A FIRE INSURANCE MAP GEOCODER FOR PRE-EARTHQUAKE SAN FRANCISCO

by

Yonatan Rosen

A Thesis Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

May 2015

ACKNOWLEDGEMENTS

I am grateful for all the guidance and advice given to me by my professors at the USC Spatial Sciences Institute, including Dr. Darren Ruddell, Dr. Jennifer Swift, Dr. Yao-Yi Chiang, Dr. Robert Vos, and my advisors, Dr. Katsuhiko Oda and Dr. Karen Kemp.

GIS Librarians Susan Powell of the Earth Sciences and Map Library at U.C. Berkeley and Andrzej Rutkowski of USC provided me with valuable feedback, and directed me towards many Sanborn-related resources.

Images of Sanborn maps were downloaded from the David Rumsey Historical Map Collection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In the years following the 1906 earthquake and fires, the streets of San Francisco were renamed, renumbered, and reshaped. These changes make it challenging to locate addresses found in historical directories, newspapers, and archives. Fire insurance maps produced by the Sanborn Map Company represent some of the most detailed sources of spatial information about early twentieth century San Francisco, but they are cumbersome to navigate.

Insurance maps contain detailed street indexes that mirror address geocoders in content and function—listing street names and address ranges. Exploiting their structure, the text of these street indexes was transcribed in order to create a geocoder that identifies map sheets. The Sanborn indexes served as reference data for an ArcMap address locator. The geocoder makes the insurance maps more navigable and provides historical context for addresses.

CHAPTER ONE: INTRODUCTION

At the turn of the twentieth century, San Francisco was a rapidly growing metropolis. The 1906 earthquake and fires destroyed the city’s urban core, erasing the city’s fabric and much of its architectural character. It is challenging to find context for historical address records like directories and business ephemera. Fire insurance maps, like those produced by the Sanborn Map Company, depict the built environment with considerable detail, helping to provide this context. Insurance maps are a means to understand the spatial characteristics of the pre-earthquake city, but they are cumbersome to navigate. GISystems serve as a spatial framework to organize the maps, making it possible to overlay the maps with other data sources, and facilitating the process of navigation.

1.1 Motivation: Finding Context

Late nineteenth and early twentieth century texts like newspapers, diaries, and directories are awash in spatial information. Before the widespread adoption of the telephone, the postal address served both as a means of contact and as locational information. The presence of addresses in these disparate sources makes it possible to conduct research on individual buildings, blocks and neighborhoods, by employing the address as a spatial attribute. An address can be used to link an anonymous classified advertisement to other records, such as directory listings, in order to create a richer story—was the address a boarding house or a the home of a prominent family? In short, postal address can serve to contextualize historical sources. Fire insurance maps produced by the Sanborn Map Company can expand the context beyond the individual address, helping users explore the broader urban environment, including the surrounding blocks and neighborhoods. In a city like San Francisco, where much of the historical context is missing, insurance maps can help to bring that environment into focus.

1.2 Problem Statement

Street addresses provide a tangible and discrete geographical reference point. However, in San Francisco, renaming, renumbering and physical changes to streets made historical addresses ambiguous. In the first decade of the twentieth century, whole blocks were renumbered, streets renamed, and roads closed or rebuilt. Identifying the precise location of a historical address can be a laborious process, requiring research and careful reading of street descriptions in street directories. Reference data from the appropriate time period can help to resolve ambiguities. Street maps, directories and indexes vary in their reliability. Insurance maps contain the greatest level of detail. They show individual structures, documenting building types, materials and uses on the basis of careful ground surveys. However, working with the multitude of individual large scale Sanborn maps is awkward and time consuming. A more efficient way to navigate these maps is needed.

1.3 Research Questions

Street addresses provide fine-grain geographical information, but, as stated above, historical street addresses in San Francisco are sometimes ambiguous. By developing an address locator based on fire insurance map indexes it is possible to verify that a geocode is correct and also provide the geocode with architectural context. This research endeavored to exploit the structure of the insurance map indexes to develop a geocoder that relies on this historical reference data. This task prompted three fundamental questions. First, can insurance map directory data meet the technical requirements of an address geocoder? Second, would such a geocoder represent an improvement over other methods of developing a historical address geocoder? Finally, what are the costs and benefits of using this approach?

1.4 Objective

Sanborn insurance maps provide a means of visually substantiating the presence of individual buildings and their listed addresses. Taking advantage of the structure of the insurance map indexes, an address locator can be developed to reference individual map sheets. The objective of this thesis was to develop a historical address geocoder that relies on this historical reference data as a means of contextualizing historical addresses. Thus, the intent was to create a geocoder that takes a street address as input and produces as output a specific Sanborn map sheet number, represented in ArcMap as a map sheet footprint, so that the location of a historic address can be examined in its contemporary context.

1.5 Methodology

Developing an insurance map-based geocoder requires four principal steps. First, the text from street indexes found in insurance maps must be recognized, corrected, and brought into a tabular form suitable for use within a database. Second, data from index maps must be vectorized. Third, the vector representation of map sheets must be linked to the text descriptions found in the street indexes. Finally, this data must be developed into a geocoder using a GISystem.

1.6 Outline of this Document

The following four chapters explain the design and implementation of the insurance map geocoder, and explore the implications of the tool. Chapter Two provides historical background on fire insurance maps and how historians have made use of them as a resource, as well as theoretical background on postal address-based geocoding. Chapter Three explains the technical process of converting the text and image based insurance maps for use as reference information in a historic geocoder, and illustrates the way that the tool functions. In Chapter Four, directory listings of bakeries are mapped using the insurance map-based geocoder, demonstrating how the tool reveals patterns in textual sources. The concluding chapter reflects on the success of the project and explores the implications of using historic resources in this manner.

CHAPTER TWO: BACKGROUND

Historical geographical information exist on a spectrum between structured, explicitly geographical documents and unstructured information, such as text references to place names, trade routes and visual depictions of place and landscape. Historical Geographical Information Science (HGIS) can bring together such structured and unstructured sources to make sense of patterns and make historical information easier to access and analyze. Structured historical resources include maps, gazetteers and directories, which organize and store geographical knowledge. Among these, fire insurance maps constitute some of the most detailed sources of geographical information about nineteenth century American cities. For urban historians, preservationists and genealogists, they provide a snapshot of cities at various stages in their development and contain valuable information to bolster research on historic buildings and land uses. The structure of the insurance map indexes parallels the data requirements of an address geocoder.

This chapter discusses how historians and other humanities researchers have employed GIScience. Then, it outlines the history of insurance maps and their use by historians and other researchers as public and scholarly access to Sanborn maps has grown. Finally, some of the technical underpinnings of geocoding, particularly as it is implemented in ArcGIS, are explained to establish the requirements for the development of an address locator using data from Sanborn indexes.

2.1 History and GIScience

GIScience grew out of the natural sciences, which view the world through an empirical lens (Gregory and Ell 2007). Historians and other researchers in the humanities rarely have the luxury of large datasets. Instead, they make sense a fragmentary evidence—material artifacts, textual sources—to construct history. GISystems allow researchers to employ location as a framework to organize disparate sources. GISystems have been deployed to organize urban archives in Atlanta (Page et al. 2013), and create a national resource for genealogists in Scotland (MacDonald and Osborne 2013). These projects underscore the utility of GISystems in helping researchers to identify locally-relevant information.

Somewhat paradoxically, GISystems are commonly employed to identify sweeping spatial patterns, but they are equally adept at finding location-specific information. Geocoding historical addresses is only useful insofar as there are other pieces of information to be gleaned from plotting the location on a map. Overlay is the process by which two or more data sources can be integrated within a GISystem. The task of comparing two maps is extremely challenging to conduct manually. It requires an understanding of each source's projection and scale. However, overlay is one of the most basic functions in a GISystem, makes a consuming manual task trivial (Gregory and Ell 2007).

2.1.1 Mapping Texts

Gregory and Hardie (2011) used corpus linguistics techniques to illustrate attitudes in the sixteenth and seventeenth century British press towards cities in Britain and continental Europe by mapping place names mentioned in texts. Similarly, the Mapping Texts project developed a GIS interface to understand regional and temporal variation in Texas newspapers, employing topic modeling techniques to determine the subjects of articles (Torget et al., 2011). Use of place names a geographical information can illustrate regional patterns. Within a city, however, street addresses can be employed to find fine-grain patterns in the distribution of heterogeneous phenomena described in newspaper articles, like crime, employment and industry.

2.2 Fire Insurance Maps

Insurance maps represent nineteenth century cities vividly, with an engaging level detail that appears almost comprehensive. However, insurance maps were created to suit insurance underwriters, not modern historians or the broader public. In fact, prior to the early 1960s, the maps were rarely accessible outside of the insurance industry (Lamb 1961; Keller 1993). Despite their prosaic purposes, map surveyors recorded details that went beyond infrastructure and risk assessment. They depict saloons, stores and hotels—recording buildings as they were used, whether this use was “genteel or disreputable,” as an early Sanborn surveyor wrote (Ristow 1968, 202). Few cartographic resources approach fire insurance maps in their ability to help modern users visualize the architectural and commercial character of nineteenth century streets.

2.2.1 Sanborn Maps of San Francisco

The Sanborn maps are not the only available large scale maps available of San Francisco before the 1906 earthquake. Plat maps are a legal record of property boundaries (Patton et al. 2005). Historical plat maps are available through the website of the San Francisco Department of Public Works (2014). However, Plat maps provide little additional information beyond block and lot numbers, dates of registrations and dimensions of lots. Real estate atlases, such as the block books published by Hicks-Judd in San Francisco, annotate lots with names of owners (1901). Plat maps and block books provide no indication of the existence of buildings on the site, or even if the street was had been constructed. In fact, many recorded lots were sited on sand dunes, steep hillsides and or under shallow bay water, land that was only developed decades later. In theory, plat maps are temporally and spatially is continuous, because all changes to properties were required to be registered to be made official. By contrast, Sanborn maps are discontinuous; they represent relevant sections of cities at discrete periods (Patton et al. 2005).

If plat maps represent property in the abstract, fire insurance maps represent the material environment, as observed by surveyors on the ground. Sanborn maps team with information about building use and construction, including building footprints, heights, materials and names of businesses. Street addresses of each building were carefully recorded—a detail lacking in both plat maps and block books.

Sanborn Fire Insurance maps of San Francisco dated between 1875 and 1991 are available at various libraries throughout the United States. The Sanborn Company produced lithographed map volumes in four major editions, 1875, 1893, 1899 and 1915. These editions were periodically updated with pasted inserts, making each surviving copy a unique source. Scans of the 1899 microfilms are available online through a ProQuest database and the San Francisco Genealogy Website (2014). The black and white microfilm scans deprive the maps of the color-coded symbology. Color photographs of a 1905 update of the 1899 maps are available through the David Rumsey collection. The color images create a clearer sense of documents' materiality. Structures are easier to distinguish and distortions are minimized. The 1899 Sanborn insurance maps of San Francisco consist of six volumes of lithographed map sheets (Rumsey 2011; Hoehn 2014). Each sheet measures twenty-five by twenty-one inches, recorded at 1":50' scale, the typical scale employed for dense urban centers (Keller 1993).

2.2.2 Historical Development of Insurance Maps

The development of fire insurance maps reflected needs and growth of the fire insurance industry. Fire insurance maps originated in seventeenth century London as a means for insurers to assess the vulnerability of their customers' properties to hazardous conditions (Keller 1993). In the early nineteenth century, fire insurance was largely conducted by local companies. Local outfits relied less on maps, because insurers were able to inspect sites themselves. By the midnineteenth century, as urban centers grew and new cities developed, insurance maps helped to provide insurers with information about properties in places they may have never been (Ristow 1968). The 1850 Perris-Hope map of New York City, a multi-volume large scale atlas of New York City, marks the beginning of the industry, and established many of the conventions of the genre. The industry grew steadily, although map production was slowed by the Civil War. In 1867, Aetna insurance company hired William Sanborn to survey cities in Tennessee. The next year, Sanborn founded his own operation, which quickly dominated the insurance mapping industry (Ibid).

Between 1868 and the early 1950s, the Sanborn Company produced maps for over 12,000 cities and towns in the United States (Ristow 1968). In small towns, the maps concentrated on industrial and commercial areas where there was greater risk of fire (and potential customers for insurance companies). In large cities like San Francisco, residential districts were mapped extensively, although neighborhoods in the periphery of the city may have been omitted.

In part because of the dominance of the Sanborn Company, the maps exhibit a remarkable uniformity in scale and representation, making it fairly straightforward to read and interpret them. Most maps were drawn at 1 inch to 50 feet. Building materials are indicated with color—yellow for wood, red for stone or brick, grey for metal, and green for special hazards (Keister 1993). Building uses were either written out or somewhat inconsistently abbreviated. Single-family residences were indicated with a “D” for “Dwelling”. Multiple families living on separate floors were indicated with an “F” for “Flats”. Buildings with multiple families living on the same floor were classified as “A” for “Apartments” (Lamb 1961; Grim and Narrow 1990). Hotels, Lodging houses, boarding houses and other housing functions like dormitories were typically written out.

Attention to commercial functions depended largely on their relevance to fire insurance underwriting. Uses that posed a hazard were given more attention, while other uses are ignored. Stores are generally indicated with an “S”. Insurance risks informed the level of detail provided about industrial uses. Uses that had implications for fire such as paint shops, blacksmiths, bakeries and received more attention than less risky functions like book binding. This focus on hazards makes the maps less reliable or comprehensive for functions outside of this purview.

Insurers subscribed to map correction services through the Sanborn Company. Maps were updated with correction stickers pasted over the original to reflect updates and new construction. Many surviving maps contain several layers of correction stickers. Researchers seeking to uncover previous layers sometimes removed the correction stickers (Lamb 1961).

2.2.3 Scholarly Access to Sanborn Fire Insurance Maps

Sanborns were a highly specialized proprietary form of geographical information, akin to high resolution satellite imagery today. Insurers paid dearly for the maps, and the heavy tomes were costly to store and maintain (Ristow 1968). A single volume of the San Francisco map cost seventy five dollars in 1893, meaning that the six volumes set was worth nearly $12,000 in 2013 dollars (Sanborn Map Company 1893). In 1925, maps of Chicago cost $1,500 and $500 a year to maintain updates (Keller 1993). Insurers also paid the Sanborn Company a subscription fee to keep the maps updated with pasted in updates as new buildings were built. Their cost and rarity limited scholarly access to the Sanborn Maps (Lamb 1961). Ninety five percent of Sanborn clients were in the insurance industry (Wrigley 1949).

By the 1930s, changes to insurance underwriting practices made maps less essential to insurers. To lower costs for their customer base, the Sanborn Company released new maps at reduced scales (Keister 1993). As Sanborn maps became outdated, they were sometimes donated to libraries and universities (Lamb 1961). However, these piecemeal collections of fire insurance maps remained fairly cumbersome to access and did not lend themselves to large-scale analysis.

Beginning in the 1860s, the Sanborn Company deposited new maps at the Library of Congress for copyright purposes, making their collection the most comprehensive. The Census Bureau accumulated a collection of 1,804 map volumes in the 1940s. In 1967, the Census Bureau transferred its collection to the Library of Congress. Following this acquisition, the Library of Congress began to distribute duplicate maps to research libraries (Ristow 1968). In 1983, Chadwyck-Healey, Inc. released microfilms copies of many of the Sanborn maps held in the Library of Congress collection. While microfilming led to wider access to the maps, limitations of the microfilming medium made them difficult to interpret. Maps were filmed at a reduced scale, and sheets were only partially visible on the screen. Microfilming also introduced distortions at edges, and the lacked the colors indicating building materials (Keller 1993). Users had to rely on index maps to assess the orientation of fragmentary map sheets. In 2001, ProQuest digitized the Chadwyck-Healey microfilms, providing online access to maps by subscription or for purchase by libraries (Lutkenhaus 2002). The digital database provided modest improvement over microfilm machine as thumbnails could be viewed simultaneously. However, the distortions and other flaws of the original microfilms remained.

2.3 Exploring Fire Insurance Maps with GISystems

GISystems have been used widely to integrate information from fire insurance maps with contemporary data for research and public use. GIScience has been applied both as means to facilitate access to maps and as a source of data. As early as 1992, the Illinois State Museum developed the Historical Hazards Geographic Information System, which took advantage of the Sanborn maps to help identify industrial hazards throughout Illinois. (Keller 1993; Colten 1992).

A number of research libraries have employed GISystems to make it easier to navigate their historic Sanborn collections. In 2010, Yale University Library digitized its large collection of insurance maps of cities throughout Connecticut. To make the maps easier to access, Yale digitized map footprints on the basis of index maps. The indexes are available as KMZ files on the library website. Additionally, they digitized building footprints from the maps of Yale's campus, using building heights indicated on the map to create a three dimensional visualization of the campus (Yale University Library 2013).

The map division of the New York Public Library has developed two excellent public participation GISystems that made insurance maps of New York City easier for the public to access. In 2010, they released the Map Warper, a map image georectification tool for the thousands of maps in their collection, including fire insurance maps of New York City dating to the 1850s.The Map Warper allows users to identify control points on map images based on landmarks. Users can also search for maps by location (Vershbow 2013; New York Public Library 2014). Georectified raster images of historical maps can be used for visual overlay within a contemporary GISystem, but the images must be vectorized for the data to be extracted. To facilitate this process, NYPL Labs, a group at the New York Public Library that creates web apps to help the public interact with its collections, developed Building Inspector. Building Inspector applies image recognition software to vectorize building footprints. Users then help to correct vectorization errors, and assign attribute data that cannot be automatically extracted, including color and building address. The data produced by Building Inspector can then be exported through a public API.

