From Text to Geographic Coordinates: The Current State of Geocoding

Daniel W. Goldberg, John P. Wilson, and Craig A. Knoblock

Abstract: *This article presents a survey of the state of the art in geocoding practices through a cross-disciplinary historical review of existing literature. We explore the evolving concept of geocoding and the fundamental components of the process. Frequently encountered sources of error and uncertainty are discussed as well as existing measures used to quantify them. An examination of common pitfalls and persistent challenges in the geocoding process is presented, and the traditional methods for overcoming them are described.*

INTRODUCTION

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

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

While significantly improving the usability, reliability, and accuracy of the geocoding process, these developments have brought with them a host of issues that a potential user must recognize and be prepared to contend with. Specific issues include the assumptions made during the interpolation process (Dearwent et al. 2001, Karimi et al. 2004), the underlying accuracy of the reference dataset (Gatrell 1989, Block 1995, Drummond 1995, Martin and Higgs 1996, Chung et al. 2004), the uncertainty in the matching algorithm (O'Reagan and Saalfeld 1987, Jaro 1984), and the choice of areal unit geocoded to (Krieger 1992, Geronimus et al. 1995, Geronimus and Bound 1998, Krieger et al. 2002a, 2003). These topics have received considerable research in recent times, and a great deal of literature is available. This article will survey the field of geocoding through a cross-disciplinary study of the geocoding literature focusing foremost on the technical aspects of the process. The changing concept of geocoding will be described, and the fundamental components of the geocoder will be outlined. Potential sources of error in the geocoding process will be explored, and particularly difficult geocoding scenarios requiring further research will be highlighted. The primary contributions of this article will be to inform the reader of the state of the art in geocoding through a discussion of its evolution over time and to warn of potentially sticky situations that can arise in the geocoding process if one is not aware of how one's decisions and assumptions can affect the geocoded results. This work should be seen as distinct from the recent work published by Rushton et al. (2006), which also offers a review of the geocoding process, but is focused on its application to health research, in particular cancer studies. Their work takes a narrow and limited view of geocoding and does not delve so deeply into the evolution or technical aspects of the geocoding process as does that presented here. As such, this paper can be seen as a more comprehensive, technically

targeted, broadly visioned journey through the geocoding process and should be used as a companion article to field-specific reviews such as that of Rushton et al. (2006).

THE CONCEPT OF GEOCODING

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

These fundamental shifts in geocoding attitudes and opportunities can be traced directly to the technological advances made to the underlying reference datasets on which they are based. The early attempts at geocoding were hindered by the lack of digital geographies to use in the assignment of codes, and were limited by their use of flat text-based files. This resulted in low-resolution nongeographic output, turning addresses and building names into the census block to which they belonged. The development of true digital geographies in the form of products such as the U.S. Census Bureau's Dual Independent Map Encoding (DIME) files enabled the assignment of true geographic codes, but their structure limited the processing that could be applied to derive the output. The introduction of the vector-based geographic datasets such as the U.S. Census Bureau's Topographically Integrated Geographic Encoding and Referencing (TIGER) (U.S. Census Bureau 2006) database have enabled new generations of geocoding algorithms to approximate representations for the geographic output using interpolation-based approaches, greatly increasing the resolution of the geographic output (Dueker 1974, O'Reagan and Saalfeld 1987, Martin 1998, Ratcliffe 2001, Nicoara 2005). Taking this a step further, the creation of precompiled geocoded national address registers such as the ADDRESS-POINT (Ordnance Survey 2006) and Geocoded National Address File (G-NAF) (Paull 2003) databases in the United Kingdom and Australia, respectively, have facilitated highly precise geocoding capabilities at national scales (Higgs and Martin 1995, Martin 1998, Ratcliffe 2001, Churches et al. 2002, Higgs and Richards 2002, Christen et al. 2004, Christen and Churches 2005, Murphy and Armitage 2005). Furthermore, the emergence of high-resolution digital parcel and property boundary files may enable even more accurate digital geographic results to be returned (Dueker 1974, Olligschlaeger 1998, Dearwent et al. 2001, Ratcliffe 2001, Rushton et al. 2006), but these developments are pushing the limits of what form the output of geocoding should take. Likewise, the development of multiresolution gazetteers defining geographic footprints for named geographic places such as the Alexandria Digital Library Gazetteer (Frew et al. 1998, Hill and Zheng 1999, Hill et al. 1999, Hill 2000) are pushing the limits of what type of geographic features can have geographic codes assigned to them (Davis et al. 2003, United Nations Economic Commission 2005), as well as the role of the geocoder in the larger geospatial information-processing context. The proliferation of a variety of diverse types of locational addressing systems throughout the world precludes a "one size fits all" geocoding strategy that will work in all cases (Fonda-Bonardi 1994, Lind 2001, Davis et al. 2003, Walls 2003, United Nations Economic Commission 2005).

