2

Geocoding Fundamentals and Associated Challenges

Claudio Owusu, Yu Lan, Minrui Zheng, Wenwu Tang, and Eric Delmelle

CONTENTS

2.1 Introduction: Geocoding and Geocoding Systems

In the twenty-first century, the ubiquitous usage of smartphones equipped with location-based services has helped millions of individuals in navigating busy traffic or finding available amenities around a particular location. Central to this technological revolution is the process of geocoding, which essentially translates text-based information about locations (address, zip code, names of localities, or even countries) into numerical geographic coordinates (e.g., longitude and latitude). Geocoding uses a spatially explicit reference dataset (e.g., digital road network) to identify the location that best

matches the input address, essentially by comparing and interpolating the address to the range of addresses for each segment of the reference dataset. Each segment contains the locations of the street center and the range of addresses between the street intersections.

Geocoding is generally incorporated in commercial geographic information systems (Bichler and Balchak 2007), where geocoded data can collectively be used for mapping, visualization, and spatial analysis of events. In the past few years, however, the democratization of internet-based mapping services such as Google Maps or MapQuest has facilitated the use of online geocoding services for non-GIS users (Wu et al. 2005; Roongpiboonsopit and Karimi 2010a).

2.1.1 Applications of Geocoding

There is a myriad of domains that have benefitted from geocoding. Geocoding has been a critical element for the delivery of parcels (Jung, Lee, and Chun 2006) and for emergency dispatching management (Derekenaris et al. 2001) where locating the destination in a timely manner is critical. In health studies, geocoding has been used extensively in research with geographic themes such as health disparities (Krieger, Chen et al. 2002; Rehkopf et al. 2006), accessibility to health care (Luo and Qi 2009; Delmelle et al. 2013), disease mapping (Law et al. 2004; Delmelle et al. 2013; Delmelle, Dony et al. 2014), and environmental exposure assessment (Chakraborty and Zandbergen 2007; Zandbergen 2007). In crime analysis, geocoding technology serves as one of the important procedures to obtain data for planning, monitoring, and evaluation of targeted responses to reduce crime in communities (Chainey and Ratcliffe 2013). The process is therefore seen as a means of achieving intelligence-led policing (Ratcliffe 2002; Chainey and Ratcliffe 2013). In addition, geocoding has been used in transportation studies (Park et al. 2011; Qin et al. 2013) for the purpose of planning efficient transportation systems and preventing traffic crashes.

2.1.2 Motivation

In this chapter, we explore geocoding fundamentals, and a myriad of challenging issues that are intimately associated with the procedure, such as spelling sensitivity, accuracy, efficiency, and automation. We also focus on the assessment of the impact of uncertainties related to these geocoding issues on the discovery of spatially explicit patterns. Further, we highlight the significance of geomasking, which is particularly important to preserve confidentiality and minimize the risk of success in reverse geocoding. We then conduct a discussion on web-based geocoding and its benefits, limits, and computational hurdles. We integrate alternative web-based geocoding services together with a cross-validation approach to facilitate the impact assessment of uncertainties associated with geocoding.

In the next section, we briefly describe geocoding fundamentals and illustrate the challenges experienced when attempting to geocode our sample data (see, illustrative dataset in Section 2.1.4). In Section 2.3, we discuss geocoding quality, including sources of errors and the impact of low geocoding quality on spatial analysis. Section 2.4 is devoted to the topic of web-based geocoding, which has recently received a lot of attention. In Section 2.5, we evaluate the merits of two web-based geocoding services as an alternative to commercial geocoding software. Efforts to model and visualize the errors are also presented. In Section 2.6, we address the issue of reverse geocoding, and discuss geomasking and aggregation, two techniques particularly useful to address privacy concerns. We conclude our chapter in Section 2.7 and present avenues for future research.

2.1.3 Contributions

Besides describing and illustrating the process of geocoding, this chapter makes a series of important contributions: (1) strategies to increase the match rate for datasets that include incomplete input addresses (reengineering incomplete addresses in an effort to increase the match rate), (2) use of online geocoding services to cross-validate geocoding results obtained from commercial GIS (and estimating uncertainties in geocoding results), and (3) modeling geocoding errors.

2.1.4 Illustrative Dataset

We use a subset of historical paper records of private water well permits from Gaston County, North Carolina (from 1989 to the present, $n = 7920$) to illustrate the geocoding concepts (subset $n = 285$). Historical records were made available as part of an effort funded by the Centers for Disease Control and Prevention, aiming to establish a public digital database of the county's wells and promote the protection of private well water supplies and quality, ultimately protecting and monitoring a key portion of the county's water supply.

The dataset is particularly salient since historical records pose serious challenges such as (1) incomplete addresses or (2) paper damage. First, a complete address should have all the key components such as house number, street name, street type as well as other directional attributes when possible (e.g., 826 Union Rd, Gastonia, NC 28054). We define an address to be incomplete when any of the key components is not available in the dataset. Second, some permits have faded, making it difficult to transcribe all the address information needed for geocoding. These two problems introduce uncertainties in the datasets.

Private well permit records were scanned and information encoded in a database; each record contains information about the owner of the well, residential location, details of the parcel, ground sketch of the water well

GASTON COUNTY HEALTH DEPARTMENT
ENVIRONMENTAL HEALTH DIVISION
991 W. HUDSON BLVD.
GASTONIA, NC 28052
PHONE (704) 853-5200

Permit void after 12 months

WELL INSTALLATION OR REPAIR PERMIT

Permit Nº 5742

TO BE FILLED IN BY APPLICANT:
Owner or Builder
Date 5-23-93
Mailing Address of Applicant
Phone
Lot Area 0.737Ac Subdivision/Park ASHLEY PLACE Lot 6 Block #
Location DINNER CT.
Signature of Applicant or Authorized Agent

Type Drilled Size 6¼" PVC Depth 200' Casing Depth 64'
Grout Cement Plaster mix Yield 20 GPM Level 34'
Contractor/Driller Suburban Well Telephone

DISTANCES:
1. Water tight sewer line ..... 50'
2. Ground Absorption Sewage System ..... 100'
3. Building Foundations. ..... 50'
4. Any other possible sources of contamination. ..... 100' (Underground storage tanks, animal lots, etc.)