Insurance maps are commonly employed a source data for studies related to the built environment, particularly when Raymond (2011) employed Sanborn maps to reconstruct a neighborhood in Seattle that was destroyed by urban redevelopment. Leonard and Spellane (2013) made use of fire insurance maps dating to 1850 to identify potential historical sources of contamination to Newtown Creek, which straddles Queens and Brooklyn.

2.3.1 Web Access to Insurance Maps of San Francisco

In 2011, the map collector David Rumsey worked with the San Francisco Public Library to digitize its 1905 revision of Sanborn maps for San Francisco. The maps were partially damaged in the fires of 1906. The maps represent a vast improvement on previous microfilm scans. The high-resolution images are clear, and contain marginalia that dramatize the use of the documents, including pencil markings and notes. In response to the release of the color insurance maps, Michal Migurski developed a web interface to allow members of the public to georectify map scans (Sommer 2011). The resulting website, maptcha.org, has a clickable webmap, allowing access to large scale map sheets. Map sheets are represented by a thumbnail graphic on the map and georectified map sheets are not visible on the site (Migurski 2011). Separate from the maptcha.org site, the David Rumsey site also provides an interface for georectification of map images through georeferencer.org.

Rectification of the insurance map images has great potential, as demonstrated by the New York Public Library project. However, the maptcha.org project only modestly improves the navigability of the insurance maps. In part, the problem stems from a quirk in the San Francisco map sheets: street widths are not drawn to scale. This means that each block would need to be independently georectified to minimize distortions. With at least four blocks depicted on each map, this would require a more intensive process to create accurately rectified maps.

2.4 Geocoding

Goldberg, Wilson and Knoblock (2007) offer the most comprehensive review article covering the state of the art in geocoding research. They define geocoding, in its most basic form, as the process of transforming textual geographical references into a spatial representation within a GISystem. These textual references can include place names, relative location description, or areal units like postal codes, but the most common form of input data is the postal address (Goldberg, Wilson and Knoblock 2007). Most address systems in the United States employ house number, street name, as well as city and state to identify location.

Geocoders must accomplish three distinct tasks to correctly identify the location of an address. First, they must parse the address, distinguishing, for example, the house number from the street name, street type and directional suffixes. Second, they must use these elements to identify a geographic feature with matching attributes. Third, they must create a geocode—the spatial representation of the address (Goldberg, Wilson and Knoblock 2007). The latter two steps introduce much of the potential for error.

2.4.1 Geocoder Data Models for Street Addresses

Geocoders can be used to identify states, geographical zones like zip codes, or even geographical features. Each type of geocoder has distinct data models and outputs. There are three data models that are used for street address geocoders. The first, most commonly employed model uses street centerlines—line data that represents the street system of a city. An address range is associated with each line segment—corresponding to a block range, with an odd or even range assigned to either side of each segment. Once the correct segment is identified, linear interpolation is employed to identify the location of the address along the segment. By this method, on a street segment with the range 800 to 899 (i.e. the 800 block), the address 875 would appear at a position three quarters of the way along the segment. In this method, all of the addresses within a range could potentially geocode correctly—all addresses from 800 to 899 are assumed to have the same dimensions. In an urban environment, the potential for error is limited by the relatively short street segments. The magnitude of error is equal to half of the length of a segment (Goldberg et al, 2007)

Two alternatives to the centerline-based data model have been developed to introduce greater precision into the geocoding process. Parcel geocoders employ the polygon geometry of parcel data, with a discrete address assigned to each lot. By this method, an address must match correctly with both the lot number and street name to be located. Similarly, address point geocoders employ point geometry to offer more precision on large lots, or when multiple addresses are associated with a single lot (Goldberg et al. 2007; Zandbergen 2008). While a centerline geocoder is more flexible and forgiving; address point and parcel geocoders are more precise. A centerline geocoder will return more false positives—geocoding for addresses that may not exist in reality, while address point or parcel geocoders are more likely to return false negatives (Zandbergen 2008).

2.5 ArcGIS Address Locator Styles

Customized geocoders in ArcGIS are termed “Address Locators”. ArcMap version 10.2 offers users twelve basic address locator styles, reflecting distinct reference data requirements and specific outputs. In addition to street address locators, these styles include city/state, zip code and place name locators. An address locator style designed to identify the location of a city and state could not be used to locate house numbers. Among these twelve choices, only three styles can be used to geocode street addresses: “US Address—Dual Ranges”, “US Address—One Range”, and “US Address—Single House” (Esri 2015). The Single House locator style is used with reference data that links addresses to discrete objects, represented as points or polygons. This style is suitable for address point or parcel geocoders described by Goldberg et al. (2007). The Dual Ranges and One Range locator styles are distinguished by how they account for polarity of addresses. The Dual Ranges style requires separate ranges for each side of a street segment, meaning that one line segment has two pairs of range attributes for the left and right sides of the streets. By contrast, the One Range locator style requires that the each street segment be designated either Left or Right. In this way, the address locator is able to distinguish odd-numbered ranges from even-numbered ranges (Esri 2015).

2.6 Assessing Geocoding Error

As shown in Section 2.4, mitigating and quantifying error is a central concern in the geocoding literature. Incorrect reference data contribute to ambiguous results. In this vein, an important distinction exists between precision and accuracy. Precision is a technical measure of an instrument. Just as a precision watch can be trusted to tell time to the closest millisecond, precision of a geocoder relates to specificity of the measurement. Can a geocoder be trusted to identify a feature to the nearest inch or the nearest yard? A linear interpolation geocoder may return the correct street segment for an address, but limitations in the data can hobble its reliability for finding the correct position on that segment. By contrast, accuracy is the degree to which points correspond to the real world position being represented (Bolstad et al. 1990).

Precision is an important goal but the emphasis on precision is misplaced for historical addresses, because reliable reference data are difficult to acquire. It is more important to have a method of verifying geocode than having fine-level geocode.

2.7 Developing a Historic Geocoder

The development of historical address locators has received limited scholarly attention, even among researchers in HGIS who have employed addresses in their geocodes. The requirements of address locators are context specific. Debats and Lethbridge (2007) created a geocoder to accommodate sequentially recoded tax records from Alexandria, Virginia, from a time when there were no house numbers. In some instances, historic address ranges and street names have not changed enough to warrant the development of new address locators. Contemporary street centerline data can be used, employing alias tables to reference modifications to street names. However, physical changes to the streets stemming from redevelopment, landfill and other infrastructure modification make editing of the centerline data necessary. Editing street centerline data is a laborious process, and potentially introduces error.

2.8 Geocoding with Insurance Maps

The significance of Sanborn insurance maps as a historical geographic resource is evidenced by the extensive scholarly engagement detailed in this chapter. Insurance maps began as a rarefied resource, but as they became outdated they became more accessible to scholars. Efforts to reproduce and digitize them have increased availability and public and scholarly interest. GIScience has been deployed to extract information from insurance maps, and to help make map collections more navigable. Address geocoding allows users to transform text into spatial representation in a GISystem. The following chapter demonstrates how a geocoder can be developed in order to improve the usability of insurance maps within a GISystem to provide historical context.

CHAPTER THREE: DESIGN AND IMPLEMENTATION OF INSURANCE MAP GEOCODER

Fire insurance maps present spatial information to users with varying degrees of on-the-ground knowledge. Indexes of streets and key maps help users navigate the volumes. These indexes create a structure that allow users to find spatially relevant information. Making use of this structure facilitates the process of data development and capture for use within a GISystem. This project required identifying the elements of the indexes that could be harvested for use in a GISystem.

As stated above, the objective of this process is to create a geocoder that takes as input a street address and produces as output a specific Sanborn map sheet number, represented in ArcMap as a map sheet footprint so that a historic address can be examined in its contemporary context. As such, the address ranges found in the Sanborn indexes had to be adapted to meet the requirements for reference data of US Address – One Range style locator, discussed in Section 2.5. Section 3.1 describes the navigational elements of the Sanborn map sheet and street indexes, and outlines how these elements have been adapted to suit the requirements of an address locator. The section ends with a flow chart that summarizes the process of development of the address locator reference data that is described in greater detail in the following three sections. Section 3.2 explains how the Sanborn street indexes were digitized and restructured to serve as reference data for the locator. Section 3.3 outlines the process used to georeference the index maps and create map sheet footprints. Section 3.4 explains how a dummy grid was created to link the address ranges to the geometry of the map sheet footprints. In Section 3.5, the creation of the address locator in ArcMap is described. Finally, Section 3.6 demonstrates the functionality of the address locator.

3.1 Conceptual Model of the Sanborn Insurance Map Indexes

In order to adapt the index structure of insurance maps to GISystems tools, it is worthwhile to delineate their basic elements. Each of the six volumes of the San Francisco Sanborn maps contains map sheets for a roughly contiguous region of the city. These regions have no explicit social or political significance, although their boundaries tend to be defined by major streets. Each volume contained an index or key map, an example is shown in Figure 3.1, which provides a visual means of identifying the location of a map sheet in relation to other sheets. Each volume contains over one hundred sheets.

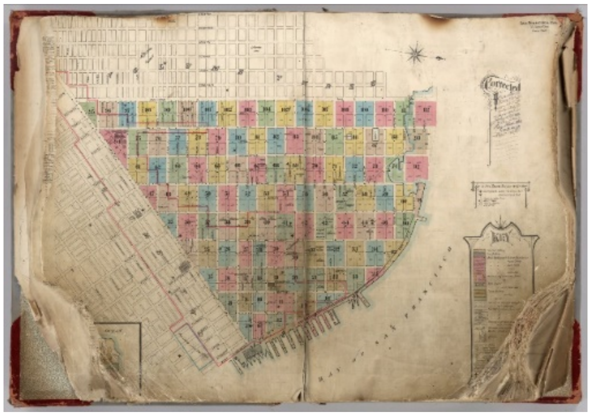


Figure 3.1 Sanborn Index Map for Volume 1[1]

Two index pages consisting of a street index, specials index, block index, and miscellaneous report is found at the front of each volume. The street index lists the streets found in the volume, their address ranges and the corresponding map sheet. A table titled, "List of Streets on Old Maps Appearing under New Official Names" details the name changes for streets between the 1893 edition and the 1899 maps. The specials index identifies sheets for major landmarks, buildings and significant sites. Figure 3.2 illustrates the various navigational tools

[1] This figure from the David Rumsey Map Collection is reused under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. Attributions for subsequent figures can be found on page 71.

included in each map volume used to identify each map sheet. The structure present in the insurance maps requires a user to take a linear set of steps to find a map of a specific address. First, she must identify the volume where the address was found. She can then look to the index map to find a map sheet visually, or consult the index of streets. The index of street simply lists the street names, the address ranges and their corresponding map sheet. If a street name is missing from the index of streets, she can look to the list of old street names to determine if the street name had changed. However, a missing street may simply be found in a different volume. The separation among the volumes complicates the process of identifying locations, because each volume has its own index page and index map.

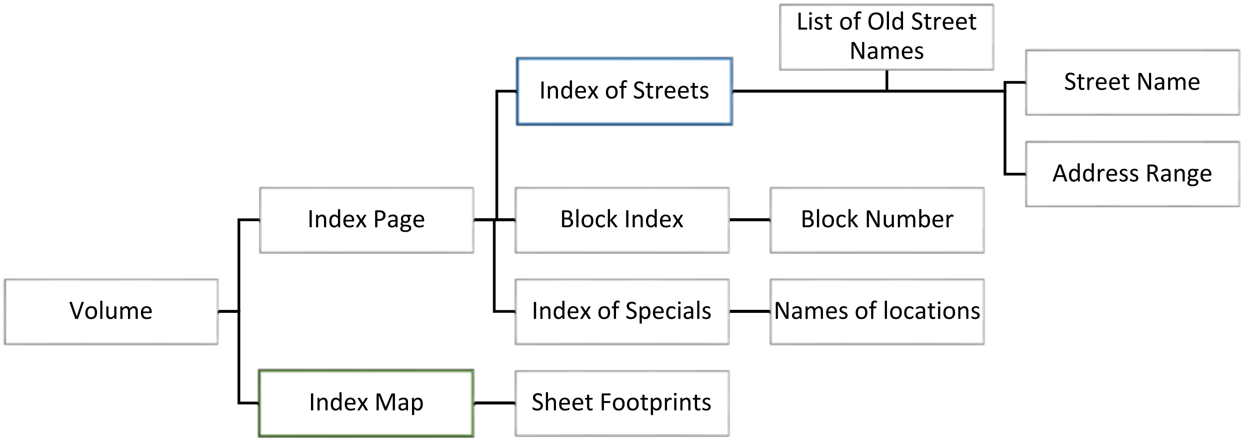


Figure 3.2 Conceptual model of the Insurance maps

3.1.1 Transforming the index structure into an address locator

The San Francisco fire insurance maps comprise 688 sheets. Each sheet depicts an area of four city blocks, or roughly twelve acres. The regularity in size of maps sheets makes it possible to think of them as a small areal unit. The footprints can be represented in a GISystem as polygons which correspond to the location of the map sheet numbers referenced in the street index.

The navigational elements found in each volume of the insurance map create a robust means of identifying locations depicted in the insurance maps by hand, but they lack the consistency and data integrity for computational interpretation. The elements of the index of streets have clear analogues in the data model of a street address geocoder. By digitizing and manipulating these elements, they can be transformed into an address locator.

The street indexes share three of the required elements of an ArcGIS street Address Locator: A street name, followed by an address range, followed by a sheet number. Several steps are required to fulfil data model requirements of the “U.S. Address—One Range” Address Locator style. First, the street name field must be divided into three attributes StreetName, StreetType, and Directional Suffix. Second, the numerical address ranges must be separated into ‘From’ and ‘To’ values. Third, the polarity—the side of the street—must be assigned. Additionally, the list of old street names can be used to create an alternate name table, providing aliases for names that have changed. Finally, the sheet number can be used to link the nominal attributes of the locator to their spatial representations in the GISystem.

The “U.S. Address—One Range” style requires lines as reference data, because it uses linear interpolation to estimate the position of an address along a line segment. The data derived from the street indexes refer to sheets, which would be better represented by polygons. However, in order to develop reference data that would function within the requirements the street address locator style, a pseudo-grid consisting of multiple line segments falling within each map sheet footprint was created (the process for developing this grid is described in Section 3.4). The One Range locator style is preferable here to Dual Range style because it makes it possible to distinguish odd-numbered segments from even numbered segments, which often occur on separate map sheets. This workaround made it possible to employ ArcMap’s geocoding interface. The address range transcription and index map digitization processes are both labor-intensive. While they exploit a robust structure, the incongruities of the source documents and imprecision of digitization tools mean that neither process can be entirely automated.

Figure 3.3, below, depicts the how reference data for the address locator was developed by combining the text address ranges from the street indexes (highlighted in blue) with the map sheet footprints from the index maps (highlighted in green). The street indexes were transcribed automatically to create a table of street names and address ranges associated with each map sheet. The index maps were brought into ArcMap and used to assign sheet numbers to modern parcel data. Map sheet footprints were created, and a false grid of lines was created to represent street segments. Finally, the address ranges and dummy line segments were each assigned a unique identification number so that they could be joined in ArcMap, creating reference data for the address locator. The steps of this process are detailed in Sections 3.2 through 3.4. Sections 3.5 and 3.6 detail the development and implementation of the address locator in ArcMap.

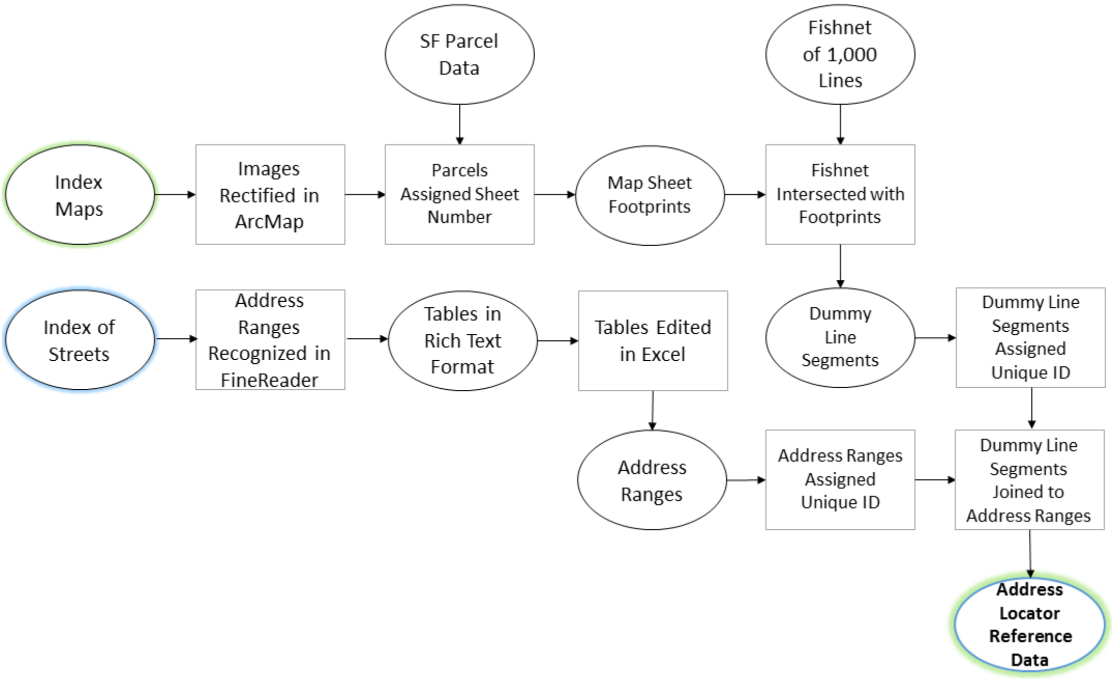
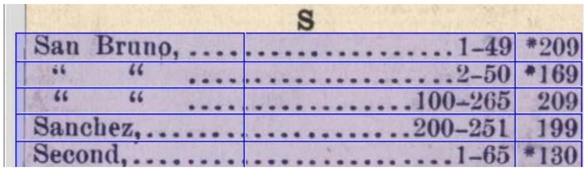

SF Parcel Data
Fishnet of 1,000 Lines
Index Maps
Images Rectified in ArcMap
Parcels Assigned Sheet Number
Map Sheet Footprints
Fishnet Intersected with Footprints
Index of Streets
Address Ranges Recognized in FineReader
Tables in Rich Text Format
Tables Edited in Excel
Dummy Line Segments
Dummy Line Segments Assigned Unique ID
Address Ranges
Address Ranges Assigned Unique ID
Dummy Line Segments Joined to Address Ranges
Address Locator Reference Data

Figure 3.3 Process of development of Address Locator reference data

3.2 Capturing Street Segment Address Ranges

At the front of each volume, an index page lists street names and address ranges and the sheet where that address can be found. The index pages have a roughly tabular format that can be exploited for use within a geocoder. The tables are not entirely dissimilar to the attributes found in a modern geocoder. They contain a street name, street type, address range and a spatial attribute in the form of the sheet number, shown in Figure 3.4.



