. 2002	●	●	●	●	●		●	●	●	●	
McElroy et al. 2003			●	●				●	●	●	
Mechanda and Puderer 2007		●	●	●				●	●	●	●
Miner et al. 2005				●							
Ming et al. 2005			●	●	●		●	●	●		
Murphy and Armitage 2005	●	●	●	●				●	●		
Nicoara 2005	●	●	●	●			●	●	●	●	
Noaki and Arikawa 2005a	●		●						●		
Noaki and Arikawa 2005b	●		●						●		
Nuckols et al. 2004			●	●				●	●	●	
O'Reagan and Saalfeld 1987		●	●	●				●	●	●	
Oliver et al. 2005			●	●				●	●		
Olligschlaeger 1998	●	●	●	●			●	●	●	●	
Openshaw 1989		●	●	●				●	●	●	
Oppong 1999								●	●	●	●
Paull 2003		●	●	●			●	●	●	●	
Purves et al. 2005	●							●	●		
Ratcliffe 2001		●	●	●	●			●	●		
Ratcliffe 2004		●	●	●				●	●		
Rauch et al. 2003	●							●	●	●	
Reid 2003	●	●		●				●	●		
Reinbacher 2006	●	●		●				●	●		
Reinbacher et al. 2008	●	●		●				●	●		
Revie and Kerfoot 1997	●	●		●				●	●		
Riekert 2002	●										
Rose et al. 2004			●	●				●	●	●	
Rull et al. 2006			●	●					●	●	
Rushton et al. 2006		●	●	●				●	●	●	
Rushton et al. 2008b			●	●				●			

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Schlieder et al. 2001	●	●		●				●	●		
Schockaert et al. 2005				●				●	●		
Schootman et al. 2004			●	●				●	●		●
Sheehan et al. 2000			●	●				●	●		
Shi 2007			●	●				●	●		
Smith and Crane 2001	●										
Smith and Mann 2003	●										
Smith et al. 1999				●				●			
Soobader et al. 2001				●				●			
Southall 2003	●	●		●				●	●		
Stevenson et al. 2000		●		●				●	●	●	
Strickland et al. 2007		●	●	●	●			●	●		
Temple et al. 2005			●	●		●	●	●	●	●	●
Thrall 2006			●	●							
Toral and Muñoz 2006	●							●	●	●	
UN Economic Commission 2005	●	●	●	●			●	●	●	●	●
Vaid et al. 2005	●	●	●	●							
Van Kreveld and Reinbacher 2004				●							
Veregin 1999			●	●				●	●		
Vestavik 2004	●	●	●	●							
Vine et al. 1998			●	●	●		●	●	●	●	●
Vögele and Schlieder 2003	●	●		●				●	●		
Vögele and Stuckenschmidt 2001	●	●		●				●	●		
Vögele et al. 2003	●	●		●				●	●		
Waldinger et al. 2003	●	●	●	●							
Waller 2008		●	●	●				●	●		●
Walls 2003	●	●	●						●	●	●
Ward et al. 2005			●	●	●			●	●		
Werner 1974		●	●	●				●			
Whitsel et al. 2004			●	●				●	●	●	
Whitsel et al. 2006			●	●				●	●		
Wieczorek 2008	●	●	●	●	●			●	●	●	
Wieczorek et al. 2004	●	●	●	●	●			●	●	●	
Wilson, et al. 2004	●	●		●				●	●		
Woodruff and Plaunt 1994	●			●				●	●		
Wu et al. 2005		●	●	●		●	●	●	●		
Yang et al. 2004			●	●				●	●		
Yildirim and Yomralioglu 2004			●								
Yu 1996			●					●	●		
Zandbergen 2007			●	●				●	●		
Zandbergen 2008		●	●	●				●	●	●	●
Zandbergen and Chakraborty 2006			●	●					●		

	Reference Data Source										
	Type					Process		Accuracy			
	Gazetteer	Point-Based	Line-Based	Area Unit-Based	Imagery	Conflation	Adding Value	Resolution	Spatial	Temporality	Non-Spatial Attributes
Zimmerman 2006		●	●	●	●			●	●		●
Zimmerman 2008		●	●	●				●	●	●	●
Zimmerman et al. 2007		●	●	●	●			●	●		●
Zimmerman et al. 2007		●	●	●	●			●	●		●
Zong et al. 2005	●	●									

Table 48 – Previous geocoding studies classified by topics of feature-matching approach