The result of this evolution is a somewhat "fuzzy" concept of geocoding, tailored to the specific requirements and data availability of the person performing the geocoding. For example, almost everyone involved in or using geocoding today would agree that turning a postal address into a geographic point is most certainly included in the set of geocoding operations. Likewise, they would probably agree that turning a portion of the postal address such as the post code (zip code) into a geographic point or polygon is also part of the geocoding process. However, continuing this line of reasoning presents a slippery slope because a series of fundamental questions arise. What should the point returned as representative of the postal code be? Should it be the center of mass (centroid)? Should it be weighted by the population distribution? Furthermore, if the digital boundary of the postal code is available, why not return it instead of just a single point? Questions such as these are just the beginning. If the postal code can be geocoded, can the city be as well? If so, what is the difference between the geocoder returning a geographic representation of the city and the gazetteer doing the same? And if they are, in fact, performing the same operation, why is it commonly understood that a gazetteer can provide geographic representations for a wide variety of geographic features such as rivers, mountains, and shorelines, while these are seldom thought of as candidates for the geocoding process? We can see through this discussion that the term *geocoding* can mean different things to different people, and their perception will be based on their primary experience or usage with a particular geocoding tool. To some, "geocoding" is synonymous with "address matching" (e.g., Drummond 1995, Vine et al. 1998, Bonner et al. 2003), highlighting its prevalent use of transforming postal addresses into geographic representations (Drummond 1995, 250). For others, "geocoding" is understood to produce a valid geographic output, but its input is not necessarily limited to simple postal addresses

(e.g., Levine and Kim 1998), and still further distinctions can be drawn between the two terms (Johnson 1998a, 25). Taken literally, geocoding means "to assign a geographic code." This definition stems from the two root words: *geo,* from the Latin for earth, and *coding,* defined as "applying a rule for converting a piece of information into another" (similar to that defined early on in the geocoding literature [Dueker 1974, 320]). Notice that this literal definition does not imply nor constrain in any way the input to the geocoding system, the processes or data sources used to assign the geographic code, or even what the geographic code returned as output must be. It is precisely this relaxation of formal constraints on the geocoding process that has allowed it to mature and prosper to the many forms that we use today, and that will in turn drive the technological advances of tomorrow.

GEOCODING FUNDAMENTALS

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

As previously noted, however, address data are not the only type of locational data that can or should be geocoded. Even the earliest geocoding systems of the U.S. Census accounted for the geocoding of named buildings (O'Reagan and Saalfeld 1987), but the task of associating geocodes with geographic features other than addresses is most commonly associated with the services provided by a gazetteer (Hill 2000). The problem with this, though, is that a gazetteer typically does not contain the functionality to generate the geocodes that it returns, instead acting as a storage mechanism after the geocodes have already been determined using other methods. As such, the geocoder is commonly employed to produce the geocodes for features in the gazetteer that are address-based, emphasizing the crucial connection between the two components as part of a larger spatial query and analysis framework. This situation is displayed in Figure 1, where the geocoder is shown to be one of many possible sources of footprint data for a gazetteer, with itself being composed of several data sources.