S		
San Bruno, ...	1-49	*209
" "	2-50	*169
" "	100-265	209
Sanchez,.......	200-251	199
Second,........	1-65	*130

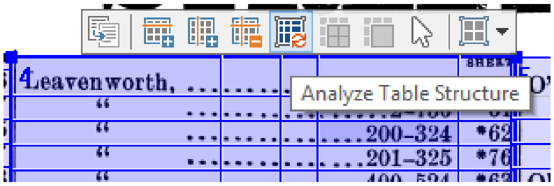
Figure 3.4 The table tool in ABBYY FineReader

ABBYY FineReader is a document management and Optical Character Recognition (OCR) software tool. Like Adobe Acrobat Pro, it can be used to convert images of text into machine readable text, but it is also able to identify page structure, and distinguish page element including text, tables, and images. Scans of the index pages were downloaded from the David Rumsey website at 490 dots per inch (dpi). The images were loaded into FineReader, and automatically preprocessed according to the default settings, which reduced the resolution to 350 dpi. Figure 3.5 shows a sample of the scale and resolution of the recognized text.

Julian, ...	1-198	203
Juniper, ...	1-156	172

Figure 3.5 Sample scale and resolution of recognized text in FineReader

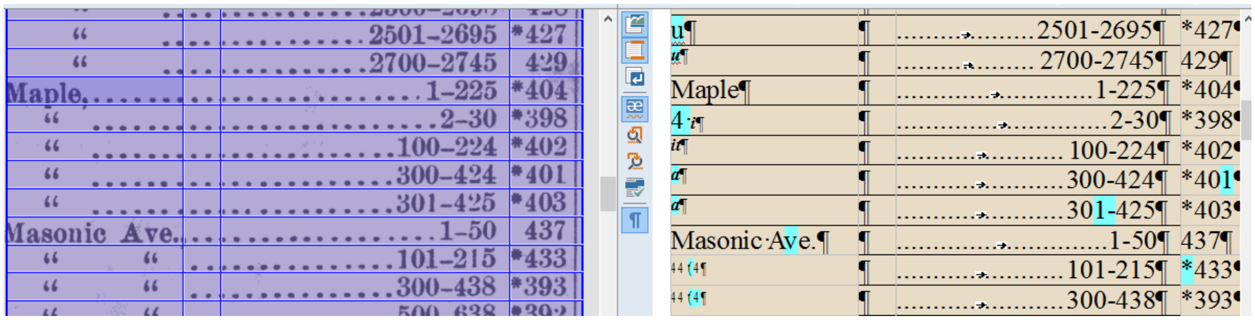
The Analyze Page tool can identify page elements automatically, but it was more reliable to draw a table over each column on the index page using the Draw Table Area tool. Once each column was defined, the Analyze Table Structure tool was used to identify the elements of the table, as shown in Figure 3.6, on the following page. The table structure had to be touched up using the Delete separator tools.



Leavenworth,		
"	200-324	*62
"	201-325	*76

Figure 3.6 The table toolbar in FineReader

Once the table structure was properly drawn, the text could be identified using FineReader's OCR capabilities. The results of the OCR must be reviewed for errors. FineReader provides a means to compare the text image to the recognized text in two windows, as shown in Figure 3.7. The recognized text are displayed next to the text image. FineReader identifies text characters based on their similarity to known fonts. Characters are assigned a confidence score. "Low confidence" characters are highlighted in blue, which facilitates the manual correction process. Extra attention was paid to numerals, because errors in recognition of numeric characters are more difficult to identify than errors in alphabetical characters that affect the spelling of words.



"		2501-2695	*427
"		2700-2745	429
Maple,		1-225	*404
"		2-30	*398
"		100-224	*402
"		300-424	*401
"		301-425	*403
Masonic Ave.		1-50	437
"	"	101-215	*433
"	"	300-438	*393

u¶	¶	.........→.........2501-2695¶	*427
u¶	¶	.........→......... 2700-2745¶	429¶
Maple¶	¶	.............→.............1-225¶	*404
4·i¶	¶	.............→.............2-30¶	*398
i¶	¶	..........→.......... 100-224¶	*402
a¶	¶	..........→..........300-424¶	*401
a¶	¶	..........→..........301-425¶	*403
Masonic·Ave.¶	¶	.............→.............1-50¶	437¶
44 (4¶	¶	..........→..........101-215¶	*433
44 (4¶	¶	..........→..........300-438¶	*393

Figure 3.7 Screenshot of ABBYY FineReader correction process

Images in the David Rumsey Collection of index pages for volumes one through four have missing sections due to damage at the edges. These lacunae were supplemented with text from microfilmed copies available through the ProQuest database. However, due to poor image quality, the microfilmed portions were transcribed by hand.

After completing optical character recognition for the index pages of each volume, the files were exported in Rich Text Format (RTF). RTF preserves the tables identified by ABBYY FineReader. In Microsoft Word, the six separate files were combined manually into a single document. The resulting tables contained four columns: the street name, text descriptions of ranges, numerical ranges, and sheet number. In order to develop an address locator in ArcMap, the four columns needed to be further subdivided into Street Name, Street Type, Directional Suffix, From, To, and Sheet number.

Extraneous formatting was removed using the find and replace function. The hyphen character between the numerical ranges was replaced with a tab character. The tables were then converted to tab delimited values (TDV) and the file was saved as plain text. The TDV file was then imported into Excel. In Excel, the resulting file contained four columns: Street, From, To, and Sheet number. A combined total of 5,153 street segments are listed in the street indexes.

The structure of the street indexes lends itself to digitization, but the format was designed for human interpretation. Inconsistencies in formatting had to be manually corrected to meet the data requirements of an address locator. These inconsistencies varied enough to make automated correction impractical. However, simple functions in Excel make it possible to correct values that fail to meet data requirements. In the street index, repeated street names were represented with ditto marks. In Excel, the names of repeated streets were filled by dragging the fill handle. Using column filters, it is straightforward to identify incorrect values found in a column.

3.2.1 Assigning Street Type

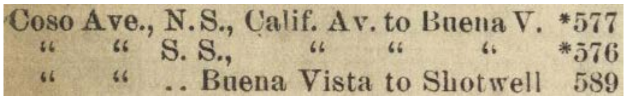
The street type value is easy to assign using Column filters. The term “Street” was omitted from most streets in the directory. By filtering the Address column by other street types (i.e. Avenue, Way, Alley, Lane, Place), it was possible to assign to correct street types to large groups of streets at once. The remainder (3,426 of 5,151) were filled with the term “Street”. A small number of streets (151) also included a directional suffix, which were identified by filtering.

3.2.2 Assigning Sides to Address Ranges

Single-range Address Locators are also required to have an attribute that denotes the side of the street where the address is found. Most ranges were arbitrarily assigned the value of L, corresponding to the left side of the street. In the street indexes, streets that were split across multiple sheets were demarcated with an asterisk next to the sheet number. Using column filters in Excel, one-sided street numbers were isolated. Columns were sorted by address name and numerical range. Odd numbered ranges were assigned the letter R for the right side.

3.2.3 Text Descriptions of Street Ranges

Not all of the street segments included in the street indexes include a numerical address range. Address ranges for smaller streets and alleys were often omitted. Some 763 street segments contain no address range or text description. Most of the names of the segments do not occur on multiple pages. These segments correspond to smaller streets. Other street segments lack a numerical address range, but contain text descriptions of streets bounding the segment. For example, Coso Avenue appears on three sheets, shown in Figure 3.8.



Coso Ave.,	N.S.,	Calif. Av.	to	Buena V.		*577
“	“	S. S.,	“	“	“	*576
“	“	.. Buena Vista	to	Shotwell		589

Figure 3.8 Text descriptions of Coso Avenue

The first segment depicts the north side of Coso Avenue between California Avenue and Buena Vista. Just 341 of the street segments contain such text descriptions. Examination of these occurrences reveals that these street segments appeared either undeveloped or lacked numbered buildings represented on the map sheets. To deal with these missing numerical ranges, it is possible to find the corresponding address values using other data sources like street directories. However, it is not possible to automate this process, because available street directories from San Francisco of this period lack necessary structure. The fact that these address ranges tend to correspond to unbuilt (or unrepresented) streets on the insurance maps suggests that addresses in these ranges would occur infrequently. As an alternative measure, the Address Locator can be set to match addresses without a house number in the Locator Preferences. Selecting this option slows the performance of the Locator slightly, because it creates a greater number of possible candidate matches. However, the tool makes it easy to check candidate map sheets to identify the correct sheet.

3.3 Index Map Georectification and Creation of Map Sheet Footprints

At the front of each volume, an index or key map appears in order to help users navigate the map sheets. The six images of the index maps were downloaded from the David Rumsey website. Using the slice tool in Adobe Photoshop, images of the index maps were divided into smaller tiles, in order to make them easier to manipulate within ArcGIS.

The tiled images were loaded into ArcMap. Employing the Georeferencing Tool, control points were assigned to each map image tile. Georeferencing is an imprecise process, requiring trial and error to adjust map images to fit the coordinate system. By georeferencing tiled areas of the images, distortions caused by discrepancies in projections could be minimized.

Next, parcel data from the City of San Francisco data portal were loaded into ArcMap. While some individual lots may have changed, the shape and dimensions of the blocks have remained consistent for the most part. This allowed the corners of blocks to be used as control points. Street centerline data were loaded to provide labels to streets. Figure 3.9, on the next page, shows the alignment of one index map tile with modern parcel data.

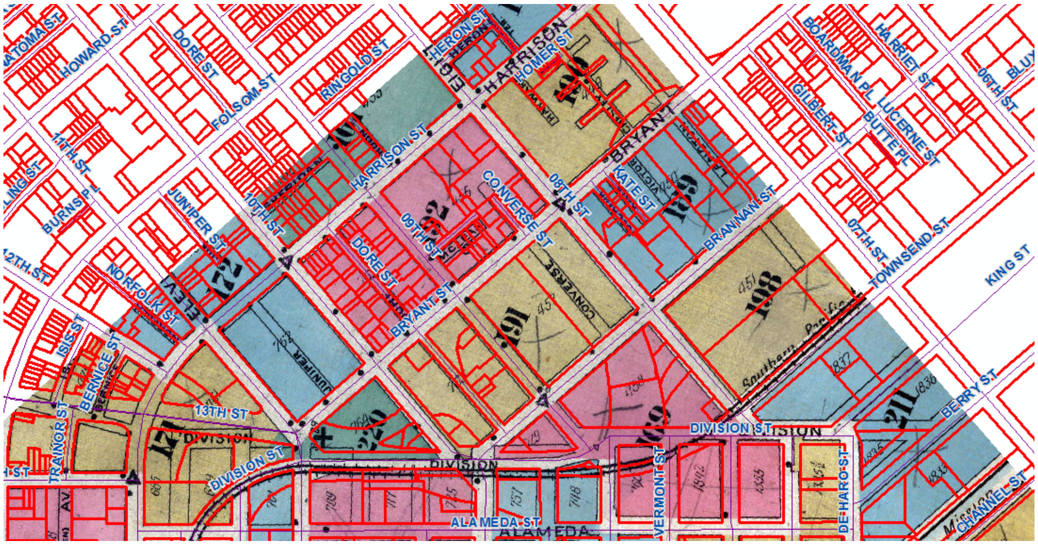


Figure 3.9 Detail of Volume II Index Map in ArcMap

The process of georectification was repeated for each of the six volumes of the map sheets. Figure 3.10 shows a composite of the index pages. There is no map coverage for large portions of the city, including the Presidio, Golden Gate Park, and large areas of the Sunset and Richmond Districts on the city’s west. These absences tend to correspond to areas that were not relevant to insurance mappers—parks, cemeteries, military bases and undeveloped land.

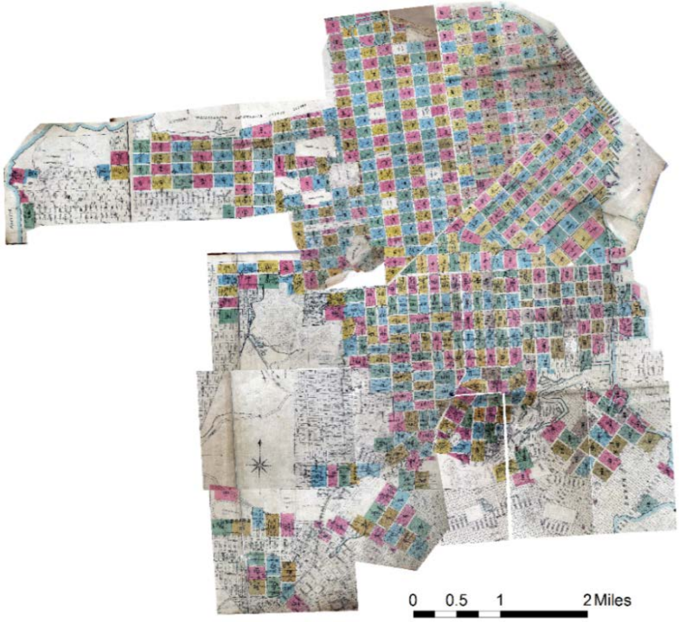


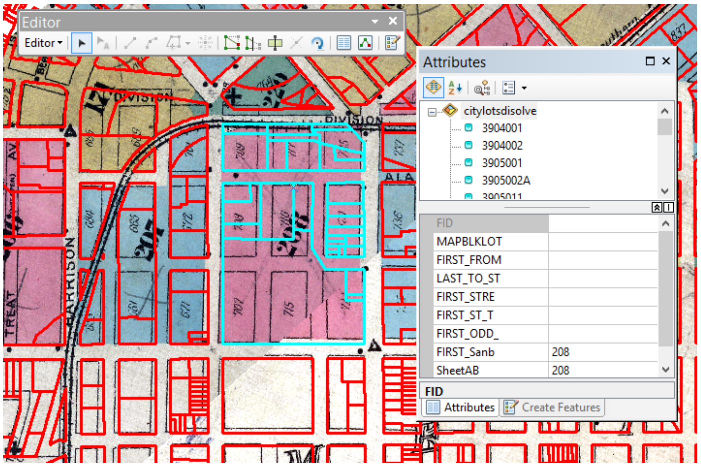
Figure 3.10 Composite image of map index pages.

On their own, the georectified index map images cannot be used for analysis. In order to be used as part of the address locator data model, the illustrated map sheet footprints must be represented with vector data to create discrete objects that can be manipulated within the database.

3.3.1 Creating Map Sheet Footprints

Once the task of georeferencing of index maps was complete, a vector representation of the map sheets needed to be created. Chiang et al. (2009) employed raster analysis techniques with historical maps to automate the process of digitization. While the task of digitization could have been accomplished partially by identifying regions through raster analysis, the scale and generalization of the index maps make them unreliable for overlay with modern data. Instead, map sheet numbers were manually assigned to the modern parcel data geometry, insuring that the referenced map footprints reflect the geography of the city.

In ArcMap, parcel data were overlaid on top of the tiled images of the index maps, as shown in Figure 3.11. All parcels falling within a single color-coded map sheet were selected. The selected parcels were then assigned the corresponding sheet number as an additional attribute. Some blocks of the city aligned cleanly with parcel data, but in other cases, careful examination of the map sheets themselves was required to determine which parcels to code to which map sheet. This was particularly true in the peripheral tracts of the city where the orthogonal structure of the city’s street grids were not maintained, such as in Bernal Heights or the Fairmount Tract.



citylotsdisolve
- 3904001
- 3904002
- 3905001
- 3905002A
- 3905011

FID	
MAPBLKLOT	
FIRST_FROM	
LAST_TO_ST	
FIRST_STRE	
FIRST_ST_T	
FIRST_ODD_	
FIRST_Sanb	208
SheetAB	208

Figure 3.11 Selecting parcel data and recoding

In limited regions of the city, the geometry of the parcels did not correspond properly to the maps sheets. For example, the Marina District, which was not developed until after 1915, did not continue the street grid of the surrounding streets. In these instances, the parcels were edited to fit the historical street grid structure. These areas were undeveloped in the 1899 and 1905 Sanborn Editions. In fact, much of the property in these parcels was still unfilled bay and marshland. Figure 3.12 depicts the map sheet footprints generated by the coded parcel data. The spatial regularity and relative continuity of the sheets is evident.