	Matching					
	Type		Process		Accuracy	
	Deterministic	Probability-Based	String Comparison	Relaxation	Match Type	Match Rate
Abe and Stinchcomb 2008	●	●	●	●	●	●
Agouris et al. 2000		●				
Alani 2001	●					
Amitay et al. 2004		●				
Arampatzis et al. 2006	●				●	●
Armstrong et al. 2008					●	
Armstrong and Tiwari 2008	●	●	●	●	●	●
Bakshi et al. 2004	●					
Beal 2003	●		●		●	
Beaman et al. 2004					●	
Beyer et al. 2008					●	
Bichler and Balchak 2007	●	●	●	●	●	●
Bilhaut et al. 2003		●			●	
Blakely and Salmond 2002	●	●	●	●	●	●
Block 1995					●	●
Bonner et al. 2003					●	●
Boscoe et al. 2002					●	●
Boscoe 2008	●	●	●	●	●	●
Bow et al. 2004	●		●		●	●
Cayo and Talbot 2003						●
Chavez 2000	●					
Chen, M.F. et al. 1998			●			●
Chen, W et al. 2004					●	●
Chen et al. 2008					●	
Chiang and Knoblock 2006		●				
Chiang et al. 2005		●				
Christen and Churches 2005		●			●	●
Christen et al. 2004		●			●	●
Chua 2001	●		●		●	●
Chung et al. 2004						
Churches et al. 2002	●	●	●	●	●	●
Clough 2005	●	●	●		●	●
Davis Jr. 1993					●	●
Davis Jr. and Fonseca 2007	●		●		●	●
Davis Jr. et al. 2003	●		●		●	●
Dearwent et al. 2001	●		●	●	●	●
Densham and Reid 2003	●	●				
Drummond 1995	●	●	●	●	●	●
Durr and Froggatt 2002					●	●
Efstathopoulos et al. 2005	●		●			
Eichelberger 1993	●				●	
El-Yacoubi et al. 2002		●				

	Matching					
	Type		Process		Accuracy	
	Deterministic	Probability-Based	String Comparison	Relaxation	Match Type	Match Rate
Fu et al. 2005a	●	●			●	
Fu et al. 2005b	●	●			●	
Fulcomer et al. 1998	●		●	●	●	●
Gabrosek and Cressie 2002						●
Gilboa et al. 2006	●		●	●	●	●
Goldberg et al. 2007	●	●	●	●	●	●
Goodchild and Hunter 1997						
Gregorio et al. 1999	●		●	●	●	●
Gregorio et al. 2005	●		●	●	●	●
Grubesic and Matisziw 2006					●	
Grubesic and Murray 2004	●	●	●	●	●	●
Han et al. 2004						●
Hariharan and Toyama 2004		●				
Haspel and Knotts 2005					●	●
Higgs and Martin 1995b	●				●	●
Higgs and Richards 2002					●	●
Hill 2000	●		●			
Hill and Zheng 1999	●		●			
Hill et al. 1999	●		●			
Hurley et al. 2003					●	●
Jaro 1984	●	●	●	●	●	●
Jaro 1989	●	●	●	●	●	●
Johnson 1998a	●		●		●	●
Johnson 1998b	●		●		●	●
Jones et al. 2001	●				●	
Karimi et al. 2004	●	●	●	●	●	●
Kimler 2004	●	●	●			●
Krieger 1992						●
Krieger 2003					●	●
Krieger et al. 2001					●	●
Krieger et al. 2002a					●	●
Krieger et al. 2002b					●	●
Krieger et al. 2003					●	●
Krieger et al. 2005					●	●
Krieger et al. 2006					●	●
Kwok and Yankaskas 2001					●	●
Laender et al. 2005	●	●	●		●	●
Lam et al. 2002	●					
Lee 2004	●		●		●	
Lee and McNally 1998					●	
Leidner 2004	●	●			●	
Levine and Kim 1998	●		●	●	●	●
Li et al. 2002	●	●				
Lind 2001					●	