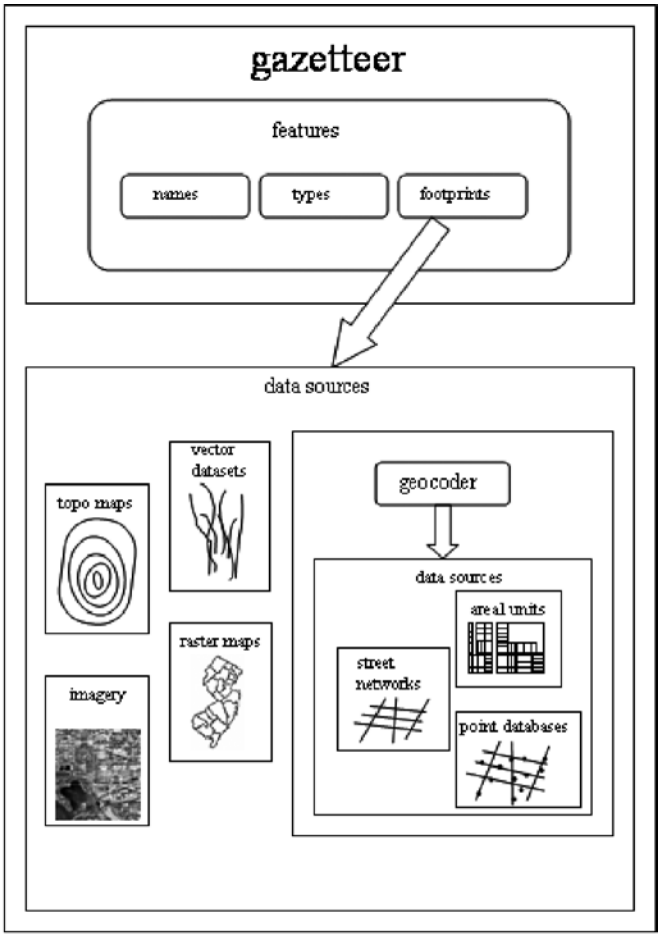

Figure 1. Relationship between the gazetteer and geocoder

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

The processing algorithm determines the appropriate geographic code to return for a particular input based on the values of its attributes and the values of attributes in the reference dataset. This is by far the most complicated portion of the geocoding process in which the most research has been invested. The key topics involved in the process include the standardization and normalization of the input into a format and syntax compatible with that of the reference dataset (Johnson 1998b, Churches et al. 2002, Laender et al. 2005, Nicoara 2005), the matching algorithm