Site Sketch

Tax Book # 14
Tax Map # 85
Tax Parcel # 1
Grid # 12

FIGURE 2.1
A typical private well permit with information of the owner (masked), location, and a sketch of where the well is built.

position, well specification, and the tax location code of the parcel. Figure 2.1 shows an example of a scanned permit. For illustration purposes, we selected a random sample of n = 285 (3%) well samples.

2.2 Geocoding Fundamentals: Input and Reference Data

Accurate reference datasets and valid addresses are the two required inputs for geocoding. *Reference datasets* typically include street network, parcel, and address points data (Zandbergen 2008). In this chapter, we use all three reference datasets and set up hierarchic rules to geocode the illustrative dataset.

Figure 2.2 shows an instance of two different reference datasets (address points and parcel centroid). It can be seen that address points reference data depicts the centroid of the buildings, making it more accurate than the other reference datasets.

For a myriad of reasons such as protecting confidentiality, *addresses* are sometimes made available at different scales, including the street level (Rushton et al. 2006; Goldberg, Wilson, and Knoblock 2007), names of buildings (Davis and Fonseca 2007), closest intersection (Levine and Kim 1998; Park et al. 2011; Delmelle, Zhu et al. 2014), neighborhood level (Casas, Delmelle, and Varela 2010), ZIP code (Krieger, Chen et al. 2002; Krieger, Waterman et al. 2002), textual descriptions of localities (Goldberg and Cockburn 2010), and cities or counties. The scale at which addresses are made available will affect the location of the output feature. For example, addresses at the ZIP code level will be geocoded at the centroid of a postal zip code instead of the residential location.

2.2.1 Geocoding Process

The geocoding process relies on a matching algorithm, which essentially attempts to determine the location of the input address over the range of addresses in the reference dataset. The reference dataset used for the

FIGURE 2.2
Example of two reference datasets: address point in red (most accurate) and parcel centroid (less accurate).

geocoding process determines the technique used in matching the spatial information to geographic coordinates. In most commercial GIS software packages, the matching algorithm is embedded in an address locator. An address locator is a model used to create geometry for textual descriptions representing addresses in the reference data (ESRI Redlands CA, USA). In the United States, a dual range address locator is used when street network is chosen as reference data.

Street geocoding is the most widely used technique due to the readily available TIGER files from the U.S. Census Bureau; here, the algorithm performs a linear interpolation of the input address within the range of address numbers and polarity of the street segment. The process can be decomposed in multiple stages. First, the algorithm attempts to match the street name of the input address with street names from the reference dataset. Next, it will determine the side of the street the address is at, based on whether the address number is even or odd. Third, the correct position of the address is determined after computation of the proportion of the address range associated with the correct side of the street segment. This proportion is then added to the start of the segment to obtain the correct coordinate. Finally, for most commercial GIS software, an optional offset from the street centerline is added. Figure 2.3 shows the interpolated distance (v) and the offset distance (d) used to determine an address along Union Road. The address range along Union Road starts from 101 to 199 on the odd parity side, and from 102 to 200 on the even parity side.

In *parcel geocoding*, the input address is matched to the centroid of the parcel. The returned geographic feature is therefore a point feature with a geographic coordinate (Zandbergen 2008). Although the technique is generally assumed to return more accurate results, it also has been found to introduce positional errors, particularly for a large parcel, since the true address location may not necessarily be at the center of that parcel.

Address point geocoding has been introduced to alleviate this problem. The input address is matched directly to a point feature, which represents the

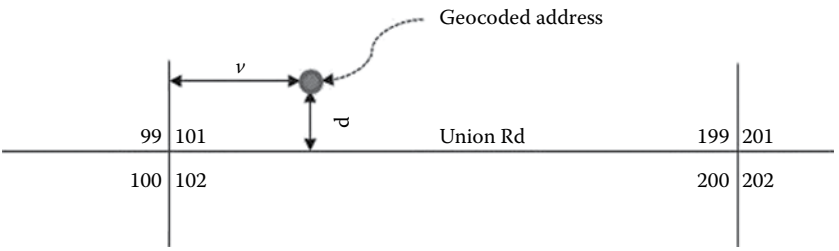

FIGURE 2.3
Interpolation algorithm using address range between the start and end of the street centerline segment for an input address as 117, Union Road.

center of the rooftop of buildings making it more accurate. Emergency calls (e.g., 911 in the United States) use such a geocoding approach.

2.2.2 Match Rate

The success of the geocoding procedure can be determined by its match rate, which is the percentage of records in the input dataset that was correctly geocoded (Zandbergen 2008). A high match rate is often desirable because geocoded results are further used as the sample during spatial investigations (Goldberg, Wilson, and Knoblock 2007; Zimmerman 2008; Ha et al. 2016). Zimmerman (2008) showed that in some instances up to 30% of addresses may need to be excluded if only geocoded records were considered during the analysis. This exclusion of unmatched records reduces the sample size, thereby weakening the generalization of the analytical results due to selection bias and reducing statistical confidence (Zimmerman 2008; Ha et al. 2016).

Geocoding is now a key research methodology and efforts to increase the match rate will help to reduce unmatched addresses that are excluded from the spatial analysis. It is important to note that an increased match rate does not automatically translate into improved geocoding quality. Different strategies exist to increase the geocoding match rate. First, *varying the spelling sensitivity* essentially increases the degree to which a street name is allowed to change. One drawback of this approach is that it will augment the set of potential matches at the cost of potentially selecting a wrong match. The second strategy consists of using *different reference datasets* (McElroy et al. 2003; Yang et al. 2004). A couple of recent studies combined parcel and street network geocoding techniques as a strategy to increase the match rate of the output geographic features (Roongpiboonsopit and Karimi 2010b; Murray et al. 2011; Delmelle et al. 2013). For instance, Delmelle et al. (2013) used different U.S. Census reference datasets to increase the number of geocoded children with birth defects in a study estimating travel impedance to health care centers.