	Matching					
	Type		Process		Accuracy	
	Deterministic	Probability-Based	String Comparison	Relaxation	Match Type	Match Rate
Lind 2005					●	
Lovasi et al. 2007	●	●			●	●
MacDorman and Gay 1999						●
Maizlish and Herrera 2005	●	●	●	●	●	●
Markowetz 2004	●	●				
Markowetz et al. 2005	●	●				
Martin and Higgs 1996						●
Martins and Silva 2005	●					●
Martins et al. 2005a	●					●
Martins et al. 2005b	●					●
Mazumdar et al. 2008	●	●			●	●
McCurley 2001	●	●	●			●
McElroy et al. 2003	●	●	●	●	●	●
Mechanda and Puderer 2007	●				●	
Meyer et al. 2005	●	●	●	●	●	●
Michelson and Knoblock 2005		●	●		●	●
Ming et al. 2005		●				
Murphy and Armitage 2005	●	●	●	●	●	●
Nicoara 2005	●	●	●	●	●	●
Nuckols et al. 2004						●
O'Reagan and Saalfeld 1987	●	●	●	●	●	●
Oliver et al. 2005					●	●
Olligschlaeger 1998	●				●	●
Paull 2003						●
Porter 1980	●		●			
Purves et al. 2005	●				●	●
Raghavan et al. 1989		●			●	●
Ratcliffe 2001	●		●			●
Ratcliffe 2004	●		●		●	●
Rauch et al. 2003	●	●	●			
Reid 2003	●					
Reinbacher 2006	●	●			●	
Reinbacher et al. 2008	●	●			●	
Revie and Kerfoot 1997	●					
Riekert 2002	●					
Rose et al. 2004					●	●
Rull et al. 2006						●
Rushton et al. 2006	●	●	●	●	●	●
Rushton et al. 2008b					●	
Schootman et al. 2004					●	●
Schumacher 2007	●	●	●	●	●	●
Shi 2007				●	●	●
Soobader et al. 2001						●

	Matching					
	Type		Process		Accuracy	
	Deterministic	Probability-Based	String Comparison	Relaxation	Match Type	Match Rate
Strickland et al. 2007	●	●			●	●
Tezuka and Tanaka 2005		●				
Thrall 2006	●	●	●	●	●	●
Vine et al. 1998					●	●
Waldinger et al. 2003	●					
Waller 2008					●	
Walls 2003	●		●		●	●
Ward et al. 2005	●		●	●	●	●
Whitsel et al. 2004					●	●
Whitsel et al. 2006	●	●	●	●	●	●
Wieczorek 2008	●					●
Wieczorek et al. 2004	●					●
Wilson, et al. 2004	●					
Winkler 1995	●	●	●	●	●	●
Wong and Chuah 1994	●	●	●	●	●	●
Woodruff and Plaunt 1994	●		●			
Wu et al. 2005		●	●	●	●	●
Yang et al. 2004	●	●	●	●	●	●
Yu 1996	●	●	●	●	●	●
Zandbergen 2007					●	●
Zandbergen 2008	●	●	●	●	●	●
Zandbergen and Chakraborty 2006					●	●
Zimmerman 2006	●				●	●
Zimmerman 2008					●	●
Zimmerman et al. 2007	●				●	●
Zimmerman et al. 2008					●	●
Zong et al. 2005	●				●	●

Table 49 – Previous geocoding studies classified by topics of feature interpolation method

	Interpolation					
	Type			Accuracy		
	Point-Based	Line-Based	Area Unit-Based	Point-Based	Line-Based	Area Unit-Based
Abe and Stinchcomb 2008	●	●	●	●	●	●
Agouris et al. 2000			●		●	
Agovino et al. 2005		●				
Alani 2001			●			●
Amitay et al. 2004	●		●			
Armstrong et al. 2008		●	●		●	●
Armstrong and Tiwari 2008		●	●		●	●
Bakshi et al. 2004		●			●	
Beal 2003	●	●	●	●	●	●
Beyer et al. 2008			●			●
Bichler and Balchak 2007		●			●	
Bilhaut et al. 2003	●		●	●		●
Boscoe et al. 2002		●	●			
Boscoe 2008	●	●	●		●	●
Bow et al. 2004	●	●	●	●	●	●
Brody et al. 2002	●	●	●			
Cayo and Talbot 2003		●	●		●	●
Chalasani et al. 2005	●	●	●	●	●	●
Chen, C.C. et al. 2003			●			●
Chen, C.C. et al. 2004	●	●	●	●	●	●
Chen, M.F. et al. 1998			●			●
Chen, W et al. 2004			●			●
Chen et al. 2008				●	●	●
Chiang and Knoblock 2006	●	●		●	●	
Chiang et al. 2005	●	●		●	●	
Christen and Churches 2005			●			●
Christen et al. 2004			●			●
Chua 2001	●	●	●	●	●	●
Chung et al. 2004		●	●		●	●
Churches et al. 2002			●			●
Clough 2005	●		●	●		●
Collins et al. 1998	●	●	●	●	●	●
Cressie and Kornak 2003				●	●	●
Curtis et al. 2006	●	●	●	●	●	●
Davis Jr. 1993	●	●	●	●	●	●
Davis Jr. and Fonseca 2007	●	●	●	●	●	●
Davis Jr. et al. 2003	●	●	●	●	●	●
Dearwent et al. 2001		●	●		●	●
Densham and Reid 2003	●			●		
Drummond 1995		●	●		●	
Dueker 1974	●	●	●			
Durr and Froggatt 2002	●	●	●	●	●	●