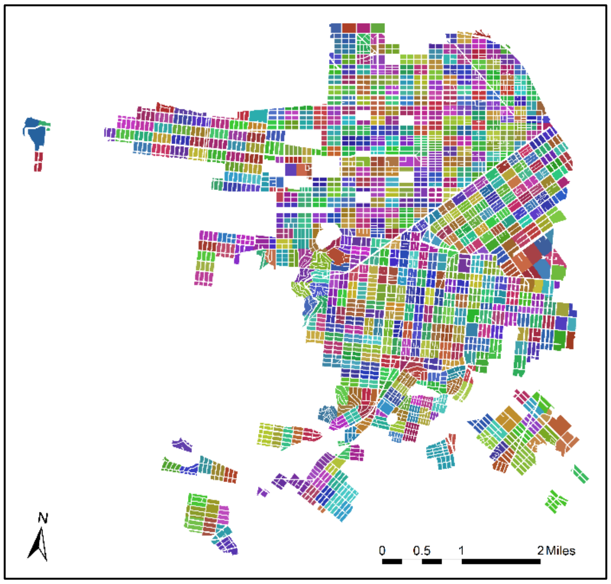


Figure 3.12 Overview of map footprints

3.3.2 Joining Sheets to Links of Map Images

While not a part of the process to develop an address geocoder, eventual use of the geocoder required the map footprint data to be related to images of the original Sanborn maps. Images of the map sheets are available on three different databases. ProQuest's Digital Sanborn Maps, 1867-1970 provides access to scans of the Chadwick-Healy microfilm of the 1899 edition of the maps in PDF format. The ProQuest database requires a subscription for access, and the database interface makes linking impractical. The website SFGenealogy.com provides public access to the same 1899 edition with scans of a distinct microfilm. The SFGenealogy scans are superior in clarity and legibility, but evidence substantial distortion at page edges. The full-color images available from on the David Rumsey website supersede both microfilm sources in quality, notwithstanding the fire-damaged sections of the pages. Contrasts between the 1899 edition and the 1905 update also provide further insights into the development of the city during that period.

Links to the images from SFGenealogy.com and the David Rumsey website were extracted by editing the relevant index HTML pages in a text editor. In both cases, the links to the map sheets contained the sheet number in the URL. Using the sheet number attribute, a table containing links was joined to the map sheet footprints. This allows the relevant map image to be opened within ArcMap by using the Identify tool on a particular map sheet footprint.

3.4 Creating a Dummy Street Grid

Theoretically, a geocoder can use an address to identify any type of object. A geocoder can return a polygon, corresponding to a zip code region, for example. However, ArcGIS requires line data for Address Locator styles that employ address ranges, like the "U.S. Address- One Range" style. In order to meet the data requirements, arbitrary line street segments were created to correspond to map sheets. The objective in creating this locator was to make geocodes that identify the correct map sheet.

After examining options for creating lines within a polygon, a more straightforward method was selected. A fishnet consisting of 1000 columns and no rows was generated, using the extent of the map sheet footprints layer. The fishnet was intersected with map sheet footprints creating 14,777 segments (many more than the 5,151 segments in the directory), each coded with the number of a corresponding map sheet. Figure 3.13 shows the grid intersected with the fishnet.

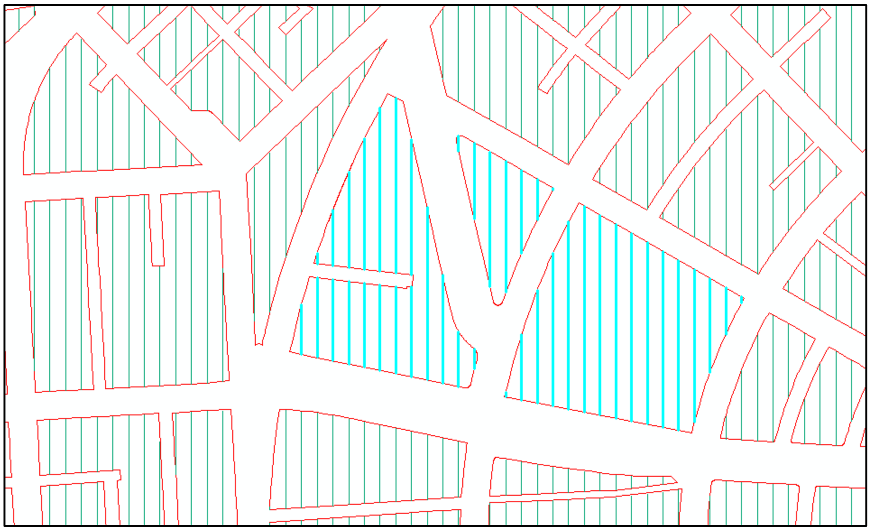


Figure 3.13 Intersected lines for sheet 174

Line segments in each map sheet footprint needed to be numbered sequentially to serve as a unique identifier, in order to be able to join the address range data. Using Feature to ASCII tool in ArcMap, the attributes of the intersected line features were exported into a text file. The Feature to ASCII tool automatically includes a coordinate pair and length attribute for each feature, but only the ObjectID and Sheet number attribute are necessary. The text file was loaded into RStudio. A short code, shown in Appendix A, was run. The code sorts the lines by their sheet number, and creates a sequence number for each feature. The sequence number is then concatenated with the sheet number to create a unique id number, separated by an underscore character. The resulting table is saved as a CSV for import into ArcMap. Using the ObjectID attribute, the resulting table is joined to the intersected fishnet lines in ArcMap. A new feature class is created.

A similar code was used for the table containing address ranges, also shown in Appendix A. The lines are sorted based on sheet number, then a sequence number is generated, and a unique identifier is created for each feature. The resulting table is imported into ArcMap and joined to the numbered line segments, based on the unique identifier previously created. While the line number tables and address description tables could be merged in R, joining in ArcMap allows for more flexibility as street description files must be edited periodically.

The footprint for sheet 174, depicted in Figure 3.13, above, contains twenty nine line segments. Each line segment is numbered sequentially: 174_1, 174_2, up to 174_29. In the street directories, nine address ranges are associated with the Sheet 174, shown in Table 3.1 on the next page. Employing the sequentially numbered value generated using the R code, the attributes developed from the index maps can finally be associated with a spatial feature in ArcMap.

Table 3.1 Street segments found on Sheet 174

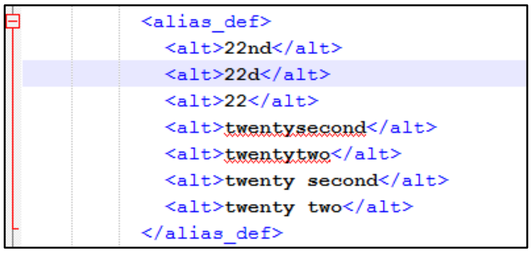
UniqueID	ST_NAME	ST_TYPE	FROM	TO	JoinID
174_1	BOND	STREET			1390
174_2	FRANKFORT	AVENUE			1392
174_3	GLEN PARK	STREET			1393
174_4	TONNINGSEN	STREET			1397
174_5	HOWARD	STREET	1600	1699	1394
174_6	MISSION	STREET	1601	1699	1395
174_7	FOLSOM	STREET	1699	1640	1391
174_8	THIRTEENTH	STREET	100	290	1396
174_9	TWELFTH	STREET	100	256	1398

3.5 Building the Address Locator in ArcMap

In ArcMap, a US Address locator requires linear features with these attributes: name, type, direction, joinID. The fishnet line features with the address ranges assigned to them meet these requirements. Default setting for the one-range U.S. Address locator functioned sufficiently. A slight modification was required to allow for matching addresses without an address.

3.5.1 Alias Table

Changes to street names are listed in on index pages and within street directories. An alias table was developed by digitizing the street list provided in the fire insurance map indexes and finding the corresponding street segments. Just eighty two street name changes were identified in this manner. Another group of changed names were identified by consulting street directories. The alias table simply requires a join ID for each segment and the modified name. Some aliases were created to correspond to frequently seen abbreviations. These included the ordinals "second" and "third", which are abbreviated "2d" and "3d", and other abbreviations unique to this period. Figure 3.14 shows how to edit the alias of the Address Locator style in the XML file "USAddress.lot", which is found in the Geocode folder of ArcMap system folders.



```
<alias_def>
  <alt>22nd</alt>
  <alt>22d</alt>
  <alt>22</alt>
  <alt>twentysecond</alt>
  <alt>twentytwo</alt>
  <alt>twenty second</alt>
  <alt>twenty two</alt>
</alias_def>
```

Figure 3.14 Editing the Address Locator style in XML

3.6 Employing the Address Locator

A group of addresses reflecting a discrete area of the city can illustrate the utility of the Address Locator in identifying correct map sheets. Sheet 43, which includes Washington Square in the district now known as North Beach was selected as the focus of the study area. The region contains many of the prominent streets of San Francisco of the period, including Montgomery Avenue, now known as Columbus Avenue and Dupont, now known as Grant Street. Seven adjacent map sheets, Sheets 31, 32, 42, 44, 55, 56, and 57, were also included in the study area, shown in Figure 3.15. Figure 3.16, shown on the next page, is a composite of georectified map sheets for the study area. Figure 3.17, also on the next page, shows the study area in relation to the city as a whole.

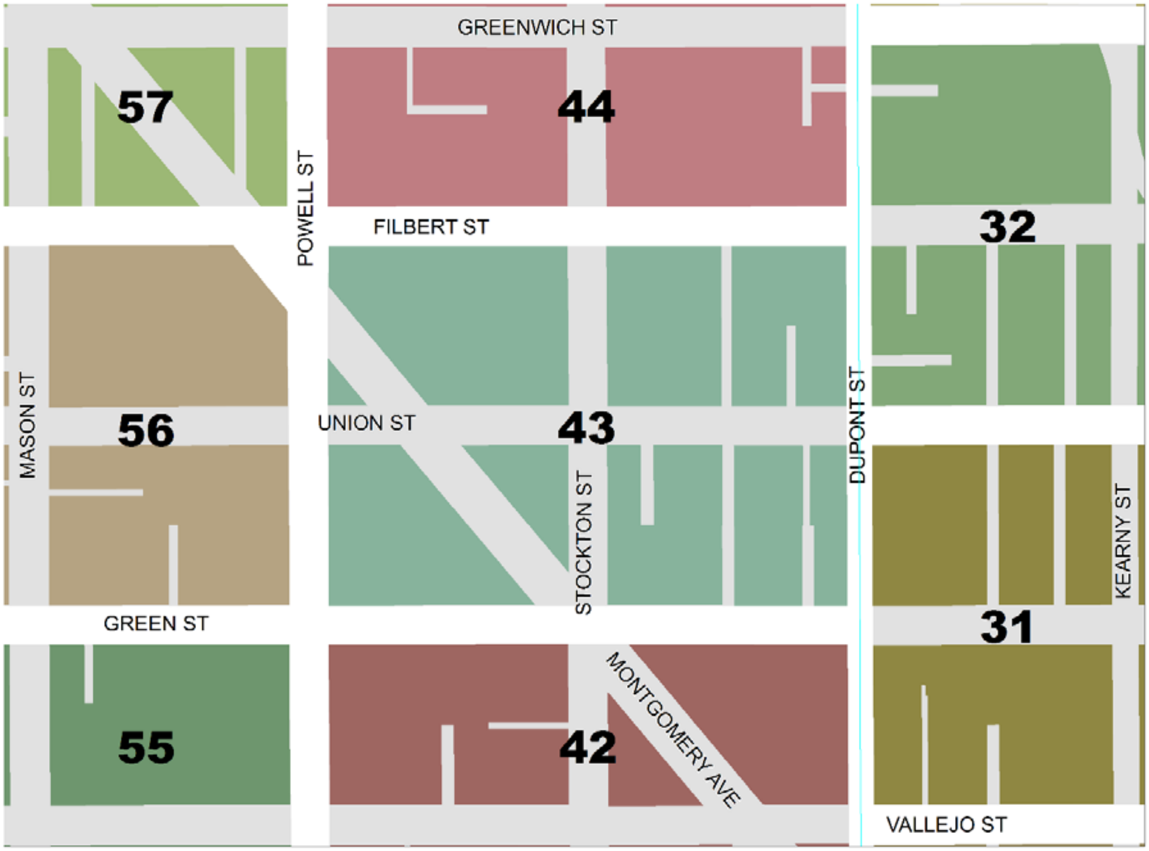


Figure 3.15 Study area

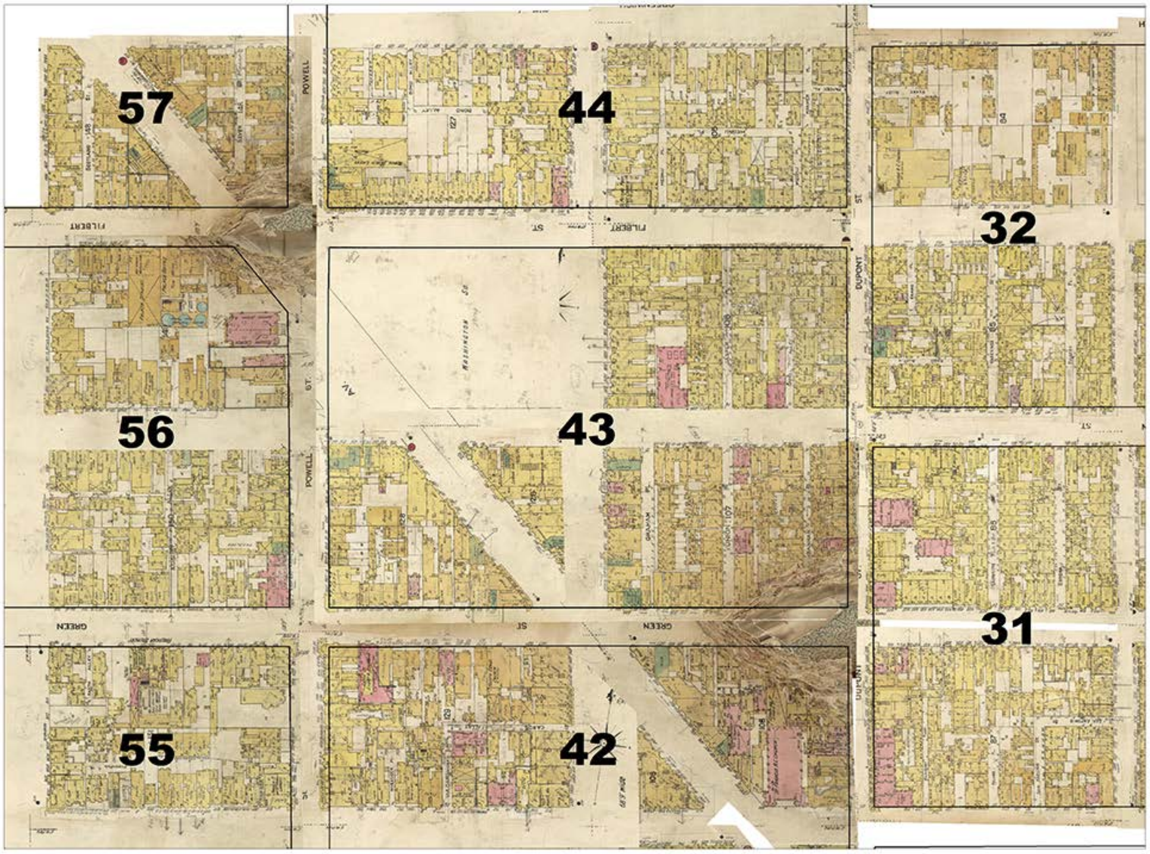


Figure 3.16 Composite of rectified map sheets

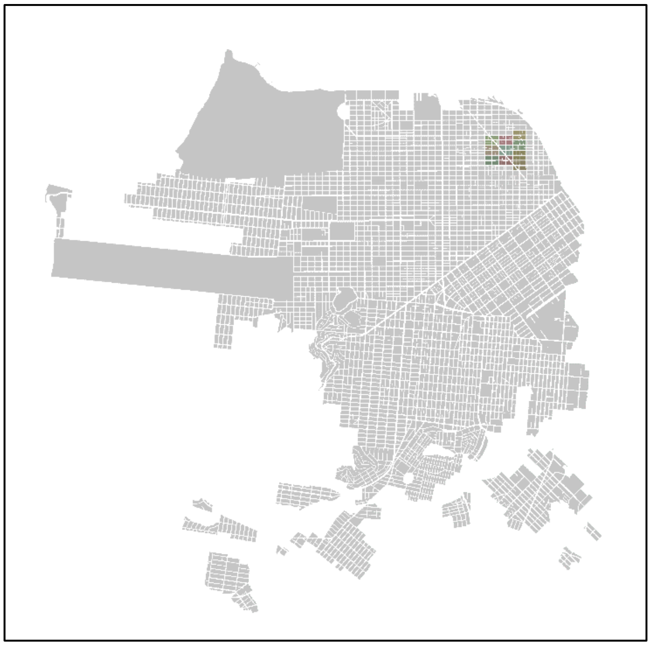


Figure 3.17 Study area in context

A list of addresses, shown in Table 3.2, were created as a table and added to ArcMap. The addresses are designed to illustrate the way that the locator responds to various conditions, including, addresses falling within a known address range, addresses outside of known ranges, and addresses that match multiple ranges.

Table 3.2 List of Test Addresses

No.	Address	Geocoding Result
1	1499 Dupont Street	Matched to Sheet 43
2	622 Green Street	Matched to Sheet 43
3	650 Green Street	Not Matched
4	635 Green Street	Matched to Sheet 42
5	501 Union	Tied Candidates
6	502 Union	Tied Candidates
7	8 Union Place	Matched to Sheet 43
8	541 Montgomery Ave	Matched to Sheet 43
9	543 Montgomery Ave	Tied Candidates
10	1001 Jasper Place	Matched to Sheet 43

The table was geocoded in ArcMap. Five of the addresses were matched correctly, and the other five matched but had other candidate matches. The address 1499 Dupont Street was matched readily, and the correct map sheet identified, shown in Figure 3.18.