2.2.3 Illustration

In the context of our illustrative dataset, we used multiple datasets from the Gaston County Planning & GIS Department and developed a two-phase geocoding approach as shown below in Figure 2.4. First, during an *automated phase*, different reference datasets (address point, parcel, and street network datasets) are combined in a hierarchical manner into a single composite locator in ArcGIS, a commercial GIS. The rationale to impose a hierarchy among different datasets is to increase the match rate while reducing the odds of positional error. Second, the *improvement phase* consists of using additional datasets such as bacteria test results of the wells and deed records to reengineer the unmatched addresses. Three main strategies are adopted in this

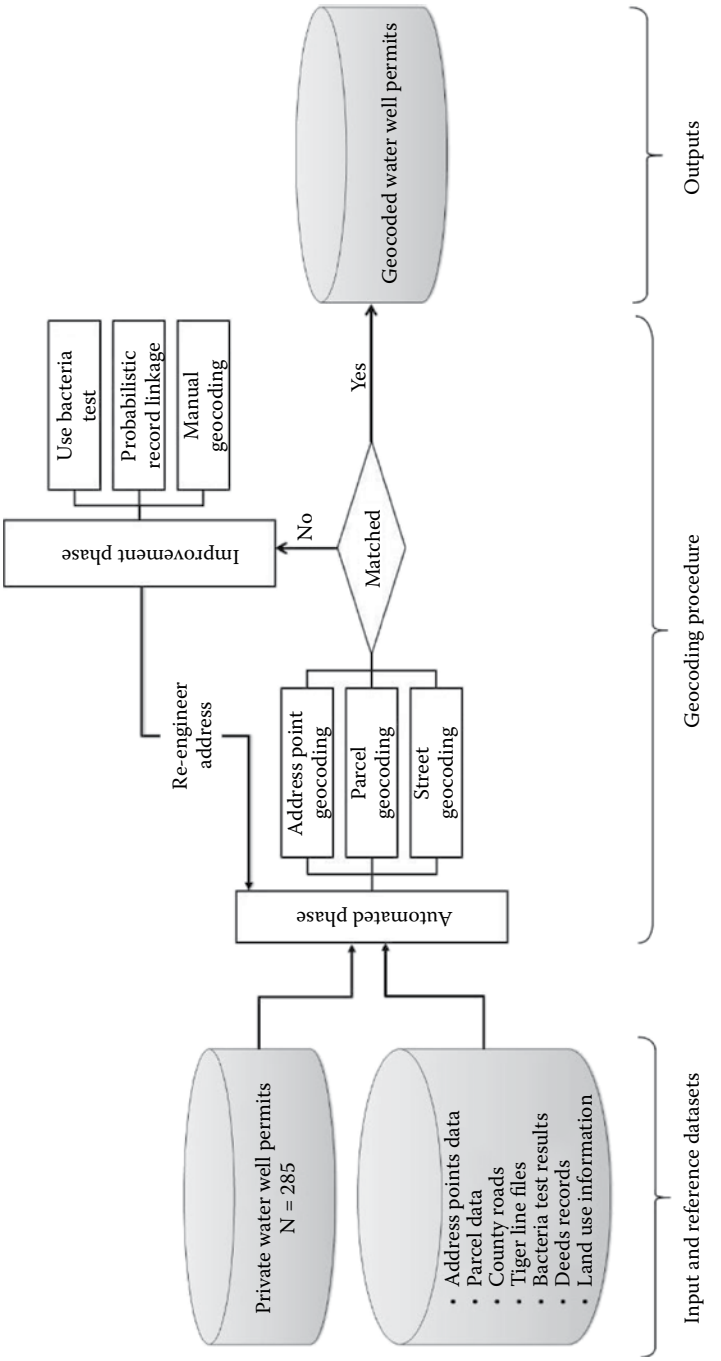

FIGURE 2.4
Flowchart: input and references datasets, geocoding procedure, and outputs.

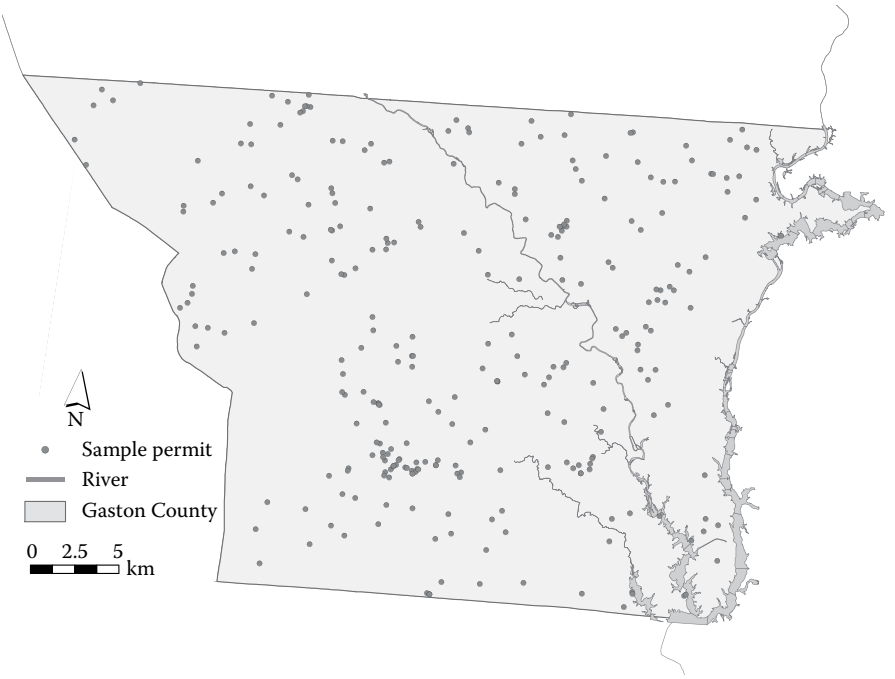

FIGURE 2.5
N = 285 geocoded private well permits in Gaston County, North Carolina.

phase. First, the unique permit number is linked to and cross-checked with the bacteria test results. Second, non-successful records are then subject to a probabilistic record linkage, using information such as tax location codes, name of the well owner, subdivision name, lot size, and block number information. Third, manual geocoding is implemented as the final step, which involves manually interpreting the descriptive address, using additional information such as lot area, lot number, and block number. Once an address has been determined, the commercial GIS attempts to re-geocode using the composite address locator. Figure 2.5 shows the locations of the n = 285 wells that were geocoded with address points reference data.