	Interpolation					
	Type			Accuracy		
	Point-Based	Line-Based	Area Unit-Based	Point-Based	Line-Based	Area Unit-Based
Fortney et al. 2000		●	●		●	●
Fu et al. 2005a	●		●	●		●
Fu et al. 2005b	●		●	●		●
Fulcomer et al. 1998		●	●			
Gatrell 1989			●			●
Gilboa et al. 2006		●			●	
Goldberg et al. 2007	●	●	●	●	●	●
Goodchild and Hunter 1997		●	●		●	●
Gregorio et al. 1999		●	●			●
Gregorio et al. 2005			●			●
Grubesic and Matisziw 2006		●	●			●
Grubesic and Murray 2004		●			●	
Han et al. 2004		●			●	
Haspel and Knotts 2005		●				
Henshaw et al. 2004	●		●			
Higgs and Martin 1995b	●	●	●			
Higgs and Richards 2002	●		●	●		●
Hill 2000	●	●	●			
Hill and Zheng 1999	●	●	●			
Hill et al. 1999	●	●	●			
Hurley et al. 2003		●	●		●	●
Hutchinson and Veenendall 2005a	●	●	●			
Hutchinson and Veenendall 2005b	●	●	●			
Johnson 1998a	●	●	●			
Johnson 1998b	●	●	●			
Karimi et al. 2004		●	●		●	●
Kennedy et al. 2003			●			
Kimler 2004	●			●		
Krieger et al. 2001			●			●
Krieger et al. 2002a			●			●
Krieger et al. 2002b			●			●
Krieger et al. 2003			●			●
Krieger et al. 2005			●			●
Krieger et al. 2006		●	●		●	●
Kwok and Yankaskas 2001		●	●			
Laender et al. 2005	●	●	●			
Lam et al. 2002	●		●	●		●
Lee 2004		●			●	
Lee and McNally 1998		●	●			
Levine and Kim 1998	●	●	●		●	
Li et al. 2002	●		●	●		●
Lind 2001	●	●	●			
Lind 2005	●	●	●			
Lovasi et al. 2007		●	●		●	●

	Interpolation					
	Type			Accuracy		
	Point-Based	Line-Based	Area Unit-Based	Point-Based	Line-Based	Area Unit-Based
Markowetz 2004	●		●			
Markowetz et al. 2005	●		●			
Martin 1998		●	●		●	●
Mazumdar et al. 2008	●	●	●	●	●	●
McCurley 2001	●	●	●	●	●	●
McEathron et al. 2002	●	●	●	●	●	●
McElroy et al. 2003		●	●		●	●
Miner et al. 2005			●			
Ming et al. 2005		●	●		●	●
Murphy and Armitage 2005	●		●			
Nicoara 2005	●	●	●			
Noaki and Arikawa 2005a		●				
Noaki and Arikawa 2005b		●				
Oliver et al. 2005		●	●			●
Olligschlaeger 1998		●	●			
Ratcliffe 2001		●	●		●	●
Ratcliffe 2004		●	●		●	●
Rauch et al. 2003	●			●		
Reid 2003	●		●			
Reinbacher 2006	●		●	●		●
Reinbacher et al. 2008	●		●	●		●
Revie and Kerfoot 1997	●		●	●		●
Rose et al. 2004		●	●		●	●
Rull et al. 2006		●				
Rushton et al. 2006	●	●	●	●	●	●
Rushton et al. 2008b		●	●			
Sadahiro 2000			●			●
Schlieder et al. 2001	●	●	●	●	●	●
Schockaert et al. 2005			●			●
Schootman et al. 2004		●	●		●	●
Sheehan et al. 2000		●	●		●	●
Shi 2007		●	●		●	●
Southall 2003	●					
Stevenson et al. 2000	●		●			
Strickland et al. 2007	●	●	●	●	●	●
Thrall 2006		●	●			
Tobler 1972	●	●	●			
Van Kreveld and Reinbacher 2004			●			●
Vine et al. 1998		●	●			
Vögele and Schlieder 2003	●	●	●	●	●	●
Vögele and Stuckenschmidt 2001	●	●	●	●	●	●
Vögele et al. 2003	●	●	●	●	●	●
Waller 2008	●	●	●	●	●	●
Walls 2003	●	●		●	●	