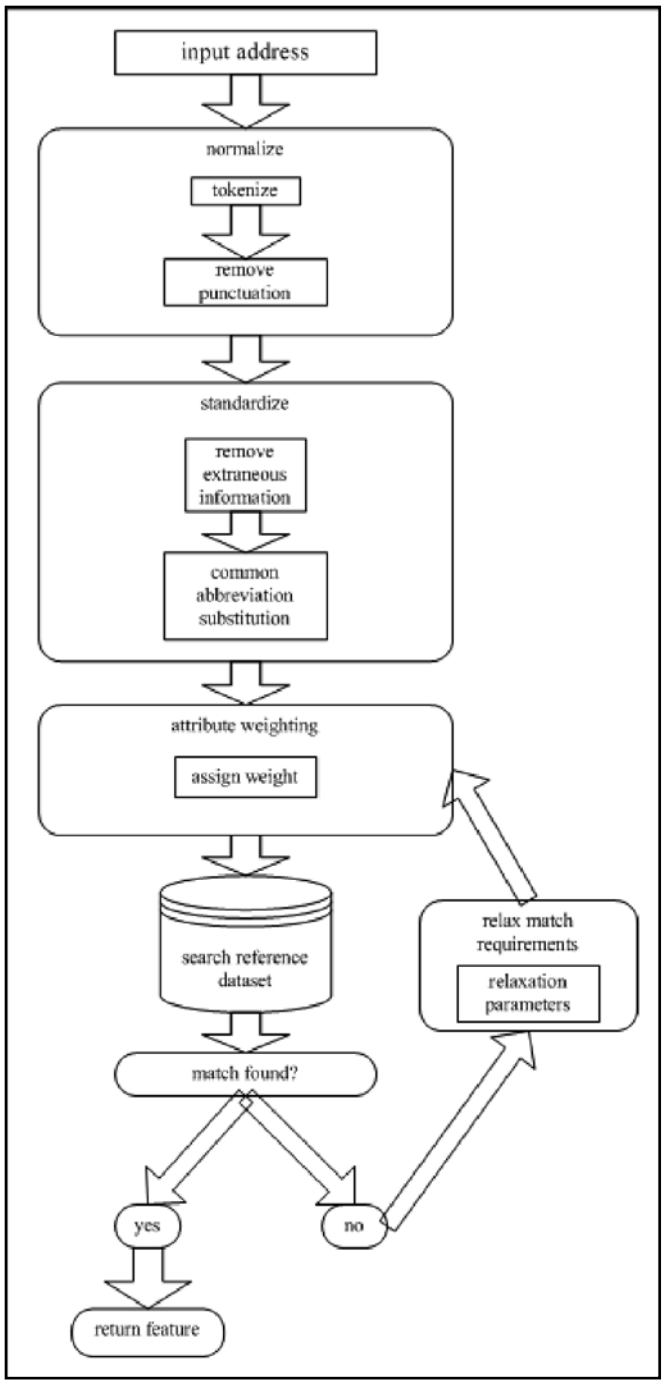

Figure 2. Schematic of deterministic address matching with attribute relaxation

that picks the best feature in the reference dataset (Drummond 1995, Vine et al. 1998, Davis et al. 2003, Bakshi et al. 2004), and the final geocode generation mechanism that determines what to return based on the reference feature selected as the best match (Drummond 1995, Levine and Kim 1998, Ratcliffe 2001, Cayo and Talbot 2003, Davis et al. 2003). Figure 2 shows a schematic diagram of how a simple deterministic processing algorithm could proceed using standardization, normalization, and attribute relaxation. The standardization and normalization process can vary in complexity from simple token parsing with lookup tables for standardizing abbreviations to advanced probabilistic methods using machine learning techniques such as hidden Markov models that can handle attribute misspellings and misplacements (O'Reagan and Saalfeld 1987, Fulcomer et al. 1998, Churches et al. 2002, Christen et al. 2004, Yang et al. 2004, Christen and Churches 2005, Nicoara 2005). In general, the key role performed in this step is to determine what each piece of the input is and to turn each into versions consistent with those in the reference dataset.

Once the input has been sufficiently massaged to be compatible with the reference dataset, the matching process picks the best candidate to be used to derive the final output. Tricks such as word stemming, using Soundex, and relaxing the requirement of matching all attributes can be used to improve the probability of finding a match in the reference dataset (O'Reagan and Saalfeld 1987, Drummond 1995, Fulcomer et al. 1998, Johnson 1998a, Levine and Kim 1998, Gregorio et al. 1999, Boscoe et al. 2002, Churches et al. 2002, Beal 2003, Christen et al. 2004, Yang et al. 2004, Christen and Churches 2005, Nicoara 2005). Here the issue may arise that zero, one, or more than one reference features can be the best possible match. In the case of one match, the algorithm will use it to determine a geocode. In the case of zero, the matching algorithm may prompt the user for more information, attempt to geocode at a lower resolution with additional datasets, or try to find additional information in other datasets to enable a match (Laender et al. 2005). Likewise, in the case of multiple matches, the algorithm may prompt the user to determine the appropriate one or consult additional datasets for more information to use in breaking the tie (Hutchinson and Veenendall 2005b, a).