2.3 Geocoding Quality: Sources of Errors

The success of the geocoding procedure is merely a function of the completeness of the addresses and the quality (i.e., spatial and temporal accuracy, completeness) of the local and regional road network that is used as the reference dataset (O'Reagan 1987; Krieger, Waterman et al. 2002; Zandbergen 2008; Goldberg 2011), and uncertainty with the matching algorithms (Rushton

et al. 2006; Goldberg, Wilson, and Knoblock 2007; Zandbergen 2008, 2011). Over the past decades, however, the accuracy and availability of reference datasets have been improved (Dueker 1974; Werner 1974; Griffin et al. 1990; Boscoe, Ward, and Reynolds 2004).

Although street networks continue to be the most widely used referenced data, the availability of parcel datasets and the introduction of address points from emergency 911 calls in the United States have increased the accuracy and match rate (Zandbergen 2008). The input datasets have expanded from postal addresses (O'Reagan 1987) to include descriptive addresses of locations (Levine and Kim 1998; Davis and Fonseca 2007).

2.3.1 Positional Accuracy

Although the match rate indicates the percentage of addresses that are successfully geocoded, it does not inform us whether the coordinates obtained from the geocoding procedure are the true coordinates. Positional accuracy is a measure of the nearness of the geocoded output from the true location on the ground. Delmelle, Dony et al. (2014) compared geocoded cases of dengue fever in an urban environment of Colombia to locations measured from GPS devices (ground truth). In the context of our illustrative dataset, positional accuracy is estimated by comparing address points that represent the center of the rooftop of buildings with water wells obtained by geocoding from a commercial GIS.

Positional accuracy can be improved by more accurate addresses and reference datasets that are spatially and temporally accurate. Practically taking measurements with GPS devices for the events being investigated can also improve the positional accuracy, but this may be costly and timely ineffective, especially when gathering large datasets. Lastly, using alternative reference datasets for geocoding different environments may minimize the errors. For example, in rural areas where large parcels is the norm, it may be helpful to use aerial photos to generate an address point that better represents the center of the rooftop of the buildings *(if an address point dataset is not already available)* than using parcel or street network datasets.

2.3.2 Impact of Geocoding Quality

Geocoding challenges mentioned in the previous section affect the geocoding quality in terms of match rate and the positional accuracy (O'Reagan 1987; Boscoe, Ward, and Reynolds 2004). Such issues are particularly important in health studies (Bonner et al. 2003; Whitsel et al. 2004; Rushton et al. 2006; Zandbergen 2007; Mazumdar et al. 2008; Chainey and Ratcliffe 2013). Positional accuracy has been found to be critical in studies of environmental exposure as errors can lead to mischaracterization in the risk analysis (Bonner et al. 2003). Positional errors in residential addresses pose a serious challenge for spatial analysis (O'Reagan 1987; Jacquez and

Jacquez 1999; Bonner et al. 2003; Harada and Shimada 2006; Goldberg, Wilson, and Knoblock 2007; Bichler and Balchak 2007; Mazumdar et al. 2008; Zandbergen 2008; Goldberg and Cockburn 2010; Zimmerman and Li 2010; Zimmerman, Li, and Fang 2010), since it may result in (1) underestimation of local risk, (2) misplacement of high-risk areas of a disease, (3) mischaracterization in the analysis of exposure risk, (4) misevaluation of spatial association, and (5) biased evidence for decision makers. When estimating access to health care, positional errors may introduce bias in the estimation of travel impedance, especially for individuals geocoded at the ZIP code for instance.

2.4 Web-Based Geocoding

The costs to prepare reference data and standardize addresses can be prohibitive when using commercial GIS software. With the rapid development of cyber-enabled technology, a myriad of web-based providers (such as Google Maps, Bing Maps, and MapQuest, to name a few) have made the process of geocoding more accessible and faster through their online geocoding services (Roongpiboonsopit and Karimi 2010a). The preparation and maintenance of reference data, address standardization, and algorithm implementation and update for geocoding are hidden in these online services (accessible as APIs). Online geocoders typically use street network data that are more up to date, which is likely to result in lower positional errors. Online geocoders, however, have limits on the number of records that can be processed (e.g., 2500 for Google Maps and Bing Maps on a daily basis, 15,000 per month for MapQuest), suffer from a lack of transparency about the geocoding algorithm (including address interpretation) and lack of metadata on the update of reference data (an issue that may vary spatially). Another important issue is that the use of online geocoders may raise important ethical issues such as confidentiality since addresses are uploaded to remote servers. In the United States, this may violate the Health Insurance Portability and Accountability Act, which protects individuals' medical records and other personal health information (DeLuca and Kanaroglou 2015; Kirby, Delmelle, and Eberth 2017; Mak et al. 2012). Different strategies exist to circumvent this issue, such as geocoding at a coarser scale, or bundle the batch of addresses to be geocoded with random addresses (Gittler 2007; Goldberg 2008).

When using geocoding APIs, users or developers need to call functions and obtain authentication from corresponding online geocoding providers. Then these online geocoding services will use their own algorithms to calculate the coordinates that will be returned to the user (e.g., in pure text or XML-based formats). In most occasion, users can type the address that they want to geocode and click a button, to display the results on the map (i.e., in

an interactive manner). Besides being available to non-GIS users, web-based geocoding systems are particularly helpful to evaluate the accuracy of the geocoding results obtained from commercial GIS software, such as ArcGIS. The accuracy evaluation is typically conducted by comparing the geocoded coordinates (Duncan et al. 2011).

2.5 Using Web-Based Geocoding Services for Cross Validation

In this study, we follow an approach similar to Duncan et al. (2011) that is based on online geocoding services (Google and MapQuest here) to validate the geocoding results from those obtained by a commercial GIS (ArcGIS). Each address record may exhibit differences in the coordinates among these geocoding options; the distance between coordinates from online geocoding services and ArcGIS-based results (referred to as error distance) is calculated. We estimate the error for the n = 285 geocoded samples. The distances are grouped into different "deviation categories" (<50, <100, <150, <200, <250, <300, and >300 m). For each category, we report the match rate, defined as "the percentage of the successfully geocoded records in relation to the total number of records originally subjected to the geocoding process, regardless of the positional accuracy" (Kounadi et al. 2013). Table 2.1 shows the percentage of geocoding results located in certain deviation categories according to different web-based geocoding services.