	Interpolation					
	Type			Accuracy		
	Point-Based	Line-Based	Area Unit-Based	Point-Based	Line-Based	Area Unit-Based
Ward et al. 2005		●	●		●	●
Whitsel et al. 2004		●	●		●	●
Whitsel et al. 2006		●	●		●	●
Wieczorek 2008	●	●	●	●	●	●
Wieczorek et al. 2004	●	●	●	●	●	●
Wilson, et al. 2004	●		●	●		●
Woodruff and Plaunt 1994			●			●
Wu et al. 2005		●			●	
Yang et al. 2004		●	●		●	●
Yu 1996		●			●	
Zandbergen 2007		●	●		●	●
Zandbergen 2008	●	●	●	●	●	●
Zandbergen and Chakraborty 2006		●			●	●
Zimmerman 2006		●	●		●	●
Zimmerman 2008				●	●	●
Zimmerman et al. 2007		●	●		●	●
Zong et al. 2005	●			●		

Table 50 – Previous geocoding studies classified by topics of accuracy measured utilized

	Accuracy							
	Measures				Estimates			
	Spatial	Temporality	Resolution	Bias Introduction	Quality Codes	Spatial	Temporality	Resolution
Abe and Stinchcomb 2008	●	●	●	●	●	●	●	●
Agouris et al. 2000	●	●	●			●	●	●
Agovino et al. 2005	●	●	●					
Alani 2001	●	●	●			●		
Alani et al. 2003	●	●	●					
Amitay et al. 2004	●		●					
Arampatzis et al. 2006	●		●			●		●
Arikawa et al. 2004	●		●	●	●	●		●
Armstrong et al. 2008	●		●			●		●
Armstrong and Tiwari 2008	●		●			●		●
Axelrod 2003	●	●	●					
Bakshi et al. 2004	●					●		
Beal 2003	●					●		●
Beaman et al. 2004	●				●			
Berney and Blane 1997		●					●	
Beyer et al. 2008	●	●	●			●	●	●
Bichler and Balchak 2007			●	●		●		
Bilhaut et al. 2003	●		●			●		●
Blakely and Salmond 2002					●			
Bonner et al. 2003	●	●	●	●		●	●	●
Boscoe et al. 2002				●				
Boscoe 2008	●	●	●		●	●		●
Bow et al. 2004	●		●	●		●		●
Brody et al. 2002	●	●	●	●	●	●	●	●
Casady 1999	●	●	●			●	●	●
Cayo and Talbot 2003	●			●		●		●
Chalasani et al. 2005	●		●					
Chavez 2000	●	●					●	
Chen, C.C. et al. 2003	●		●					
Chen, C.C. et al. 2004	●					●		
Chen, W et al. 2004	●		●	●		●		●
Chen et al. 2008	●		●			●		●
Christen and Churches 2005			●		●			
Christen et al. 2004			●		●			
Chua 2001			●					
Churches et al. 2002	●		●	●				
Clough 2005	●		●			●		●
Collins et al. 1998	●	●	●			●		●
Cressie and Kornak 2003	●	●	●	●				
Curtis et al. 2006	●		●			●		●
Davis Jr. 1993	●		●		●	●		●
Davis Jr. and Fonseca 2007	●		●	●	●	●		●