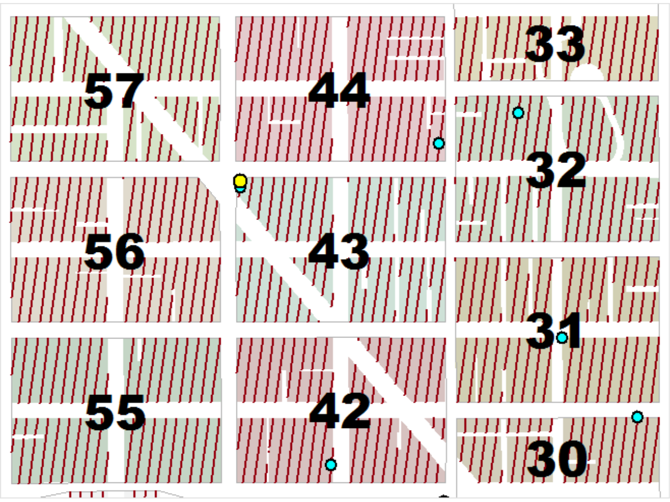


Figure 3.18 Interactive rematch for 1499 Dupont Street

Note that the candidate matches, shown as blue points, are scattered across the study area. This is an artifact of the dummy lines that fall arbitrarily on each map sheet footprint. The matched point, shown in yellow, is more than two blocks away from Dupont Street itself. Figure 3.19 shows Dupont Street in pink, and the corresponding dummy lines found on each of the adjacent map sheets, in blue.

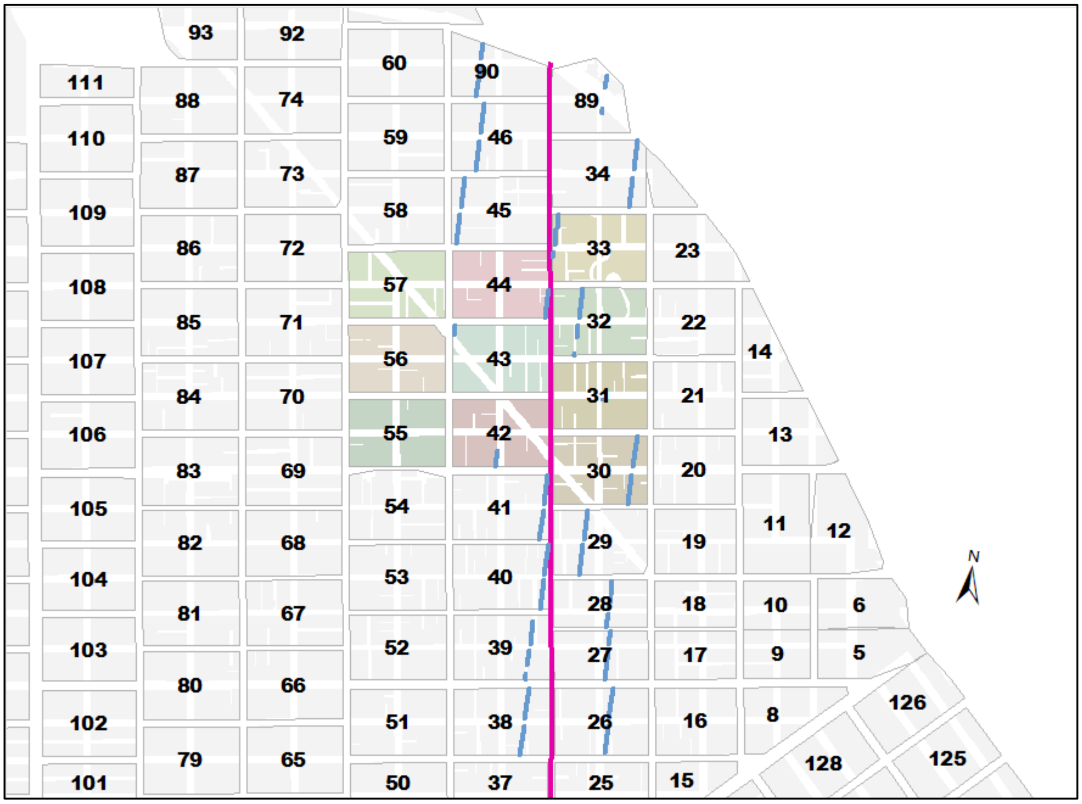


Figure 3.19 Dupont Street, shown in pink, and corresponding dummy lines, in blue.

The three addresses on Green Street are illustrative of how the locator deals with addresses falling within and outside of the address range associated with a street segment. The 500 and 600 Blocks of Green Street are divided between sheets 42 and 43. The upper limit of the 600 Block is 640, reflecting the numbering of buildings shown in the Sanborn map, shown in figure 3.20. Note that this figure is oriented with north at the bottom.

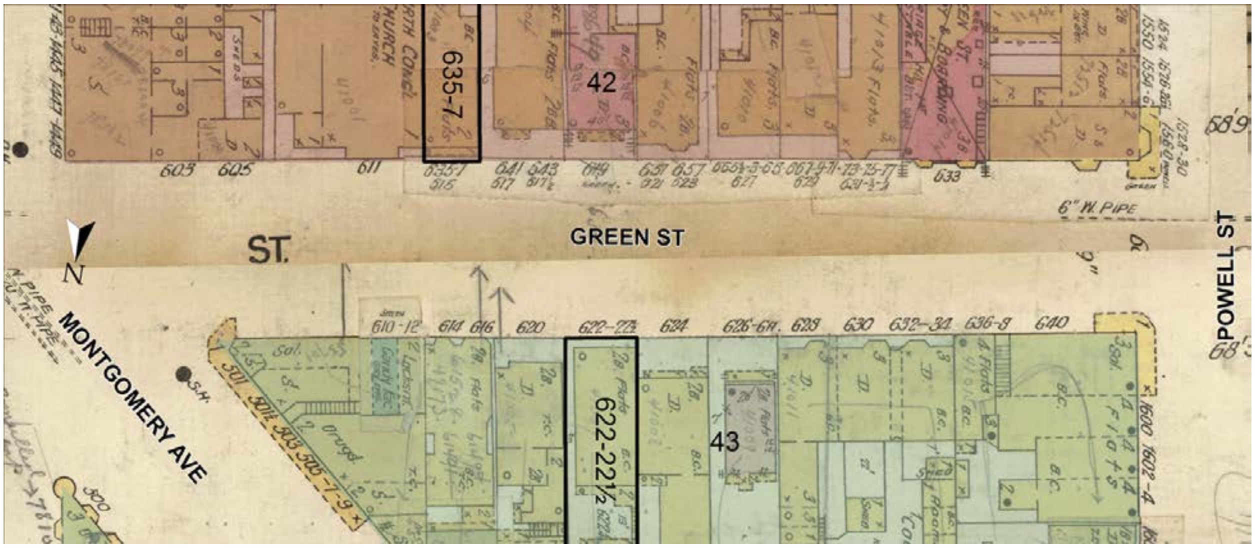


Figure 3.20 Detail of sheets 42 and 43, 600 Block of Green Street

The address 622 Green Street falls within the correct range. However, 650 Green Street is not found by the locator. Instead, the locator offered the twenty line segments associated with Green Street, as it did with Dupont Street. The task of identifying the correct range falls on the user. The odd range of the line segment, 501-635 Green Street, falls on sheet 42. Figure 3.21 illustrates the geocoded coordinates of 622 and 635 Green Streets. The correct locations on each map sheet is outlined in black. The locator is able to correctly identify the correct map sheet for odd and even ranges appearing on separate sheets.

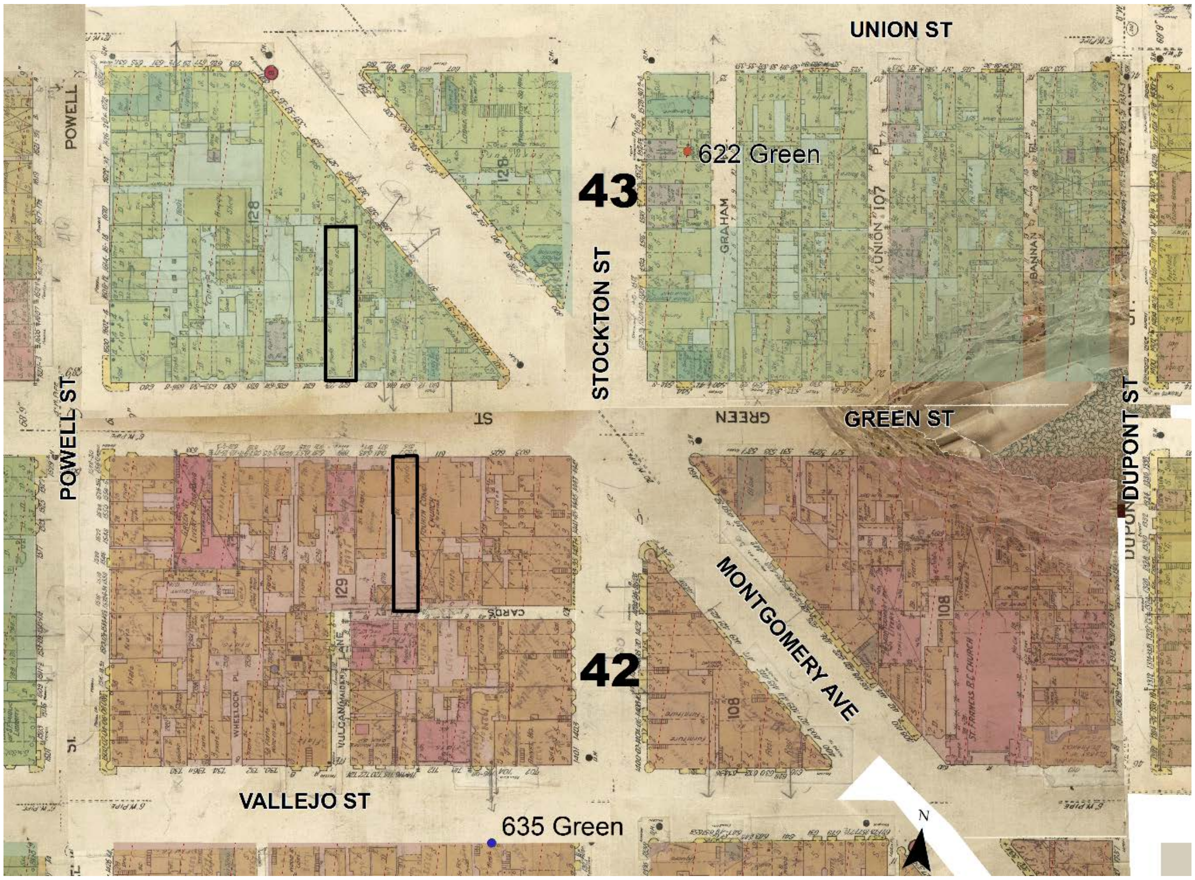
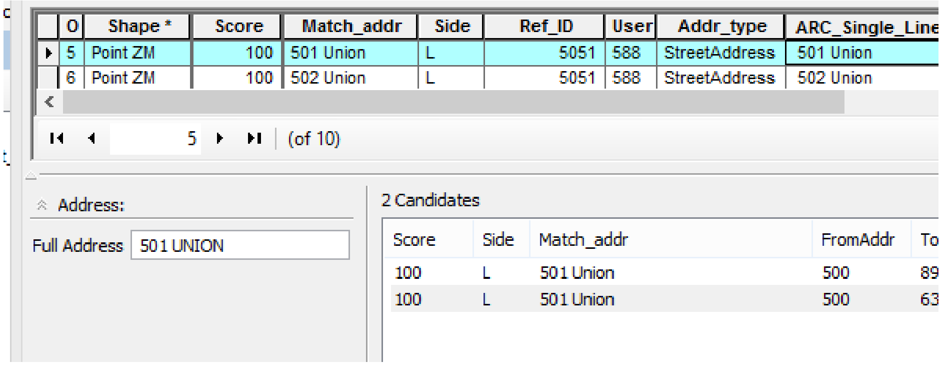


Figure 3.21 The positions of 635 and 622 Green, sheet as identified by geocoder

Addresses on Union Street illustrate a different problem. An entirely different street called Union also existed in Bernal Heights. For this reason, there are two equal candidate matches in the locator for “501 Union” without a street type specified: the Union Street falling on sheet 42, and the one falling on sheet 588, as seen in Figure 3.22.



	O	Shape *	Score	Match_addr	Side	Ref_ID	User	Addr_type	ARC_Single_Line
▸	5	Point ZM	100	501 Union	L	5051	588	StreetAddress	501 Union
	6	Point ZM	100	502 Union	L	5051	588	StreetAddress	502 Union

5 (of 10)

Address:

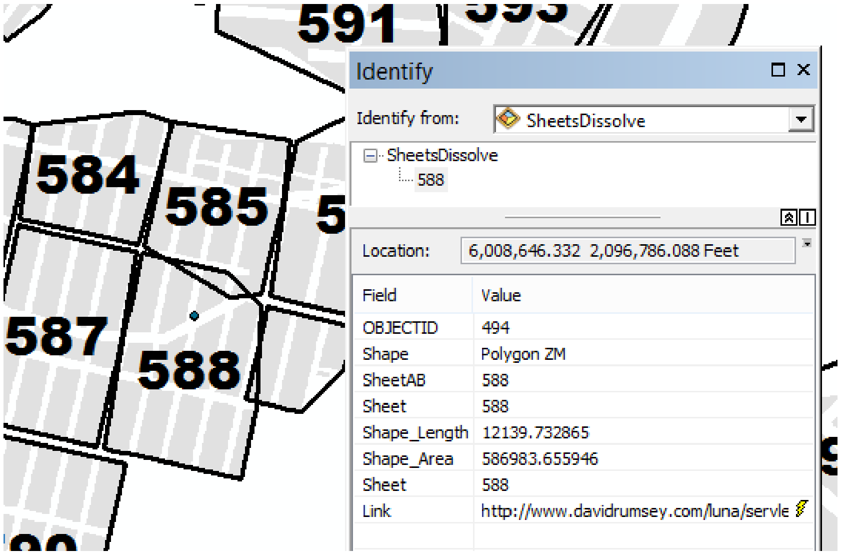
Full Address 501 UNION

2 Candidates

Score	Side	Match_addr	FromAddr	To
100	L	501 Union	500	89
100	L	501 Union	500	63

Figure 3.22 Tied candidate matches

To decide between these matches, the user can examine the insurance maps themselves, which provide more evidence to corroborate the presence of an address at the given location. Using the identify tool, candidate links to images of candidate map sheets can be accessed, shown in Figure 3.23.



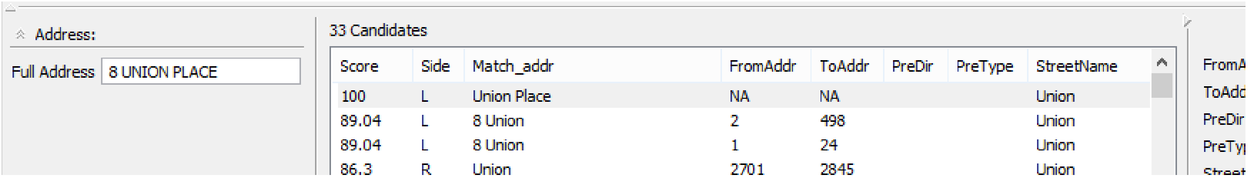
Identify from: SheetsDissolve

Location: 6,008,646.332 2,096,786.088 Feet

Field	Value
OBJECTID	494
Shape	Polygon ZM
SheetAB	588
Sheet	588
Shape_Length	12139.732865
Shape_Area	586983.655946
Sheet	588
Link	http://www.davidrumsey.com/luna/servle

Figure 3.23 Using the Identify Tool to access map sheet image

Examination of the map sheet revealed that no location existed at 501 Union in Bernal Heights. This context helps to resolve some of the ambiguities that the locator itself is unable to. However, addresses with similar names but different street types (e.g. Street, Avenue, Alley) match correctly. Despite the ambiguity between the different Union addresses, the address 8 Union Place matched to the correct segment, as shown in Figure 3.24. The match takes place without regard for the address number.



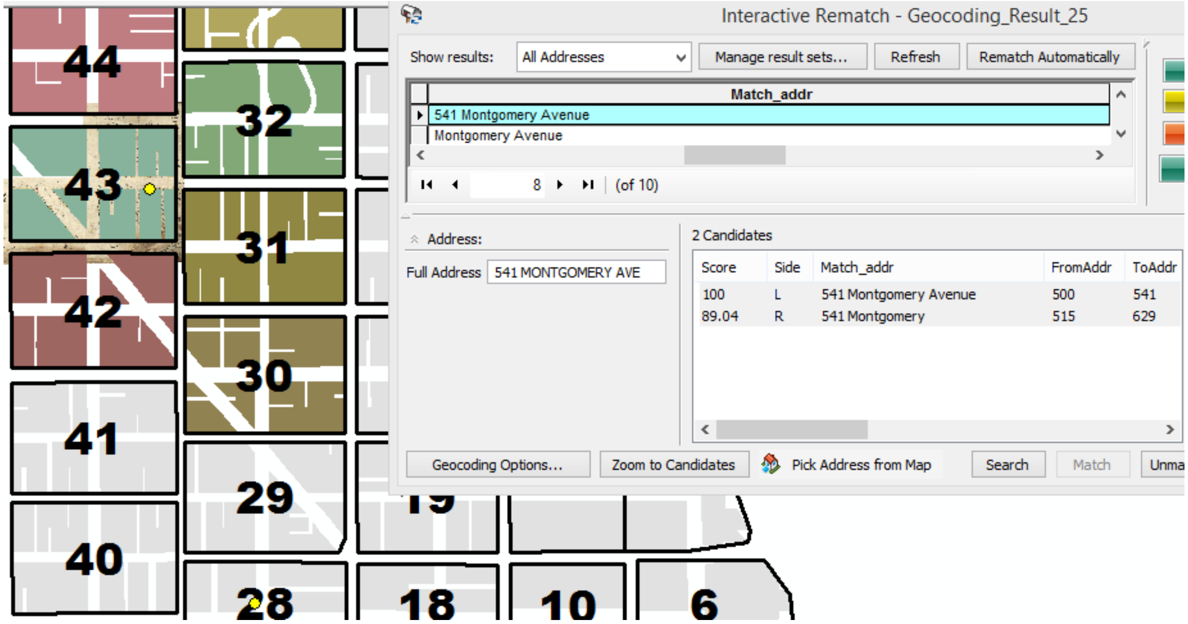
Full Address: 8 UNION PLACE

33 Candidates

Score	Side	Match_addr	FromAddr	ToAddr	PreDir	PreType	StreetName
100	L	Union Place	NA	NA			Union
89.04	L	8 Union	2	498			Union
89.04	L	8 Union	1	24			Union
86.3	R	Union	2701	2845			Union

Figure 3.24 Matching Union Place

Montgomery Avenue (now called Columbus Street) is a diagonal street that shares many address ranges with Montgomery Street. The locator correctly identifies the sheet for the address that falls within the correct address range (541 Montgomery Avenue), as shown in Figure 3.25, but it cannot distinguish between the ranges for 543 Montgomery Avenue, which falls outside of the correct range.