Generally, Google has a higher match rate and its geocoding results are likely to be closer to the ones obtained from ArcGIS. Depending on the purpose of the study, strict error thresholds may be necessary. In the case of studying exposure to highway pollution, a difference of 300 m may be very significant and bias the analysis (Zandbergen 2007). Further, greater distance errors are not uncommon in rural areas (Zimmerman and Li 2010). In the following section, we will analyze and model our web-based geocoding results comparing with true coordinates (in this case, we consider results of ArcGIS obtained using address point geocoding as the true coordinates).

TABLE 2.1

Variation in Match Rate for Two Online Geocoding Systems with Deviation Categories

Buffer (m)	50	100	150	200	250	300	>300
Google (%)	70.18	85.96	89.12	90.88	92.28	93.33	6.67
MapQuest (%)	62.46	82.11	89.47	90.88	92.63	92.98	7.02

2.5.1 Modeling Geocoding Error

In this study, we compare results of online geocoding services from Google and MapQuest to the ones obtained using ArcGIS. For this comparison purpose, we constructed error modeling, which consists of the following steps: (1) acquiring results from web-based geocoding services, (2) convert latitude and longitude (WGS84) into XY coordinates, (3) calculate the Euclidean error distance (in meter) between results of ArcGIS and web-based geocoding results, and (4) compare geocoding results in terms of the empirical distribution of error distance and fitted error model based on, for example, distance-decayed functions.

The error distance can be visualized in different ways. The error is represented in its simplest form as a line connecting the spatial locations of the geocoded well with the commercial solver and the online geocoders (yellow for MapQuest, red for Google) as shown in Figure 2.6a–d.

Figure 2.6e illustrates the error distance between the commercial geocoder and the Google geocoder, where a larger symbol denotes a greater error distance. Figure 2.6f compares the error distance among online providers. In pink and purple colored regions, the error distance is much lower when using Google than MapQuest, while the reverse is true for green colored regions. Figure 2.6e–f clearly suggests the presence of a spatial pattern in terms of error distance.

Table 2.2 and Figures 2.7 and 2.8 illustrate the empirical histogram and probabilistic distributions of error distance for the two web-based geocoding services (bin size: 10 m). About 95% of the Google-based results (with a median of 26.59 m) fall within a distance that is less than 250 m. MapQuest-based geocoding results (median: 39.28 m) have a longer error distance (about 360 m) than those of Google (250 m) with respect to a 95% threshold. In addition, the mode of Google-based error distance is within 10 m (covering 23.83% of the data), compared to MapQuest-based results with a mode around 30–40 m (25.62%). For the error modeling, we fitted the histograms of error distance using Pareto functions (see Morrill and Pitts 1967). Table 2.3 summarizes model fitting results. The goodness-of-fit of the error model for Google-based geocoding results (up to 88.97% of the variance explained) is much higher than that for MapQuest-based results (only 74.01% of the variance explained).

Results from both empirical distribution and the fitted error models suggest that Google's online geocoding service generally outperforms MapQuest for the geocoding task in our study area. This finding is consistent with what has been reported in the literature. For example, Roongpiboonsopit and Karimi (2010a) compared the quality of five online geocoding services (including Google and MapQuest), and found that Google provided a shorter error distance than MapQuest. Results from other relevant studies by Cui (2013), Chow et al. (2016), and Karimi et al. (2011) also indicate

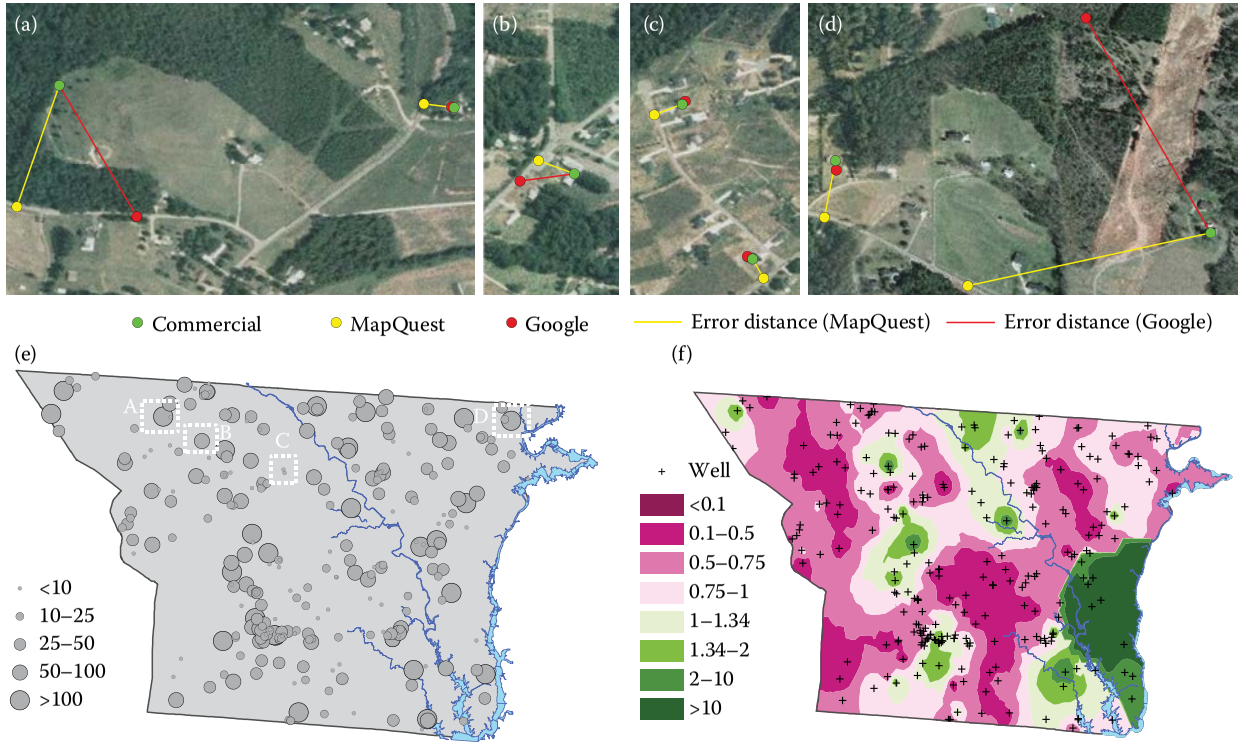

FIGURE 2.6
Visualization of a geocoding error. In a–d, geocoded addresses from the ArcGIS commercial geocoder are depicted in green, while the results from Google and MapQuest are colored in red and yellow, respectively. In e, circles of increasing sizes represent a higher error distance (units: meters). In f, is an interpolated map showing the ratio of distance errors between Google and MapQuest (in green areas, MapQuest has a lower error distance).