	Accuracy							
	Measures				Estimates			
	Spatial	Temporality	Resolution	Bias Introduction	Quality Codes	Spatial	Temporality	Resolution
Davis Jr. et al. 2003	●		●	●	●	●		●
Dearwent et al. 2001	●		●	●		●		●
Diez-Roux et al. 2001		●		●				
Dru and Saada 2001	●					●		
Drummond 1995	●	●	●	●	●	●		●
Dueker 1974	●		●					
Durr and Froggatt 2002	●		●	●		●		●
Fonda-Bonardi 1994	●		●					
Foody 2003	●	●	●					
Fortney et al. 2000	●		●	●	●	●		●
Fremont et al. 2005			●	●				
Frew et al. 1998	●	●	●		●	●	●	●
Fu et al. 2005a	●		●			●		●
Fu et al. 2005b	●		●			●		●
Fulcomer et al. 1998	●	●	●		●	●	●	●
Gabrosek and Cressie 2002	●		●	●		●		●
Gaffney et al. 2005	●	●						
Gatrell 1989	●		●			●		●
Geronimus and Bound 1998			●	●				●
Geronimus and Bound 1999a			●	●				●
Geronimus and Bound 1999b			●	●				●
Geronimus et al. 1995			●	●				●
Gilboa et al. 2006	●		●			●		
Goldberg et al. 2007	●	●	●	●	●	●	●	●
Goodchild and Hunter 1997	●		●			●		●
Gregorio et al. 1999			●	●	●			●
Gregorio et al. 2005			●	●	●			●
Grubesic and Matisziw 2006	●		●	●		●		●
Grubesic and Murray 2004	●			●		●		
Han et al. 2004	●	●				●	●	
Han et al. 2005		●					●	
Hariharan and Toyama 2004	●	●	●			●	●	●
Haspel and Knotts 2005			●					
Henshaw et al. 2004	●	●				●	●	
Higgs and Martin 1995a								
Higgs and Martin 1995b	●		●					
Higgs and Richards 2002	●		●	●		●		●
Hill 2000	●	●	●			●	●	●
Hill and Zheng 1999	●	●	●			●	●	●
Hill et al. 1999	●	●	●			●	●	●
Himmelstein 2005	●				●	●		●
Hurley et al. 2003	●		●	●	●	●		●
Hutchinson and Veenendall 2005a	●		●					
Hutchinson and Veenendall 2005b	●		●					

	Accuracy							
	Measures				Estimates			
	Spatial	Temporality	Resolution	Bias Introduction	Quality Codes	Spatial	Temporality	Resolution
Jones et al. 2001	●		●			●		●
Karimi et al. 2004	●		●		●	●		●
Kennedy et al. 2003	●	●				●	●	
Kimler 2004	●		●			●		●
Krieger 1992			●	●				●
Krieger 2003	●	●	●	●				
Krieger and Gordon 1999			●	●				●
Krieger et al. 1997			●					
Krieger et al. 2001				●				
Krieger et al. 2002a	●		●	●	●	●		●
Krieger et al. 2002b	●		●	●		●		●
Krieger et al. 2003			●	●				●
Krieger et al. 2005			●	●				●
Krieger et al. 2006			●	●		●		●
Kwok and Yankaskas 2001			●		●			●
Laender et al. 2005	●		●					
Lam et al. 2002	●		●			●		●
Lee 2004	●	●	●			●		●
Lee and McNally 1998	●	●	●					
Levesque 2003	●	●	●			●	●	●
Levine and Kim 1998	●		●		●	●		●
Li et al. 2002	●		●			●		●
Lind 2001	●		●					
Lind 2005	●		●					
Lovasi et al. 2007	●	●	●	●	●	●	●	●
Markowetz 2004	●		●					
Markowetz et al. 2005	●		●					
Martin 1998	●	●	●	●		●	●	●
Martin and Higgs 1996	●		●			●		●
Martins et al. 2005b	●		●			●		●
Mazumdar et al. 2008	●		●	●	●	●		●
McCurley 2001	●							
McEathron et al. 2002	●		●					
McElroy et al. 2003	●	●	●		●	●	●	●
Mechanda and Puderer 2007	●	●	●		●	●	●	●
Murphy and Armitage 2005			●		●			●
Nicoara 2005	●		●					
Noaki and Arikawa 2005a	●							
Noaki and Arikawa 2005b	●							
Nuckols et al. 2004	●	●	●					
O'Reagan and Saalfeld 1987		●	●				●	●
Oliver et al. 2005	●		●	●		●		●
Olligschlaeger 1998	●		●					
Oppong 1999	●	●	●					