In any case, once the appropriate reference feature has been selected, the algorithm must determine the appropriate geocode for output based on the input and the reference feature. In the case of a precompiled geocoded dataset such as the ADDRESS-POINT (Ordnance Survey 2006) and G-NAF (Paull 2003), the algorithm can simply return the existing geographic representation. However, in the case of TIGER (U.S. Census Bureau 2006), the output geography must be derived based on the line segment determined to be a match. Here interpolation algorithms deduce the appropriate output geography based on attributes of the street segment such as address ranges and polarity (Drummond 1995, Levine and Kim 1998, Ratcliffe 2001, Cayo and Talbot 2003, Davis et al. 2003). In general, these interpolation algorithms work by first identifying the correct street segment in the reference data source based on the attributes of the address to be geocoded and the attributes of the street segment (address ranges associated with both sides of the segment, street name, street suffix, etc.). Once found, the appropriate side of the street segment is ascertained using the polarity (even/odd) of the address and each of the street segment sides. The correct location along the street segment is then determined by computing where the addresses in question would fall as a proportion of the total address range associated with the appropriate side of the street

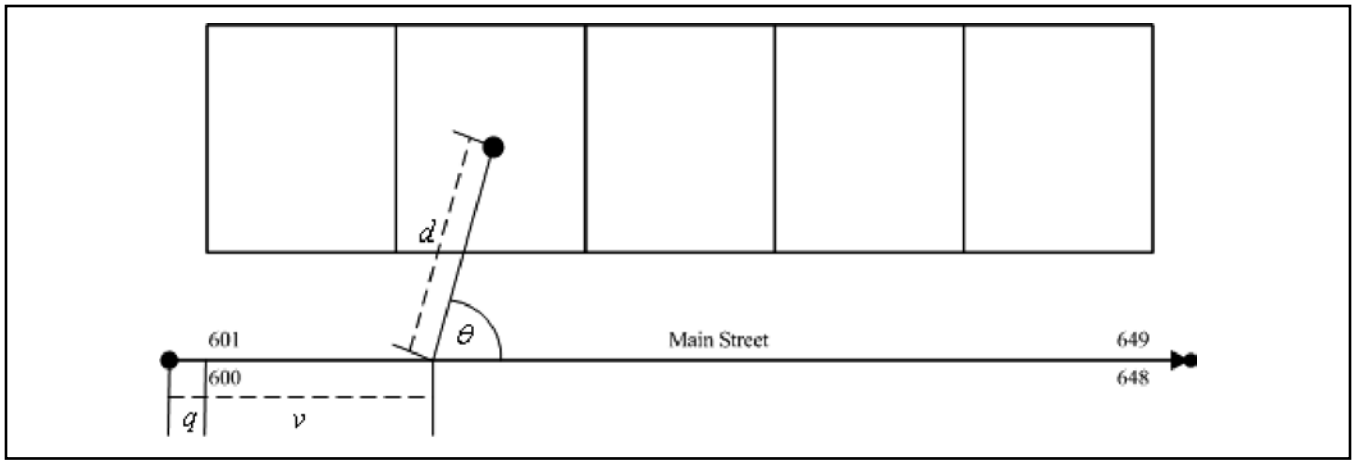

Figure 3. Sample block showing parameters of the geocoding algorithm

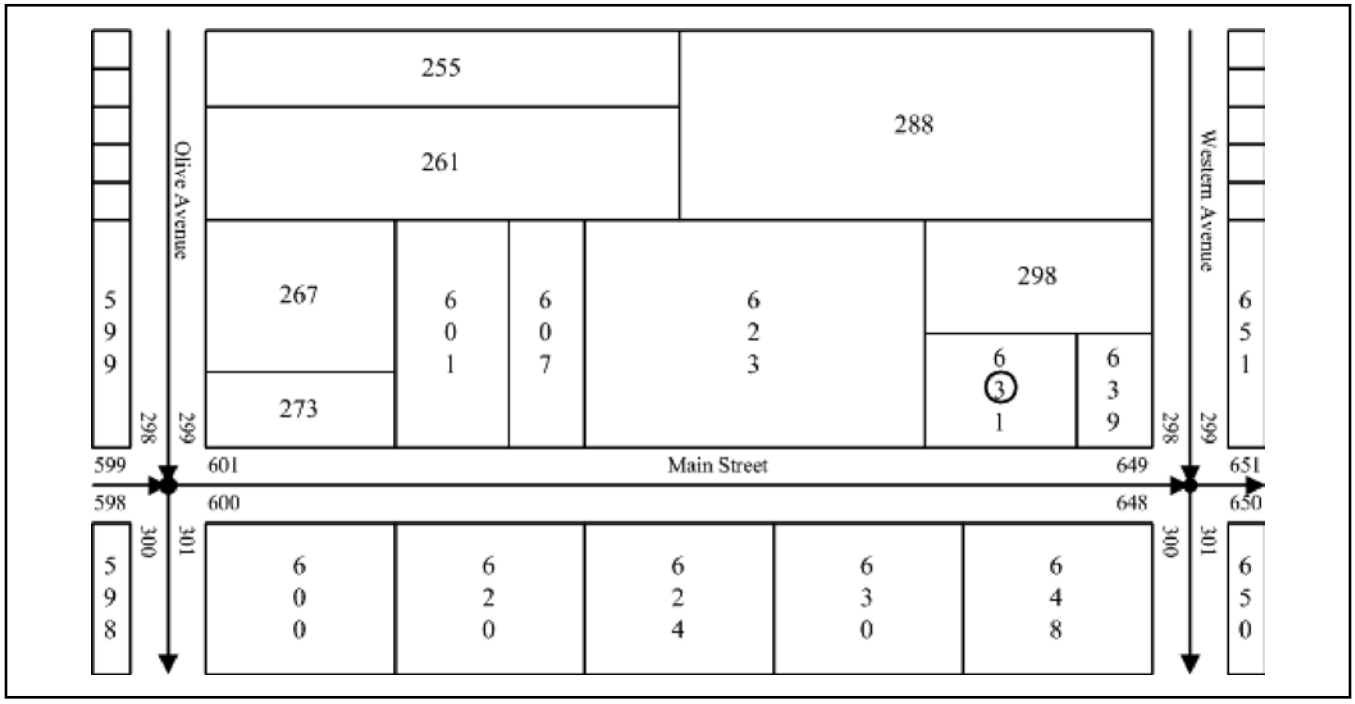

Figure 4. Sample address block with true parcel arrangement showing true geocoded point as ring