TABLE 2.2

Frequency and Probability of the Error Distance of Online Geocoding Services

	Google			MapQuest		
Bin	**Frequency**	**Percent (%)**	**Cumulative (%)**	**Frequency**	**Percent (%)**	**Cumulative (%)**
10	66	23.83	23.83	0	0.00	0.00
20	52	18.77	42.60	9	3.20	3.20
30	36	13.00	55.60	64	22.78	25.98
40	26	9.39	64.98	72	25.62	51.60
50	21	7.58	72.56	33	11.74	63.35
60	16	5.78	78.34	14	4.98	68.33
70	13	4.69	83.03	12	4.27	72.60
80	5	1.81	84.84	16	5.69	78.29
90	4	1.44	86.28	11	3.91	82.21
100	6	2.17	88.45	2	0.71	82.92
110	1	0.36	88.81	7	2.49	85.41
120	3	1.08	89.89	9	3.20	88.61
130	3	1.08	90.97	3	1.07	89.68
140	1	0.36	91.34	2	0.71	90.39
150	1	0.36	91.70	0	0.00	90.39
160	0	0.00	91.70	0	0.00	90.39
170	0	0.00	91.70	2	0.71	91.10
180	2	0.72	92.42	2	0.71	91.81
190	2	0.72	93.14	0	0.00	91.81
200	1	0.36	93.50	0	0.00	91.81
210	0	0.00	93.50	2	0.71	92.53
220	1	0.36	93.86	1	0.36	92.88
230	1	0.36	94.22	1	0.36	93.24
240	1	0.36	94.58	0	0.00	93.24
250	1	0.36	94.95	2	0.71	93.95
260	1	0.36	95.31	0	0.00	93.95
270	1	0.36	95.67	0	0.00	93.95
280	0	0.00	95.67	0	0.00	93.95
290	1	0.36	96.03	0	0.00	93.95
300	0	0.00	96.03	0	0.00	93.95
310	0	0.00	96.03	1	0.36	94.31
320	0	0.00	96.03	0	0.00	94.31
330	0	0.00	96.03	0	0.00	94.31
340	1	0.36	96.39	1	0.36	94.66
350	1	0.36	96.75	0	0.00	94.66
360	0	0.00	96.75	0	0.00	94.66
More	9	3.25	100.00	15	5.34	100.00

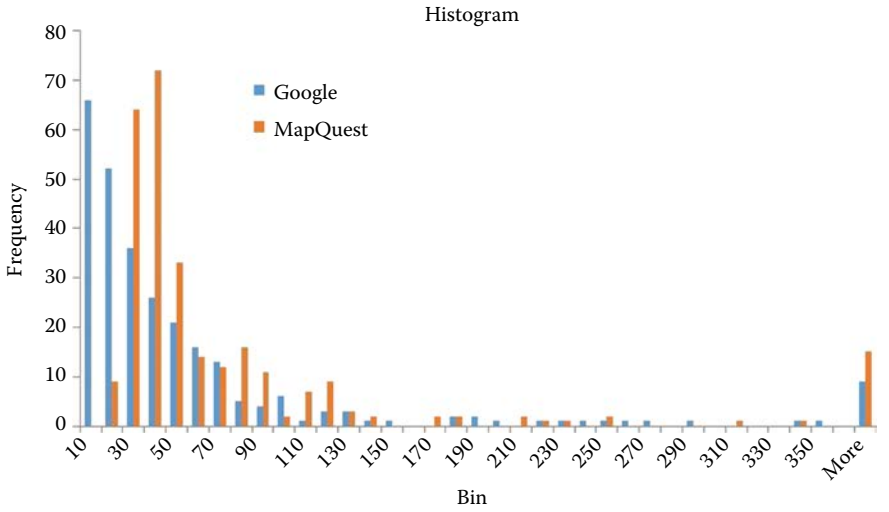

FIGURE 2.7
Histogram of error distance of online geocoding services (bin size: 10 m).

that Google's geocoding service can achieve rates that are 91.5%, 100%, and 93.64%, respectively, which are higher than other online geocoding services (e.g., MapQuest, Bing, and Geocoder.us). While multiple factors may contribute to geocoding errors, frequent update of reference data by Google may explain its high geocoding accuracy.

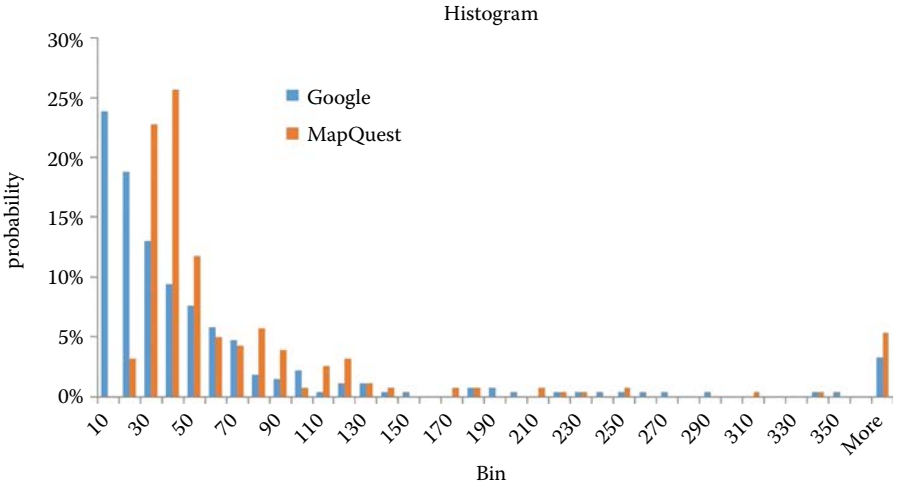

FIGURE 2.8
Empirical probabilistic distribution of error distance of online geocoding services.