	Accuracy							
	Measures				Estimates			
	Spatial	Temporality	Resolution	Bias Introduction	Quality Codes	Spatial	Temporality	Resolution
Paull 2003	●	●	●					
Purves et al. 2005	●		●			●		●
Ratcliffe 2001	●		●			●		●
Ratcliffe 2004	●		●	●		●		●
Rauch et al. 2003	●	●	●					
Reid 2003	●		●					
Reinbacher 2006	●		●			●		●
Reinbacher et al. 2008	●		●			●		●
Revie and Kerfoot 1997	●		●		●	●		●
Rose et al. 2004	●	●	●	●	●	●	●	●
Rull et al. 2006	●	●				●	●	
Rushton et al. 2006	●	●	●	●	●	●	●	●
Rushton et al. 2008b			●					●
Sadahiro 2000				●				
Schlieder et al. 2001	●		●			●		●
Schockaert et al. 2005	●		●			●		●
Schootman et al. 2004	●		●	●	●	●		●
Sheehan et al. 2000	●		●	●	●	●		●
Shi 2007	●		●	●	●	●		●
Smith et al. 1999			●	●				●
Soobader et al. 2001			●	●				●
Southall 2003	●		●					
Stevenson et al. 2000		●	●					
Strickland et al. 2007	●		●			●		●
Temple et al. 2005	●	●	●			●	●	●
Thrall 2006					●			
Vaid et al. 2005	●		●					
Vestavik 2004	●	●	●					
Vine et al. 1998	●	●	●		●			
Vögele and Schlieder 2003	●		●			●		●
Vögele and Stuckenschmidt 2001	●		●			●		●
Vögele et al. 2003	●		●			●		●
Waldinger et al. 2003	●		●			●		●
Waller 2008	●		●			●		●
Walls 2003	●	●	●					
Ward et al. 2005	●		●	●		●		●
Werner 1974			●					
Whitsel et al. 2004	●	●			●	●	●	
Whitsel et al. 2006	●		●	●	●	●		●
Wieczorek 2008	●	●	●		●	●		
Wieczorek et al. 2004	●	●	●		●	●		
Wilson, et al. 2004	●		●			●		●
Woodruff and Plaunt 1994	●		●			●		●
Wu et al. 2005	●		●	●	●	●		●

	Accuracy							
	Measures				Estimates			
	Spatial	Temporality	Resolution	Bias Introduction	Quality Codes	Spatial	Temporality	Resolution
Yang et al. 2004	●		●					●
Yu 1996	●		●			●		●
Zandbergen 2007	●		●	●		●		●
Zandbergen 2008	●	●	●			●	●	●
Zandbergen and Chakraborty 2006	●			●		●		
Zimmerman 2006	●		●		●	●		●
Zimmerman 2008	●	●	●	●	●	●	●	●
Zimmerman et al. 2007	●		●		●	●		●
Zimmerman et al. 2008	●		●			●		●

Table 51 – Previous geocoding studies classified by topics of process used

	Process				
	Manual			Auto	
	GPS	Raster Imagery / Maps	Supplemental Data	Batch-Mode	Single-Mode
Abe and Stinchcomb 2008	●	●	●	●	●
Agouris et al. 2000		●		●	
Agovino et al. 2005				●	●
Arampatzis et al. 2006				●	
Armstrong and Tiwari 2008	●				●
Bakshi et al. 2004			●	●	
Beal 2003	●				
Beaman et al. 2004				●	●
Beyer et al. 2008					●
Bichler and Balchak 2007				●	●
Bilhaut et al. 2003				●	
Bonner et al. 2003	●			●	
Boscoe et al. 2002			●	●	
Boscoe 2008			●	●	●
Bow et al. 2004				●	
Brody et al. 2002		●	●		
Cayo and Talbot 2003	●			●	●
Chalasani et al. 2005	●	●	●	●	●
Chavez 2000					
Chen, C.C. et al. 2003		●		●	
Chen, C.C. et al. 2004		●	●	●	
Chen, W et al. 2004				●	
Christen and Churches 2005				●	●
Christen et al. 2004				●	●
Clough 2005				●	
Curtis et al. 2006	●	●	●		
Dao et al. 2002	●				
Davis Jr. 1993		●			●
Dearwent et al. 2001				●	●
Dru and Saada 2001	●				
Drummond 1995				●	●
Dueker 1974				●	
Durr and Froggatt 2002		●		●	●
Fortney et al. 2000	●			●	●
Frew et al. 1998				●	
Fulcomer et al. 1998				●	
Gilboa et al. 2006				●	
Goldberg et al. 2007	●	●	●		
Gregorio et al. 1999				●	
Gregorio et al. 2005				●	
Han et al. 2004			●	●	●