segment. This proportion is then applied to the total length of the street segment to obtain a location along the centerline of the street, and additional parameters such as distance and direction from the street center and offset from the endpoints of the street can be introduced to further improve the accuracy (Ratcliffe 2001, Cayo and Talbot 2003). Additional data sources can be consulted to obtain knowledge about the number of parcels on the street and their geographic distribution (Bakshi et al. 2004) to overcome the parcel homogeneity assumption (Dearwent et al. 2001) that all parcels within an address range truly exist and have the same dimensions. In Figures 3 through 6 these points are illustrated.

Figure 3 shows the parameters for the interpolation algorithm, d and θ, the street centerline offset distance and angle, q, the corner offset distance, and v, the interpolated distance to the center of the parcel. Also shown are the address ranges for each side of the segment, 601 through 649 on the odd parity side, and 600 through 648 on the even parity side. Figure 4 shows a sample block segment with the geocoded position of 631 Main Street displayed. Figure 5 displays how the parcel homogeneity assumption divides the segment into equal portions for all addresses within the range of the street segment, placing the geocoded point for address 631 at the wrong location (shown as ring) compared to the true location (shown as shaded ring). Figure 6 also displays the parcel homogeneity assumption, but in this case the true number of parcels on the street is known and the resulting geocoded point for address 631 is at a closer location (shown as ring) to that of the true location (shown as shaded ring). When using area-based reference features such as postal code and parcel polygons to compute point geographies