Show results: All Addresses

Match_addr
541 Montgomery Avenue
Montgomery Avenue

8 (of 10)

Address:

Full Address 541 MONTGOMERY AVE

2 Candidates

Score	Side	Match_addr	FromAddr	ToAddr
100	L	541 Montgomery Avenue	500	541
89.04	R	541 Montgomery	515	629

Figure 3.25 Distinguishing Montgomery Avenue from Montgomery Street

Jasper Place, a short street between Union and Filbert Streets, shown in Figure 3.26, does not have an address range associated with it in the street index, although the map shows addresses are assigned in this block. An address far outside of the appropriate range (1001) still matches to the correct map sheet, because the locator can match addresses without house numbers.

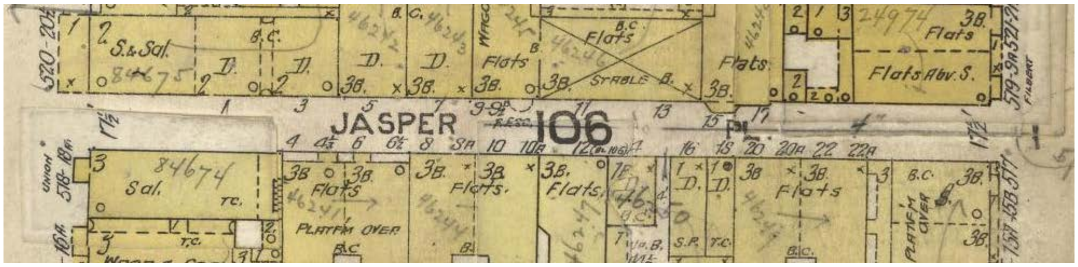


Figure 3.26 Detail of sheet 43, Sanborn Fire insurance map

3.7 Conclusion

Developing a geocoder based on insurance maps required the use of sophisticated Optical Character Recognition software to transcribe and correct the text of the address ranges. While the street indexes had a roughly tabular form, considerable effort was required to manipulate the text into a form that would match stringent data requirements of a contemporary address locator. Additionally, the map footprints were created by georeferencing index maps and recoding parcel data. The resulting geocoder allows users to visualize the rough position of geocodes within ArcMap, as well as a means to quickly find and inspect an image of the original insurance map sheets.

CHAPTER FOUR: APPLICATION AND EVALUATION OF GEOCODER

The objective in creating a historic address geocoder is to correctly identify the location of historic addresses. Mapping a large group of historic addresses can demonstrate the strengths of a geocoder, and the types of addresses that it fails to recognize. Errors can result from problems in the geocoding method, errors in transcription or problems in the source materials. Mapping a large set of addresses can shed light on the nature and characteristics of source materials that would otherwise be obscure. Using directory listings for bakeries reveals insights into the utility of business directories as well as fire insurance maps in research.

4.1 Bakeries of San Francisco

Listings in business directories are a convenient source of historical addresses, because they are structured in a way that is machine readable. The Crocker-Langley Directory was published annually, with alphabetical and classified listings within San Francisco. Researchers have relied on the Crocker-Langley business directories to identify locations of businesses or residences. Paul Groth (1994) employed listings for residential hotels and boarding houses to illustrate their distribution throughout San Francisco. Edith Sparks (2006) and Jessica Sewell (2011) mapped listings of groceries and other female-headed businesses to explain their role in commerce during the turn of the century.

Bakeries demonstrate both the limitations and the merits of the Sanborn maps as a data source. Bakeries were fire hazards, but they did not always bear the same attention paid to larger industrial hazards. They were also more dispersed throughout the city than industrial functions like paint production. Figure 4.1 shows a typical bakery found on a Sanborn map. Brick bakery ovens shown in red look distinct against the mostly timber-framed construction in San Francisco, coded yellow, making bakeries easy to identify visually.

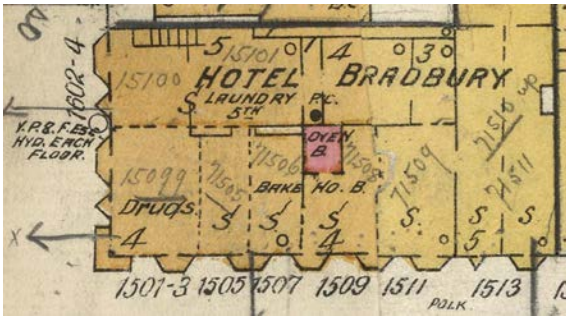
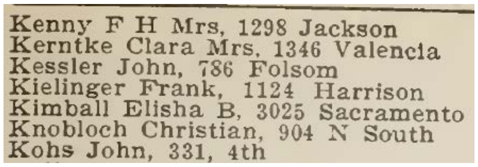


Figure 4.1 Detail of Sheet 103, Sanborn fire insurance map

4.1.1 Mapping Bakeries with the Sanborn Geocoder

To identify bakeries, listings under the heading "Bakeries" were copied from the digitized 1904 and 1905 editions of the Crocker-Langley San Francisco City Directory. The digitized text is available for download through the Internet Archive. Text was recognized using ABBYY FineReader, and reviewed for character recognition errors. Figure 4.2 shows that the listings follow a consistent structure: last name and first name, followed by the addresses separated from the name with a comma. An additional comma separated address numbers from numbered streets. The recognized document was saved as plain text, and opened within Microsoft Excel. The commas and line breaks of the listings parallel the structure of a comma-separated values (CSV) file.



Kenny F H Mrs, 1298 Jackson
Kerntke Clara Mrs, 1346 Valencia
Kessler John, 786 Folsom
Kielinger Frank, 1124 Harrison
Kimball Elisha B, 3025 Sacramento
Knobloch Christian, 904 N South
Kohs John, 331, 4th

Figure 4.2 A sample of listings from the 1904 directory shows their tabular structure.

Excel interprets the commas as column breaks when a text file is loaded. To remove the comma used as separation between numbered streets, the columns containing addresses and numbers were concatenated, forming a new, corrected address column. New lines were manually inserted for listings with multiple addresses. The names contained some clue as to the ownership of the bakeries, allowing for the creation of a third attribute by filtering the name column in Excel. Fifty nine of the names were corporate entities. One hundred twelve of the names contained the titles “Mrs” or “Miss”, identifying the proprietor as female (two additional names lacking titles were identified as female). The remaining two hundred names were labeled as male. The resulting file was imported into ArcMap as a list for geocoding.

The 1904 edition contained three hundred seventy one addresses. Of these, four addresses were duplicates, making three hundred sixty eight unique addresses. The table of addresses was geocoded using the Sanborn based geocoder. The geocoder located three hundred seven addresses, found multiple possible candidate matches for forty-two addresses, and failed to identify locations for nineteen addresses. However, geocode matches identified by the locator do not necessarily correspond to the reality on the ground. By comparing the resulting geocodes to the Sanborn maps, it is possible to verify the presence of bakeries at the mapped locations, and clarify some of the reasons for errors in locating addresses.

Following the geocoding process, the map sheet for each of the geocoded address was inspected to confirm that the address could be found. The Sanborn maps provide additional information that was used to classify the addresses into four categories: addresses with ovens, addresses labeled as a store or saloon with no oven, addresses labeled as dwellings, and addresses that were not found.

4.2 Assessing Geocode Errors

Roughly 16 per cent of addresses listed in the directories had tied candidate matches or failed to match addresses at all. Of the nineteen unmatched addresses, ten were intersections that could not be recognized in this locator, because line segments lack connectivity. Seven of the errors resulted from problems with optical character recognition and could be located once the addresses were corrected. Just two addresses could not be found at all. Problems with tied candidate address matches were more difficult to resolve. They resulted from errors in numerical ranges provided in the insurance maps index, introduced in the transcription process or from problems with distinguishing odd from even street ranges.

4.2.1 Range Overlaps

Numerical ranges provided in the index sometimes overlapped. For example, a bakery listed at 1587 Market Street matched to segments numbered 1501-1685 on sheet 144 and 1401-1599 on sheet 146. Market Street was numbered inconsistently. The block of Market depicted on sheet 144, shown in Figure 4.3 was also numbered 1201-1345. Both ranges potentially reflected the correct address. By examining the map sheets themselves, it was possible to identify the correct location, which appears on sheet 144, shown in Figure 4.4.

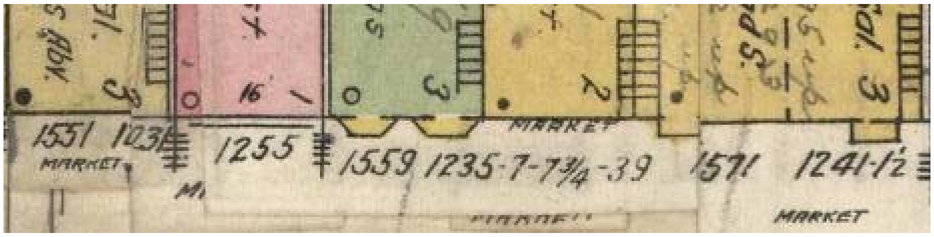


Figure 4.3: Detail of Map sheet 144 shows inconsistent numbering on Market Street.

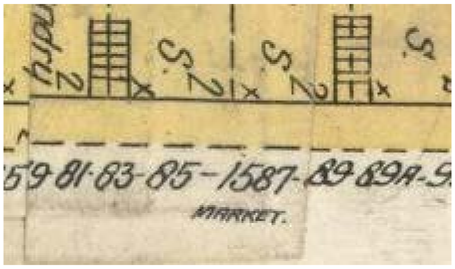


Figure 4.4: Detail of map sheet 146 shows 1587 Market Street.

Other instances of range overlaps were the result of OCR errors that were not identified through quality control, which included visual inspection of OCR results. A bakery listed at 2402

Folsom Street matched the range 2000-2748 Folsom. In this case, the numeral “1” was misrecognized as numeral “7”. Address ranges on most map sheets are rarely larger than 200. Scrutiny of disproportionate ranges can help to reduce these errors. It is possible to make changes to the geocoder as such errors are identified, to improve its accuracy.

4.2.2 Renumbering of Houses

A larger set of geocode errors resulted from house renumbering that took place between 1899 when this edition of maps were first published and 1905. Evidently, street indexes were developed by finding the upper range number for each street on a map sheet, as opposed to the theoretical address ranges of a block. Therefore, the even numbered 900 block of Valencia Street, which appears on sheet 640, appears in the index as 900-960. Having accurate upper ranges is advantageous in linear interpolation address locator, but can result in more false negatives (Zandbergen 2008). A bakery listed at 992 Valencia Street, shown in Figure 4.5, was previously numbered 948.

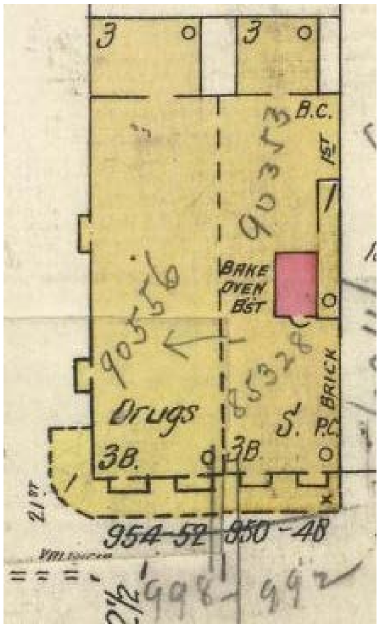
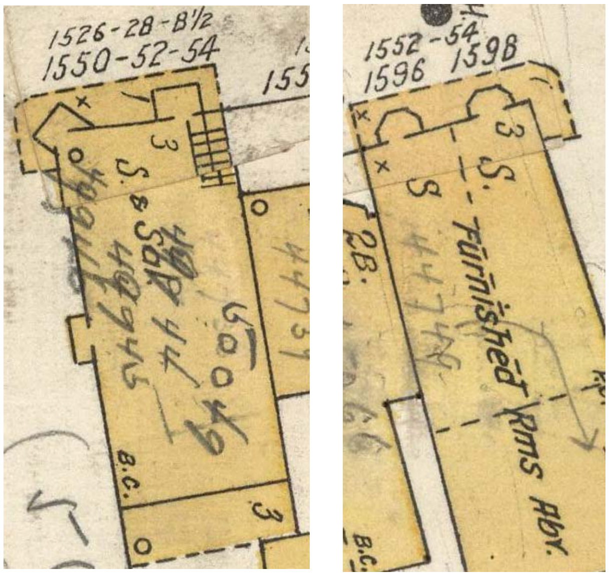


Figure 4.5 Detail of sheet 639

Errors related to changes in addresses illustrate just how unreliable a street address was in precisely identifying a location. Two buildings on one map sheet could be labeled with the same address. A bakery owned by Frank Wing at 1552 Howard Street may have been on the corner of Lafayette Street or the corner of Twelfth Street, depending if the newer of older address range was being observed. Figures 4.6a and 4.6b show two properties labeled with the same address.



Figures 4.6a and 4.6b Details of sheet 145

In total, fourty of the the listings for bakeries in the directory could not be matched to an address labeled on the maps. It is unclear if these ommisions reflect errors in the directories or errors in the Sanborn maps. One particularly confounding listing is that of Mrs. W. Waldeck of 1209 Larkin Street. Her place of business, (or residence) is shown as a block-wide salt water baths, in both the 1899 and 1904 versions of the Sanborn maps. It is possible that an address existed at that location prior to the construction of the baths, or that there was a typesetting error in the listing.

4.2.3 Classifying Bakeries

The directory listings make few distinctions among the bakeries, except for the few bakeries where Pies or Crackers were baked. However, visual inspection of the map sheets reveals important distinctions among the bakeries. While the presence of an oven might appear to be a requisite for the running of a bakery, some ovens were not shown at addresses listed in the directory. Figure 4.7 shows Christopher Aker’s Bakery at 1794 Haight. A baking oven is shown in red in the back room, but the address is also classified as a store and lunchroom (denoted with “S. & Lunch”), and a single family dwelling (denoted with a “D.”).

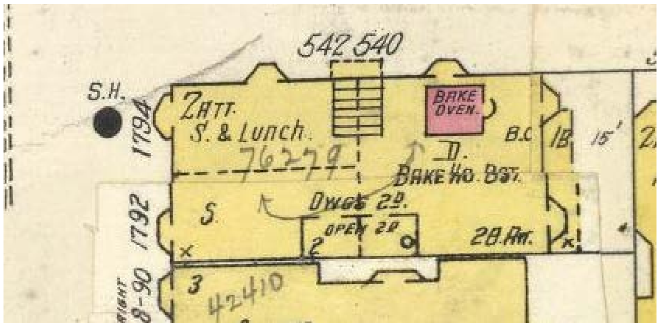


Figure 4.7 1794 Haight Street, Detail sheet 425

One hundred forty of the bakery listings are identified as a store, saloon or restaurant on the Sanborn maps, with no visible oven. Such listings may have sold bread produced off site, or their ovens may have been too small for the attention of the insurance surveyors. Two hundred and five listings matched addresses with ovens illustrated in them. Such listings corroborate the evidence that the location held a bakery. Figure 4.8 shows the distribution of bakeries throughout San Francisco based on these categories, which are not exclusive. Stores are illustrated with a small blue circle. Squares indicate culinary functions: red for lunch counters and cafeterias, teal for bakeries, blue for restaurants and green for saloons. Listings with ovens illustrated on site are circled in orange.