TABLE 2.3

Fitted Modeling Results Based on Error Distance

Geocoding Services	Fitted Models	R^2
Google	$Y = 4245\ D^{-1.508}$	0.8897
MapQuest	$Y = 6431.1\ D^{-1.523}$	0.7401

Note: Y: frequency; D: distance

2.6 Reverse Geocoding, Geomasking, and Aggregation

Although geocoded data result in a great opportunity to develop better analytical solutions, there exist some important concerns, especially in the context of epidemiology to protect privacy needs. At the core of the issue is the thread of *reverse geocoding*, which essentially determines the address based on geographic coordinates. Using a published map of geocoded records and overlaying with other layers of spatial information (such as parcel and street layers), the approximate address of the geocoded record can be traced back (Curtis, Mills, and Leitner 2006).

Several *geomasking* techniques and aggregation strategies have been developed to conceal the true identity of geocoded records and minimize the risk of success in reverse geocoding. Geomasking (Armstrong, Rushton, and Zimmerman 1999) is a spatial statistical technique used to introduce uncertainty (i.e., noise) into the spatial locations of geocoded records, which has implications for the quality of further spatial analysis (e.g., cluster detection). The main mechanism behind geomasking consists of perturbing the spatial location of a geocoded record, typically in a random distance and along a random direction. Other strategies have been developed, such as the donut geomasking method (Hampton et al. 2010) where geocoded records are moved within a random direction and within certain distance bounds. These distance bounds can be tighter in urban areas and looser in rural regions where the spacing between residences is much greater.

Finally, geocoded records can be *spatially aggregated* into census units, where all the data are moved to the geographic centroid of the unit (Tellman et al. 2010). The choice of the unit is a function of the number of cases and the population within that unit.

Despite their ability to preserve some confidentiality, *geomasking* and *aggregation* methods have some substantial limitations, such as (1) reducing the level of precision, (2) introducing statistical bias into the results, (3) blurring meaningful variations in data, and (4) weakening clustering detection. Current research attempts to find optimal geomasking strategies to preserve the spatial pattern of geocoded records while maintaining privacy.

2.7 Conclusions

In this chapter, we have discussed fundamentals of geocoding, which we illustrated on a dataset of private well addresses in Gaston County, North Carolina. We compared spatial locations estimated by a commercial geocoder to the ones obtained by two popular online providers, Google and MapQuest. We found that in most cases, coordinates from online geocoders were relatively close (26.59 m for Google and 39.28 m for MapQuest) to the ones obtained by the commercial geocoder. Generally, MapQuest geocoder yielded greater error than Google geocoder.

There remains a suite of challenges in geocoding. First, online web services provide an alternative for geocoding, but further work is needed to tackle the issue of transparency on reference datasets and geocoding algorithms. An open geocoding standard and platform may be of help. Second, massive data are increasingly available, and how to efficiently and effectively geocode these datasets (say, millions or billions of addresses) poses a big data challenge. Cyberinfrastructure-enabled high-performance computing holds promises in resolving the big data challenge. Third, the evaluation of geocoding accuracy, particularly for handling massive data, remains as a challenge. Spatial or spatiotemporal statistics may provide support for evaluating the robustness of the geocoding process.

References

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

Bichler, G., and S. Balchak. 2007. Address matching bias: Ignorance is not bliss. *Policing: An International Journal of Police Strategies & Management* 30 (1): 32–60.

Bonner, M. R., D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim. 2003. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 14 (4): 408–412.

Boscoe, F. P., M. H. Ward, and P. Reynolds. 2004. Current practices in spatial analysis of cancer data: Data characteristics and data sources for geographic studies of cancer. *International Journal of Health Geographics* 3 (1): 28.

Casas, I., E. Delmelle, and A. Varela. 2010. A space-time approach to diffusion of health service provision information. *International Regional Science Review* 33 (2): 134–156.

Chainey, S., and J. Ratcliffe. 2013. In *GIS and Crime Mapping*. London, UK: John Wiley & Sons, pp. 1–448.

Chakraborty, J., and P. A. Zandbergen. 2007. Children at risk: Measuring racial/ethnic disparities in potential exposure to air pollution at school and home. *Journal of Epidemiology and Community Health* 61 (12): 1074–1079.

Chow, T. E., N. Dede-Bamfo, and K. R. Dahal. 2016. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. *Annals of GIS* 22 (1): 29–42.

Cui, Y. 2013. A systematic approach to evaluate and validate the spatial accuracy of farmers market locations using multi-geocoding services. *Applied Geography* 41: 87–95.

Curtis, A. J., J. W. Mills, and M. Leitner. 2006. Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics* 5 (1): 44.

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

Delmelle, E., C. Dony, I. Casas, M. Jia, and W. Tang. 2014. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science* 28 (5): 1107–1127.

Delmelle, E. M., C. H. Cassell, C. Dony, E. Radcliff, J. P. Tanner, C. Siffel, and R. S. Kirby. 2013. Modeling travel impedance to medical care for children with birth defects using geographic information systems. *Birth Defects Research Part A: Clinical and Molecular Teratology* 97 (10): 673–684.

Delmelle, E. M., H. Zhu, W. Tang, and I. Casas. 2014. A web-based geospatial toolkit for the monitoring of dengue fever. *Applied Geography* 52: 144–152.

DeLuca, P., and P. S. Kanaroglou. 2015. An assessment of online geocoding services for health research in a mid-sized Canadian city. In *Spatial Analysis in Health Geography*, edited by P. Kanaroglou, E. Delmelle and A. Paez, New York, NY: Ashgate Publishing, pp. 31–46.

Derekenaris, G., J. Garofalakis, C. Makris, J. Prentzas, S. Sioutas, and A. Tsakalidis. 2001. Integrating GIS, GPS and GSM technologies for the effective management of ambulances. *Computers, Environment and Urban Systems* 25 (3): 267–278.