	Process				
	Manual			Auto	
	GPS	Raster Imagery / Maps	Supplemental Data	Batch-Mode	Single-Mode
Haspel and Knotts 2005				●	
Henshaw et al. 2004	●	●			
Higgs and Martin 1995b				●	●
Hill 2000				●	
Hill and Zheng 1999				●	
Hill et al. 1999				●	
Hurley et al. 2003			●	●	●
Hutchinson and Veenendall 2005a				●	
Hutchinson and Veenendall 2005b				●	
Karimi et al. 2004	●			●	●
Kennedy et al. 2003		●		●	●
Krieger 1992				●	
Krieger 2003			●	●	●
Krieger et al. 2001				●	
Krieger et al. 2002a				●	
Krieger et al. 2002b				●	
Krieger et al. 2003				●	
Krieger et al. 2005				●	
Krieger et al. 2006				●	
Kwok and Yankaskas 2001			●	●	
Lee 2004	●			●	
Lee and McNally 1998				●	
Levesque 2003		●			
Levine and Kim 1998			●	●	
Lovasi et al. 2007			●	●	●
MacDorman and Gay 1999				●	
Mazumdar et al. 2008		●		●	●
McEathron et al. 2002		●	●	●	●
McElroy et al. 2003			●	●	●
Mechanda and Puderer 2007			●	●	
Ming et al. 2005		●			
Nicoara 2005				●	
Olligschlaeger 1998				●	
Purves et al. 2005				●	
Ratcliffe 2001		●		●	
Rauch et al. 2003				●	
Rose et al. 2004			●	●	●
Rushton et al. 2006	●	●		●	●
Rushton et al. 2008b				●	
Schootman et al. 2004			●	●	●
Strickland et al. 2007		●		●	●
Temple et al. 2005	●	●	●		●

	Process				
	Manual			Auto	
	GPS	Raster Imagery / Maps	Supplemental Data	Batch-Mode	Single-Mode
Thrall 2006				●	
Vine et al. 1998	●	●	●	●	●
Ward et al. 2005	●	●	●	●	●
Whitsel et al. 2004				●	
Whitsel et al. 2006				●	●
Wieczorek 2008	●	●	●	●	●
Wieczorek et al. 2004	●	●	●	●	●
Wu et al. 2005	●				
Yang et al. 2004				●	●
Zandbergen 2007				●	
Zandbergen 2008				●	●
Zandbergen and Chakraborty 2006				●	
Zimmerman 2006	●	●		●	●
Zimmerman et al. 2007	●	●		●	●

Table 52 – Previous geocoding studies classified by topics of privacy concern and/or method

	Privacy				
	Type		Process		
	Data Leak	Self-Identifying	Masking	Randomization	Aggregation
Arikawa et al. 2004	●	●	●	●	●
Armstrong et al. 1999		●	●	●	●
Beyer et al. 2008		●			●
Boscoe et al. 2004			●	●	●
Boulos 2004		●	●	●	●
Brownstein et al. 2006		●	●		
Casady 1999	●	●			●
Chen et al. 2008		●	●	●	●
Christen and Churches 2005	●	●			
Churches et al. 2002	●	●	●		●
Croner 2003		●	●		●
Curtis et al. 2006		●	●	●	●
Dao et al. 2002	●	●			
Gittler 2008a	●	●	●		
Goldberg et al. 2007	●	●	●	●	●
MacDorman and Gay 1999		●			
Mazumdar et al. 2008		●	●	●	●
Miner et al. 2005	●	●			
Oppong 1999		●	●		●
Rushton et al. 2006		●	●	●	●
Rushton et al. 2008b		●			●
Stevenson et al. 2000		●			●
Sweeney 2002	●	●	●	●	●
Vine et al. 1998		●	●		●
Waller 2008					●
Zimmerman et al. 2008	●	●	●	●	●

Table 53 - Previous geocoding studies classified by topics of organizational cost

	Organization		
	Cost		
	Obtaining Reference Data	Per-Geocode	Manpower
Beal 2003		●	●
Boscoe et al. 2002		●	●
Boscoe et al. 2004	●		
Davis Jr. 1993	●	●	●
Johnson 1998a	●	●	
Johnson 1998b	●	●	
Krieger 1992		●	
Krieger 2003		●	●
Krieger et al. 2001		●	
Martin and Higgs 1996	●		
McElroy et al. 2003	●	●	●
Miner et al. 2005	●		
Strickland et al. 2007		●	●
Temple et al. 2005	●		
Thrall 2006	●	●	●
Whitsel et al. 2004		●	
Whitsel et al. 2006		●	●

North American Association of Central Cancer Registries, Inc.
2121 W. White Oaks Drive, Suite B
Springfield, IL 62704
217.698.0800
217.698.0188 fax
info@naaccr.org
www.naaccr.org