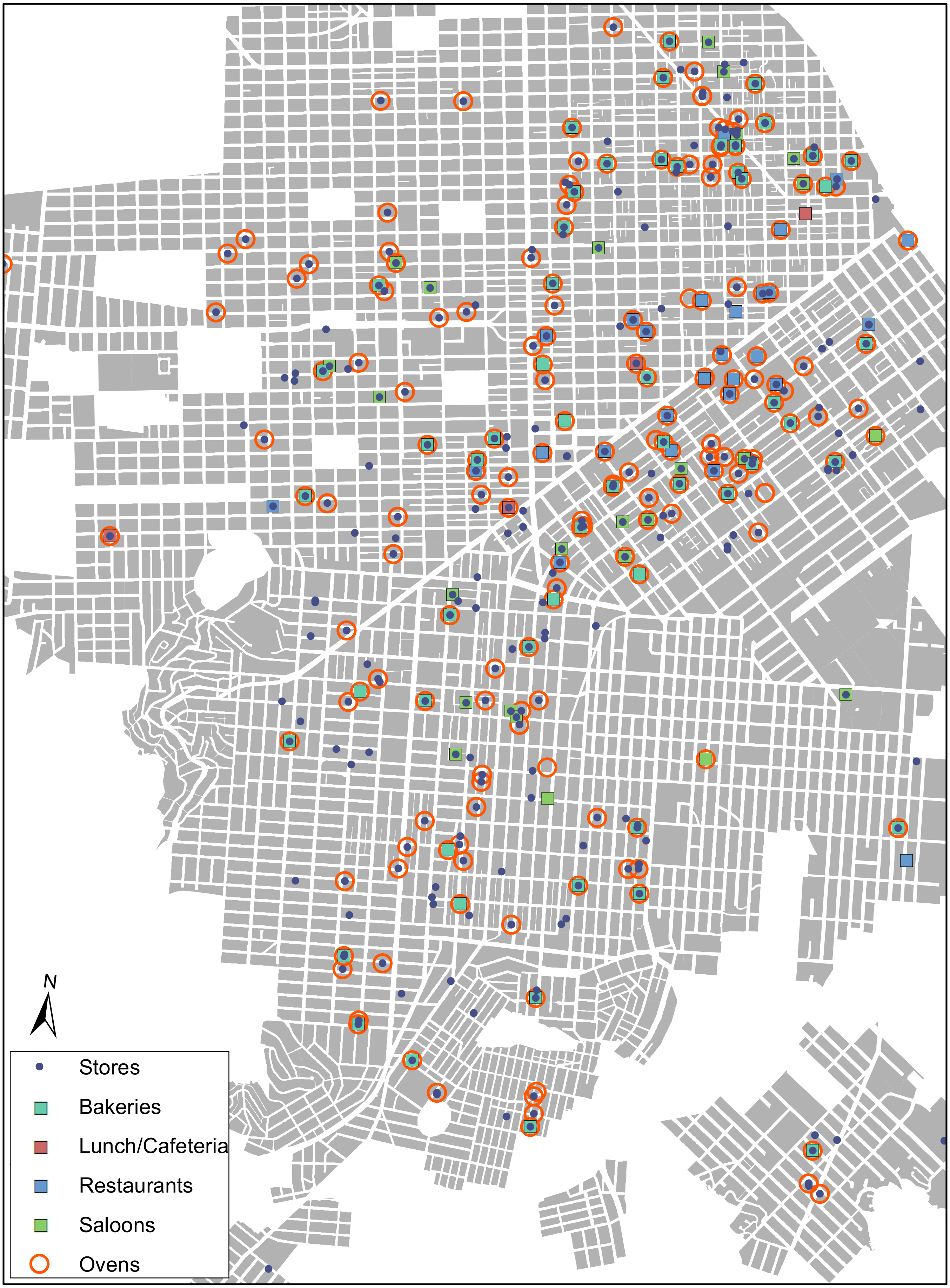


Figure 4.8 Distribution of Bakeries within San Francisco

4.3 Female Proprietorship

Thirty percent of the total listings had female proprietors. Eighty two, or 58 percent, of bakery listings depicted as stores on the Sanborn maps were headed by female proprietors. Just twelve of the two hundred five bakeries depicted with ovens on the maps were headed by women. Assuming that the directory correctly listed locations of bakeries, this discrepancy suggests that women were more likely to run smaller, storefront bakeries, while large bakeries were headed by men. Eleven of the listings corresponded to domestic buildings, with no store or factory visible on the Sanborn maps. Six of these listings, were headed by women. It is conceivable that such listings corresponded to home businesses.

4.4 Temporal Reliability of Sanborn Maps and Business Directories

The latest update present in this Sanborn maps edition were dated to 1905, but updates to the maps were conducted incrementally. Areas of the city with greater density and more change were surveyed carefully to reflect changes, while peripheral areas seemed to languish. Most updates were undated. Therefore, the maps are properly dated to a range of years—1899 to 1905. Not a snapshot, but a long exposure. The Crocker Langley directory, on the other hand, is temporally continuous. A record exists for each year, and differences between directories from year to year might represent change over time. Comparing the 1904 listings of bakeries to the 1905 listings might complicate that assumption. While they contain the same number of listings, mapping them demonstrates a dramatic change. One hundred nine bakeries were found at locations that were not found in the previous year’s listings, shown in Figure 4.9 on the following page. Much of this expansion pushes towards the western and southern edges of the city. It is nevertheless difficult to assert on the basis of this evidence that the presence of bakeries changed so rapidly.

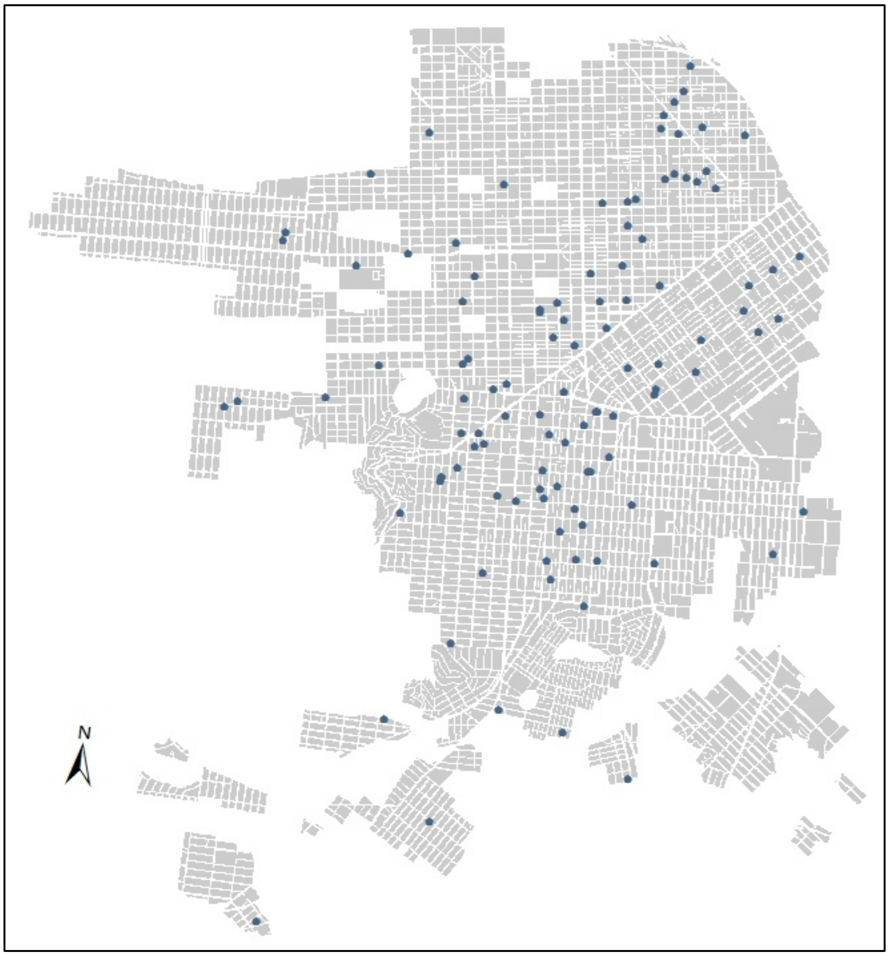


Figure 4.9 New bakeries found in 1905 directory

4.5 Comparison of Sanborn-based Geocoder and Contemporary Address Locator

A common approach to historical address geocoding is to modify contemporary street centerline files to comport with historical names, ranges and geometry (see, for example, Debats 2007). In order to compare the insurance map geocoder to a centerline geocoder, a street address locator was constructed using publicly available data from the City of San Francisco website. The shapefile contains all relevant attributes necessary for a functioning modern locator, but some of these attributes must be edited to function with historical addresses. Incorrectly identified historical addresses can indicate where the underlying data must be edited to create a functioning historical locator.

The centerline geocoder identified the correct map sheet for two hundred seventy bakeries, or 73 percent of the total. Centerline geocoded addresses are more precise than the insurance map based geocodes, when they can be verified, but this precision corresponds to an increase in plainly incorrect results. Ninety-seven were incorrectly identified. These are discussed below.

4.5.1 Renaming Errors

The simplest approach to developing a historic address locator is to create an alias table that joins renamed streets to corresponding present-day street segments using a modern centerline file. This approach is easy to accomplish, because lists of renamed streets are compiled frequently in directories and guidebooks. Sixteen of the erroneous geocodes resulted from streets being renamed. However, in many cases, the address ranges of streets changed along with the names. These errors cannot be resolved by an alias table alone. Seventeen additional errors were the result of ambiguity between numbered streets and avenues. Numbered avenues in the Richmond and Sunset Districts were in early stages of development in 1904.

4.5.2 Renumbering Errors

Fifty geocoding errors resulted from street address ranges changing, eighteen resulted in tied matches. To correct for such errors, individual line segments would have to be identified and edited to fit a broader range, or edited to match numerical ranges found on the Sanborn map sheets. Either approach would be relatively laborious.

4.5.3 Changes to Streets

Physical changes to streets sometimes resulted in unmatched geocodes when using the centerline geocoder. Certain blocks of streets were removed over time as the result of freeway construction, urban renewal, and other forms of development. Just five of the geocode errors are attributable to physical changes to streets. These errors could be ameliorated by drawing new street centerlines segments.

4.5.4 Intersections

The centerline based geocoder is able to correctly identify intersections. This is impossible using the Sanborn geocoder. However, the locator does not currently distinguish between directional indicators (e.g. “SW cor Mission and 11th”). This makes it necessary to take an extra step to match intersection geocodes to map sheets, because an intersection could potentially match to up to four distinct sheets.

4.6 Assessment of Geocoders

The performance of the two geocoders was comparable. In raw numbers, the centerline geocoder correctly identified two hundred seventy map sheets, while the insurance maps geocoder identified three hundred seven. Table 4.1, below, compares the match rates of the two geocoders. The Sanborn based geocoder found seventy one bakeries missed by the centerline geocoder. Thirteen bakeries missed by the Sanborn locator were correctly identified by the centerline geocoder. An additional ten intersections could only be located using the centerline based locator. The results suggest that the two methods could work in tandem, serving as a basis to check one result against another.

Table 4.1 Comparison of Insurance Map and Centerline Geocoders

	Insurance Map Geocoder	**Centerline Geocoder**
Correctly Matched	307 (83.4%)	270 (73.3%)
Multiple Candidate Matches	42 (11.4%)	27 (7.3%)
Unmatched	19 (5.1%)	71 (19.3%)
Total Listings	368 (100%)	368 (100%)
Matches missed by other geocoder	71	13

4.7 Summary of Geocoding Test

Mapping the Crocker Langley Directory using the insurance maps geocoder demonstrates the utility of the tool and uncovers interesting limitations in the both the directory and Sanborn maps. While both sources purport to inventory business and industry in San Francisco more or less comprehensively, neither source appears to be exhaustive. Checking the directory listings against the Sanborn maps suggests that the bakery category was more heterogeneous than the directory suggests. The insurance maps help not only to validate the geocode, but to provide visual context that informs how they can be interpreted.

The insurance map geocoder modestly outperformed the centerline geocoder in identifying the correct map sheet for bakeries listed in the directory. The centerline geocoder did not match or incorrectly identified seventy one listings, while the insurance map geocoder incorrectly identified just thirteen listings. The relatively high mismatch rates of both geocoders confirms the need to employ reference data to confirm geocodes. It is important to note that the directory listings of bakeries do not represent a random sample of historic addresses. Bakeries were commercial enterprises, and tended to be listed on major streets. It is possible that a different set of private addresses would perform better using the insurance map geocoder.

CHAPTER FIVE: DISCUSSION

Employing fire insurance map indexes to develop a geocoder demonstrates how the structure of historical data sources can be exploited within a GISystem. This process also sheds light on the implications of overlaying large-scale historical maps with modern data, and their value in providing context for historical data. Sanborn maps are widely used by researchers, but it is important to understand their peculiarities to evaluate suitability as a source document. A historic geocoder can serve as a means to make better sense of the relationship between maps, texts and the urban environment they represent.

This chapter reviews the insights gleaned through the process of developing a geocoder from the Sanborn map indexes. Section 5.1 explores the implications of viewing insurance maps with a GISystem. Section 5.2 articulates how Sanborn maps of San Francisco represent the turn-of-century built environment, delineating the types of data that can be extracted from them. Section 5.3 assesses how the geocoding approach allows for more direct access to data represented on the insurance maps. Section 5.4 suggests how the process of developing insurance map data into a geocoder can be improved and what additional steps must be taken to improve its functionality. Finally, Section 5.5 answers the research questions posed in Chapter 1.

5.1 Reading Insurance Maps through a GISystem

The process of digitization fundamentally transforms the way that researchers interact with historical sources. Insurance map volumes are heavy, awkward to maneuver and rare. Stooped over a table in a library leafing through map sheets, a researcher mirrors the posture of countless underwriters who used the same volume with different questions or intentions. The physicality of this process cannot be duplicated scrolling through microfilms or zooming and panning images on a computer screen. Manipulating fire insurance maps with a GISystem further transforms the map viewing experience. GISystems spare users the task of identifying map sheets and assessing orientation, tasks that required careful examination of index maps and street directories. GISystems make overlaying insurance maps with other sources of data straightforward.

5.1.1 Overlaying Insurance Maps with Contemporary Data

Overlay is one of the most basic functions of a GISystem. Without a GISystem, comparing two maps of the same region is a tedious, time consuming, and imprecise task (Gregory and Ell 2007). Within a GISystem, the process of data overlay is so fundamental that it almost escapes notice. Nevertheless, the decision of which data to overlay informs the types of questions that can be asked and the context for understanding. Temporal context is an important consideration. Contemporary data about San Francisco are plentiful, while historical data are rare.

Contemporary base maps display landmarks that allow users to understand how the historical maps relate to the physical world. Superimposing fire insurance maps on familiar contemporary base maps focuses the user on questions of contrast and change. The juxtaposition of the historical fabric of tenements, boardinghouses and factories with homogenous superblocks of office buildings downtown underscores the ways that San Francisco was transformed over the twentieth century. However, this contrast does not explain the process of development. Maps and other historical sources from intervening stages of change are needed to construct a narrative. Nonetheless, visual overlay of insurance maps can provide important insights. One can situate a historical building within its past context, with clues about past uses and architectural additions. Maps can also serve as a basis to discuss political and social issues like housing and labor. On their own fire insurance maps tell a limited story, with a disproportionate emphasis on hazards.

Maps are an important part of a city’s cultural heritage, not simply as a data source, but as a record of ways people once lived. Preservationists rely on fire insurance maps to demonstrate the historical or cultural significance of a building. Maps illustrate historical context in a way that other sources cannot. In areas of the city where buildings from before the earthquake still exist, they speak to the integrity of the urban fabric. In tandem with an architectural survey or inventory and archival sources, maps provide evidence of a building's historical significance.

5.1.2 Overlaying Historical Data on Insurance Maps

Comparing fire insurance maps to the present day city puts them into a familiar spatial context, making it easier to navigate the map sheets. However, such comparisons may tell us more about the contemporary city than the city of the past. Situating our understanding in the present day, we are confronted with data extraneous to the experience of early twentieth century San Francisco. Insurance maps are a rich, yet limited source of information. They depict the built environment, emphasizing hazards over other cultural or social phenomena that exist in an urban landscape. Yet insurance maps can provide much needed context for other historical data sources. Mapping historical textual sources containing addresses, such as directories, newspaper articles and advertising and business records in the context provided by the Sanborn maps can expand and complicate our knowledge of the historical city.

Ultimately, we read a fire insurance map in the present, finding significance and surprises in details intended as mater-of-fact records. It is essential to keep the essential function of insurance maps in mind. Sanborns weren't used for navigation; they were a means of risk assessment. Such limitations parallel concerns in modern GIScience literature regarding fitness of use. GISystems allow users to overlay data from a variety of sources, without regard to methods of data collection or classification. Metadata provide careful users with a means of understanding the methodology employed to derive data. Yet historical sources require more careful scrutiny than contemporary data, precisely because the origins and methods of development are obscure.

5.2 Extracting Data from Sanborn Maps

All maps have blind spots or phenomena that escape the attention of the mapmakers. The limitations of insurance maps are difficult to articulate without a thorough comparison of the maps to other historical sources. The omissions of insurance maps merit further investigation, but it is important to begin by articulating the types of data evidenced in an insurance map. A GISystem requires that data be classified into discrete categories. While insurance maps are not as cleanly categorized as digitally derived data, there are attributes that can be identified on all maps related to three broad categories: infrastructure, landscape and buildings. Hazards informed the surveying and mapping process. Keeping this focus in mind can help to explain what is being represented, and what is missing from the maps.

5.2.1 Infrastructure

Parcel and lot boundaries formed the basis of map sheets. In general, San Francisco's historical parcel boundaries correspond to modern boundaries, except in areas where streets were widened, moved or rebuilt. Names of streets were written and their widths were noted, though the maps do not show these dimensions at scale. Materials of streets are noted only when they differ from materials described at the beginning of each volume. Water mains were illustrated at street intersections with their widths. Water hydrants are marked. However, other major aspects of urban infrastructure escaped their attention. Electrical wires, communications infrastructure, and streetcar tracks were absent, while heavy rail tracks were carefully documented.

5.2.2 Environment

While Sanborn maps existed primarily to record the built environment, they illustrated significant natural features, like shorelines and other bodies of water. Despite San Francisco's complex topography, however, elevation and slope was generally ignored. Steep hillsides, rock outcroppings and other natural features were shown only occasionally, as shown in Figure 5.1. Parks, squares, and cemeteries were rarely mapped.