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

Duncan, D. T., M. C. Castro, J. C. Blossom, G. G. Bennett, and S. L. Gortmaker. 2011. Evaluation of the positional difference between two common geocoding methods. *Geospatial Health* 5(2): 265–273.

Gittler, J. 2007. Cancer registry data and geocoding. In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, edited by G. Rushton et al., Boca Raton, FL: CRC Press, pp. 195–223.

Goldberg, D. 2008. Privacy and confidentiality. In *A Geocoding Best Practices Guide*, edited by R. Borchers et al., Springfield, IL: North American Association of Central Cancer Registries, Inc. pp. 183–187.

Goldberg, D. W. 2011. Advances in geocoding research and practice. *Transactions in GIS* 15 (6): 727–733.

Goldberg, D. W., and M. G. Cockburn. 2010. Improving geocode accuracy with candidate selection criteria. *Transactions in GIS* 14 (s1): 149–176.

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

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

Ha, S., H. Hu, L. Mao, D. Roussos-Ross, J. Roth, and X. Xu. 2016. Potential selection bias associated with using geocoded birth records for epidemiologic research. *Annals of Epidemiology* 26 (3): 204–211.

Hampton, K. H., M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller. 2010. Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, 172 (9): 1062–1069.

Harada, Y., and T. Shimada. 2006. Examining the impact of the precision of address geocoding on estimated density of crime locations. *Computers & Geosciences* 32 (8): 1096–1107.

Jacquez, G. M., and J. A. Jacquez. 1999. Disease clustering for uncertain locations. *Disease Mapping and Risk Assessment for Public Health Decision Making*, edited by A. Lawson et al., London, UK: Jonh Wiley & Sons, pp. 151–168.

Jung, H., K. Lee, and W. Chun. 2006. Integration of GIS, GPS, and optimization technologies for the effective control of parcel delivery service. *Computers & Industrial Engineering* 51 (1): 154–162.

Karimi, H. A., M. H. Sharker, and D. Roongpiboonsopit. 2011. Geocoding recommender: An algorithm to recommend optimal online geocoding services for applications. *Transactions in GIS* 15 (6): 869–886.

Kirby, R. S., E. Delmelle, and J. M. Eberth. 2017. Advances in spatial epidemiology and geographic information systems. *Annals of Epidemiology* 27 (1): 1–9.

Kounadi, O., T. J. Lampoltshammer, M. Leitner, and T. Heistracher. 2013. Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science* 40 (2): 140–153.

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

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

Law, D. G., M. L. Serre, G. Christakos, P. A. Leone, and W. C. Miller. 2004. Spatial analysis and mapping of sexually transmitted diseases to optimise intervention and prevention strategies. *Sexually Transmitted Infections* 80 (4): 294–299.

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

Luo, W., and Y. Qi. 2009. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & Place* 15 (4): 1100–1107.

Mak, S., D. T. Duncan, M. C. Castro, and J. C. Blossom. 2012. Geocoding-protected health information using online services may compromise patient privacy-Comments on "Evaluation of the positional difference between two common geocoding methods" by Duncan et al.—Response. *Geospatial Health* 6 (2): 157–159.

Mazumdar, S., G. Rushton, B. J. Smith, D. L. Zimmerman, and K. J. Donham. 2008. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics* 7 (1): 13.

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

Morrill, R. L., and F. R. Pitts. 1967. Marriage, migration, and the mean information field: A study in uniqueness and generality. *Annals of the Association of American Geographers* 57 (2): 401–422.

Murray, A. T., T. H. Grubesic, R. Wei, and E. A. Mack. 2011. A hybrid geocoding methodology for spatio-temporal data. *Transactions in GIS* 15 (6): 795–809.

O'Reagan, R. T. and Saalfeld, A. 1987. Geocoding theory and practice at the bureau of the census. *Statistical Research Report Census/SRD/RR-87/29*, US. Census Bureau, Washington, DC. pp. 1–14.

Park, S. H., J. M. Bigham, S.-Y. Kho, S. Kang, and D.-K. Kim. 2011. Geocoding vehicle collisions on Korean expressways based on postmile referencing. *KSCE Journal of Civil Engineering* 15 (8): 1435–1441.

Qin, X., S. Parker, Y. Liu, A. J. Graettinger, and S. Forde. 2013. Intelligent geocoding system to locate traffic crashes. *Accident Analysis & Prevention* 50: 1034–1041.

Ratcliffe, J. 2002. Intelligence-led policing and the problems of turning rhetoric into practice. *Policing & Society* 12 (1): 53–66.

Rehkopf, D. H., L. T. Haughton, J. T. Chen, P. D. Waterman, S. Subramanian, and N. Krieger. 2006. Monitoring socioeconomic disparities in death: Comparing individual-level education and area-based socioeconomic measures. *American Journal of Public Health* 96 (12): 2135–2138.

Roongpiboonsopit, D., and H. A. Karimi. 2010a. Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science* 24 (7): 1081–1100.

Roongpiboonsopit, D., and H. A. Karimi. 2010b. Quality assessment of online street and rooftop geocoding services. *Cartography and Geographic Information Science* 37 (4): 301–318.

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

Tellman, N., E. R. Litt, C. Knapp, A. Eagan, J. Cheng, and J. Lewis Jr. 2010. The effects of the Health Insurance Portability and Accountability Act privacy rule on influenza research using geographical information systems. *Geospatial Health* 5 (1): 3–9.

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

Whitsel, E. A., K. M. Rose, J. L. Wood, A. C. Henley, D. Liao, and G. Heiss. 2004. Accuracy and repeatability of commercial geocoding. *American Journal of Epidemiology* 160 (10): 1023–1029.

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

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

Zandbergen, P. A. 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health* 7 (1): 37.

Zandbergen, P. A. 2008. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32 (3): 214–232.

Zandbergen, P. A. 2011. Influence of street reference data on geocoding quality. *Geocarto International* 26 (1): 35–47.

Zimmerman, D. L. 2008. Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics* 64 (1): 262–270.

Zimmerman, D. L., and J. Li. 2010. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics* 9 (1): 10.

Zimmerman, D. L., J. Li, and X. Fang. 2010. Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in Medicine* 29 (9): 1025–1036.