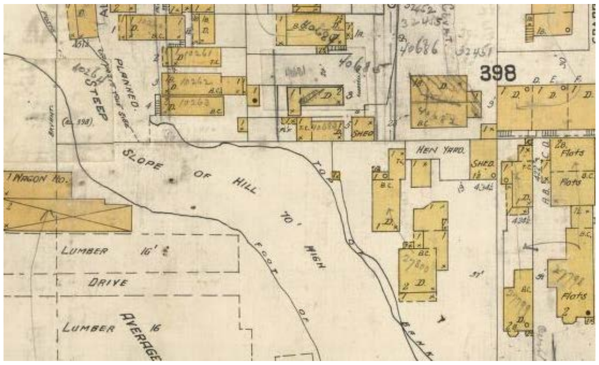


Figure 5.1 Schematic depiction of hill contours, Sheet 175, Sanborn Insurance Map

5.2.3 Buildings

Insurance maps focused on structural elements of the city. Buildings were carefully documented, and their uses carefully described. Buildings were drawn in great detail, with an emphasis on projections like bay windows or porches. Color was used to indicate building materials, and some materials and construction techniques were indicated in writing. However, some architectural elements were more carefully documented than others. Dimensions of light wells and alleyways were often recorded, but windows and doors were rarely indicated.

The attention paid to building uses depended on their relevance to fire insurance. Residential functions were categorized, and the number of housing units present in a building usually indicated. The scrutiny paid to commercial and industrial uses varied. Entertainment facilities like theaters, amusement parks and concert halls were named and documented, but other commercial functions outside of the hazard focus were portrayed with a simple abbreviations. Shops and restaurants were rarely expressly named or detailed. However, certain industrial functions, such as confectioners or candy making, evidently had significance for fire insurance.

5.3 Assessment of Geocoding Approach

Sanborn maps prove themselves to be a valuable source of reference information to contextualize historic addresses and provide a more holistic view of the architectural environment of pre-earthquake San Francisco. However, this context could be provided by simply vectorizing the index maps, without taking the additional step of digitizing the text-based street directories. The insurance map geocoder modestly outperforms the centerline geocoder in identifying the correct map sheets, but it has a distinct advantage over the centerline geocoder, in that it does not introduce extraneous modern data into the process. It situates addresses within their historical context.

Before endeavoring to create a historic geocoder, it is important to test and understand the types of errors for addresses being geocoded. Some errors can be addressed simply by editing and modifying centerlines. However, insurance maps are an unparalleled source to verify historical addresses, despite their limitations.

5.4 Further Work

A historical geocoder functions as a piece of the spatial data infrastructure necessary to develop a spatial understanding of the historical urban environment. I have shown that insurance index maps and the associated street indexes contain the principal components necessary to develop a functioning geocoder that provides historical context. The process of employing street indexes could be repeated for other cities with extensive Sanborn coverage, particularly in cities like San Francisco where changes in street numbering make street centerline geocoders unreliable.

5.4.1 Improving Geocoder Precision

While aggregating geocodes to map sheets provides context to evaluate whether or not a geocode is correct, a more precise geocoder would create more flexible point data that can be re-aggregated for sensitivity analysis. Given the variety of textual sources that can be mapped using this method, discrete points corresponding to individual records are often preferable, as opposed to a homogenous aggregated count of observations occurring on a sheet. As outlined in section 4.2, the suitability of existing centerline data can be evaluated by geocoding an address with both a centerline-based geocoder and the insurance maps geocoder. As long as both geocodes fall within the same map footprint, the centerline geocode can be assumed to be more precise.

By mapping larger sets of historical addresses using this method, it will be possible to identify streets where centerlines need to be redrawn or modified. Large areas of San Francisco were not mapped by Sanborn, but these unmapped areas exist mostly in low-density regions of the city. By identifying addresses that were not mapped in the Sanborns, a more systematic understanding of the surveyors’ blind spots and unmapped regions of the city can be developed.

Building Inspector, the public participation map vectorization project run by the New York Public Library, will ultimately develop sufficient data to create a parcel geocoder for historical addresses in New York City. To replicate this effort in San Francisco, the insurance map sheets must be satisfactorily georeferenced, with attribute data, including street addresses, assigned to each parcel. However, parcel geocoders have greater rates of false negatives, because they will only match known addresses (Zandbergen 2008). The ambiguity of San Francisco’s addresses during this period could mean that a parcel geocoder based on data derived from Sanborn maps would fail to geocode addresses that existed but were unlabeled in the Sanborn maps.

5.5 Conclusions

This thesis asked three principal questions. First, could historical data from insurance map indexes meet the technical requirements of a modern geocoder? Second, would a geocoder based on the indexes of fire insurance maps perform better than other types of historical address geocoders? Lastly, would the benefits of an insurance map geocoder outweigh the costs of the undertaking? These questions are not answered definitively by this study, but the project has demonstrated that the approach is feasible and can help to uncover valuable insights about early twentieth century San Francisco and other historical cities.

5.5.1 Exploiting the Structure of Insurance Map Indexes

Adapting Sanborn map indexes to a GISystems geocoder demands an understanding of both the historical data and modern technical requirements. The regularity of the insurance map indexes made it possible to capture the necessary data with minimal manual editing. Not all historical data sources offer structure that so readily lends itself to digitization, but the Sanborn map indexes can be digitized using available OCR software due to their consistent format. Street indexes and geocoders serve similar functions—finding locations on the basis of street names and house numbers. As such, they share similar attributes. Understanding the navigational elements of historical geographical sources can make it easier to exploit their structure within a GISystem. Fire insurance map indexes differ from street directories because they associate street attributes with a geographical object—the map sheet. Insurance map indexes are uniquely suited to adaptation because of their structure.

5.5.2 Comparison of Approaches to Historic Geocoding

The principal approach to creating historical geocoders is editing centerline files. However, editing modern centerline data is a laborious process, and depends on high quality reference data.

Small scale maps and street directories can be employed to this end, but this process cannot be automated, and requires a thorough understanding of the ways that the modern street system differs from the historical streets. Few resources parallel insurance maps in their level of detail and reliability as reference data for historical addresses. Modern centerline files are filled with historically irrelevant information that complicate the task of identifying address locations. Centerline address geocoders offer more precision than a geocoder that identifies map sheets, but such precision is moot if geocodes cannot be verified in their turn-of-the century context. Insurance map indexes can be exploited more quickly than editing centerline data.

5.5.3 Costs and Benefits of an Insurance Map Geocoder

It is clear that transcribing and digitizing Sanborn indexes is time consuming, but this process can be automated more readily than the task of researching and editing centerline files. Insurance maps provide historical context to the geocoded address, allowing the users to verify the presence of an address, and also to visualize the surrounding environment. The map sheet footprints derived by digitizing the index maps provide this context, regardless of how the location is found. However, because the street indexes were created at the same time as the maps, they ensure that geocodes are historically accurate. Identifying a map sheet provides a straightforward means to confirm the presence of an address visually. It also underscores the inherent imprecision of geocoding historical addresses.

5.6 Final Remarks

This thesis demonstrates the feasibility of using the Sanborn map indexes for the development of an address locator. Geocoding historical addresses is contingent upon the specific temporal and spatial relevance of the reference information at hand. In many cases, a modern geocoder can function well to identify addresses in cities that have not experienced major changes. However, the performance of any geocoder must be checked against temporally relevant sources to insure that the geocoded addresses are meaningful. Insurance maps are one of the most reliable historical sources of information about the built environment of nineteenth century U.S. cities. For this reason, they function as an ideal means to verify historical addresses.

REFERENCES

Bolstad, P. V., P. Gessler and T. M. Lillesand. 1990. "Positional uncertainty in manually digitized map data." *International Journal of Geographical Information Science*, 4(4): 399-412.

Colten, C. E. 1991. "Illinois Sanborn Geographic Information System." In Proceedings of the 34th Annual Engineering Geologists Meeting, Chicago, IL. October 3, 1991.

Chiang, Y. Y., C. A. Knoblock, C. Shahabi, C. C. Chen. 2009. "Automatic and accurate extraction of road intersections from raster maps." *GeoInformatica* 13(2): 121-157.

Esri. 2010. "Customizing Locators in ArcGIS 10." http://www.arcgis.com/home/item.html?id=aeb00de638f3492a93308a4a03183c7d. Accessed 18 January, 2015.

Esri. 2015. "Commonly used address locator styles." http://resources.arcgis.com/en/help/main/10.1/index.html#/Commonly_used_address_locator_styles/00250000000v000000. Accessed 18 January, 2015.

Gregory, I. N., and A. Hardie. 2011. "Visual GISting: bringing together corpus linguistics and Geographical Information Systems." *Literary and Linguistic Computing* 26(3): 297-314.

Gregory, I. N. and P. Ell. 2007. *Historical GIS: Technologies, Methodologies and Scholarship*. Cambridge, UK: Cambridge University Press.

Goldberg, D. W., J. P. Wilson, and C. A. Knoblok. 2007. "From Text to Geographic Coordinates: The Current State of Geocoding." *URISA Journal* 19(1): 33-46.

Harris, T. M., L. J. Rouse, and S. Bergeron. 2010. "The geospatial semantic web, pareto GIS, and the humanities." In *The Spatial Humanities: GIS and the Future of Humanities Scholarship*, edited by. J. Bodenhamer, J. Corrigan and T. M. Harris, 124-142, Bloomington, Indiana: Indiana University Press.

Hoehn, P. "Union List of Sanborn & Other Fire Insurance Maps." http://www.lib.berkeley.edu/EART/sanborn_union_list. Accessed October 23, 2014.

Hicks-Judd Company. 1901. *The San Francisco Block Book*. San Francisco: Hicks-Judd Company.

Karrow, R. and R. E. Grim. 1990. "Two examples of thematic maps: Civil War and fire insurance maps." In *From sea charts to satellite images: interpreting North American history through maps*, 3.ed, edited by D. Buisseret, 213-220. Chicago: University of Chicago Press.

Keister, K. 1993. "Charts of Change." *Historic Preservation* 45(3): 42-49, 91-92.

Keller, W. B. 1993. "Collecting and using fire insurance and real estate atlases: an individual perspective." *Art Reference Services Quarterly* 1(3): 31-48.

Leonard, A. E. and P. Spellane. 2013. "Using Old Maps and New Methods to Discover the Early Chemicals and Petroleum Industries of Newtown Creek in New York City." *Journal of Map & Geography Libraries* 9(1-2): 25-43.

Lamb, R. A. 1961. "The Sanborn map: a tool for the geographer." *California Geographer* 2: 19-22.

Lutkenhaus, B. 2002. "Digital Sanborn maps, 1867-1970." *Reference Reviews* 16(3): 51.

MacDonald, S. and N. Osborne. 2013. "AddressingHistory—Crowdsourcing a Nation's Past." *Journal of Map and Geography Libraries*, 9: 191-214.

Migurski, M. 2011. "Sanborn Atlas." http://www.maptcha.org. Accessed November 4, 2014,

New York Public Library. 2010. "NYPL Map Warper." http://maps.nypl.org/warper. Accessed December 26, 2014.

Oswald, D. L. 1997. *Fire Insurance Maps: Their History and Applications.* College Station, TX: Lacewing Press.

Page, M. C. K. Durante and R. Gue. 2013. "Modeling the History of the City." *Journal of Map and Geography Libraries* 9(1-2): 128-139.

Patton, D. K., A. K. Lobben and B. M. C. Pape. 2005. "Mapping Cities and Towns in the Late Nineteenth and Early Twentieth Centuries: A Look at Plat, Sanborn, and Panoramic Mapping Activities in Michigan." *Michigan Historical Review* 31(1): 93-122.

Raymond, A. 2011. "Denny Regrade, 1893–2008: A Case Study in Historical GIS." *Social Science History* 35(4): 571-597.

Ristow, W. W. 1968. "United States Fire Insurance and Underwriters maps 1852—1968." *Quarterly Journal of the Library of Congress* 25(3): 194-218.

San Francisco Department of Public Works. 2014. "Historical Block Diagrams." http://bsm.sfdpw.org/subdivision/assessor/. Accessed October 22, 2014.

San Francisco Planning Department. 2014. "San Francisco Property Information Map." http://propertymap.sfplanning.org/. Accessed October 22, 2014.

Sommer, L. 2011. "Mapping Project Reveals Pre-1906 Quake San Francisco." http://ww2.kqed.org/news/2011/08/16/mapping-project-reveals-pre-earthquake-san-francisco. Accessed October 22, 2014.

Torget, A. J., R. Mihalcea, J. Christensen, and G. McGhee. 2011. “Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers.” White paper, University of North Texas Digital Library.

Vershbow, B. 2013. "NYPL Labs: Hacking the Library." *Journal of Library Administration* 53(1): 79-96.

Wrigley, R. L. 1949. "The Sanborn Map as a Source of Land Use Information for City Planning." *Land Economics* 25(2): 216-219.

Yale University Library. 2013. "Yale University Library - The Map Collection / Sanborn Fire Insurance Maps." http://www.library.yale.edu/MapColl/print_sanborn.html. Accessed October 23, 2014.

Yuan, M. 2013. "Mapping Text.” In *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Edited by D. J. Bodenhamer, J. Corrigan and T. M. Harris, 109-123. Bloomington, Indiana: Indiana University Press.

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

ATTRIBUTIONS

Images of Sanborn maps of San Francisco from the David Rumsey Map Collection are used under the Attribution-NonCommercial-ShareAlike 3.0 license, http://creativecommons.org/licenses/by-nc-sa/3.0/.
Links to original files for each figure are included below:

Figure **Page**

Figure 3.1 19
Unmodified.
Index Map: San Francisco Sanborn Insurance Map Atlas, Vol. 1.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213889~5501108

Figure 3.4 23
Cropped, resolution reduced.
Index: San Francisco Sanborn Insurance Map Atlas, Vol. 2.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213950~5501175

Figure 3.5 23
Cropped, resolution reduced.
Index: San Francisco Sanborn Insurance Map Atlas, Vol. 2.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213950~5501175

Figure 3.6 24
Cropped, resolution reduced.
Index: San Francisco Sanborn Insurance Map Atlas, Vol. 1.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213888~5501107

Figure 3.7 24
Cropped, resolution reduced.
Index: San Francisco Sanborn Insurance Map Atlas, Vol. 4.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214074~5501427

Figure 3.8 26
Cropped, resolution reduced.
Index: San Francisco Sanborn Insurance Map Atlas, Vol. 5.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214133~5501488

Figure 3.9 28
Cropped, georectified, resolution reduced, vector overlay.
Index Map: San Francisco Sanborn Insurance Map Atlas, Vol. 2.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213951~5501176

Figure 3.10 28
Cropped, georectified, resolution reduced.
Index Map: San Francisco Sanborn Insurance Map Atlas,
Vol. 1 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213889~5501108
Vol. 2 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213951~5501176
Vol. 3 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214010~5501363
Vol. 4 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214075~5501428
Vol. 5 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214134~5501489
Vol. 6 http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214199~5501554

Figure 3.11 30
Cropped, georectified, resolution reduced, vector overlay.
Index Map: San Francisco Sanborn Insurance Map Atlas, Vol. 2.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213951~5501176

Figure 3.16 36
Cropped, georectified, resolution reduced, vector overlay.
Vol. 1, Page 31-32. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213905~5501124
Vol. 1, Page 41-42. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213910~5501129
Vol. 1, Page 43-44. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213911~5501130
Vol. 1, Page 57-58. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213918~5501137

Figure 3.20 39
Cropped, georectified, resolution reduced, vector overlay.
Vol. 1, Page 41-42. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213910~5501129
Vol. 1, Page 43-44. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213911~5501130

Figure 3.21 40
Cropped, georectified, resolution reduced, vector overlay.
Vol. 1, Page 41-42. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213910~5501129
Vol. 1, Page 43-44. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213911~5501130

Figure 3.26 42
Cropped.
Vol. 1, Page 43-44. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213911~5501130

Figure 4.1 45
Cropped.
Vol. 1, Page 103-104. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213941~5501160

Figure 4.3 47
Cropped.
Vol. 2, Page 143-144. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213961~5501186

Figure 4.4 47
Cropped.
Vol. 2, Page 145-146. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213962~5501187

Figure 4.5 48
Cropped.
Vol. 6, Page 639-640. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214207~5501562

Figures 4.6a and 4.6b 49
Cropped.
Vol. 2, Page 145-146. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213962~5501187

Figure 4.7 50
Cropped.
Vol. 4, Page 425-426. San Francisco
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~214101~5501454

Figure 5.1 61
Cropped.
Vol. 2, Page 175-176. San Francisco.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~213977~5501202

APPENDIX A: R CODE

```
#Load csv file exported from ArcMap
    fishnet <- read.csv("C:/R/f25n.csv", header=T)
#Sort fishnet features by sheet number
    fish <- fishnet[order(fishnet$Sheet),]
# Dataframe "fishdf" with Object ID and sequence number concatenated to sheet no.
    fishDF <- data.frame(
             s = sequence(rle(as.vector(fish$Sheet))$lengths),
             ObjectID = fish$OID,
             uid = paste(fish$Sheet, fishDF$s, sep = "_")
    )
    write.csv(fishDF, file = "C:/R/f25d.csv" )
#Assign unique ID to Address Ranges.
    ar10 <- read.csv("C:/R/addressranges.csv")
    arZ <- ar10[order(ar10$Sheet),]
    arDF <- data.frame(
             s = sequence(rle(as.vector(arZ$Sheet))$lengths),
             jid = arZ$JoinID,
             uid = paste(arZ$Sheet, ardf$s, sep ="_")
     ar10 <- read.csv("C:/R/addressranges.csv")
     arZ <- ar10[order(ar10$Sheet),]
     s = sequence(rle(as.vector(arZ$Sheet))$lengths)
     arDF <- data.frame(
       JoinID = arZ$JoinID,
       uid = paste(arZ$Sheet, s, sep ="_"))
     arM <- merge(arZ, arDF, by="JoinID")
     write.csv (arM, file = "C:/R/arsort.csv")
```