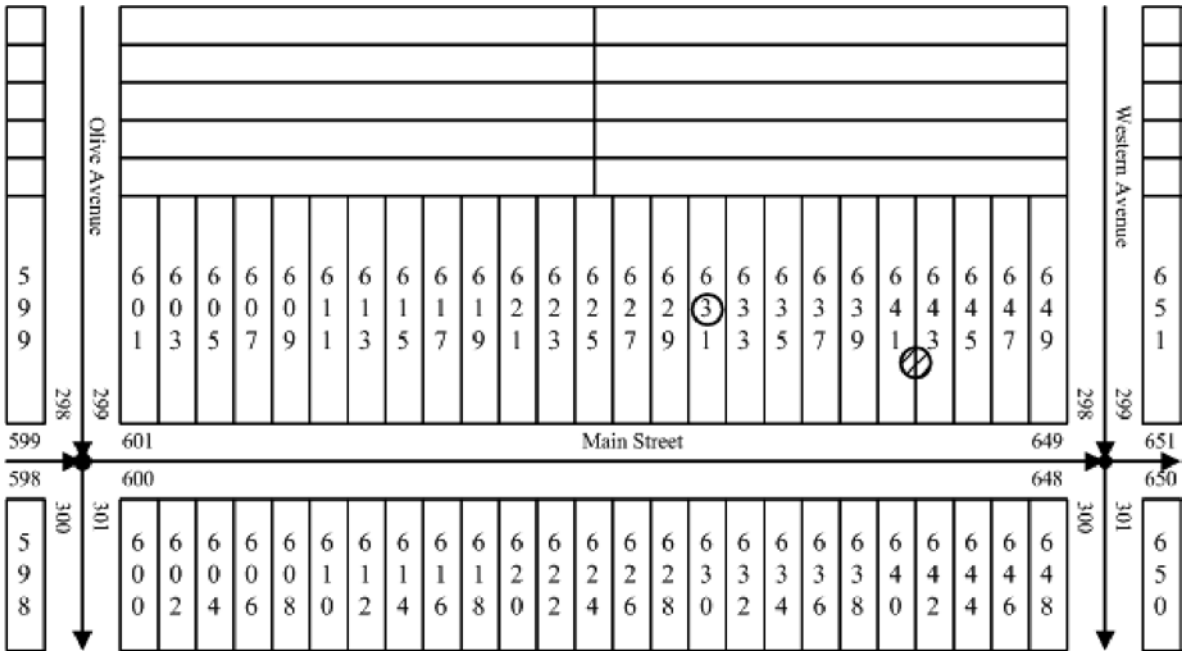

Figure 5. Sample address block with parcel homogeneity assumption using address range showing erroneous geocoded point as ring and true geocoded point as shaded ring

to return as output, the algorithm must calculate an appropriate centroid (Stevenson et al. 2000, Dearwent et al. 2001, Ratcliffe 2001). It may simply return the center of mass of the object, or it may perform more complex calculations in conjunction with other information such as population distributions across an area to determine a more representative weighted centroid (Gatrell 1989, Durr and Froggatt 2002).

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

GEOCODING ERROR

This broad definition of geocoding also brings with it a significant burden in the form of anticipating and/or quantifying geocoding error. Even simply defining what the error of the geocoding process is presents an arduous task. When speaking of geocoding error, is reference made to the positional accuracy of the returned geographic object, the probability that the feature returned is the one that was desired, or the validity of one or more assumptions used by the geocoding algorithm? Further definitions could include the error caused by the match rate, the weighting and relaxation techniques used in the standardization process, or the confidence cutoffs used during probabilistic matching. Common causes and effects of errors in each stage of the geocoding process are listed in Table 1.

Table 1. Common Causes and Effects of Errors in Stages of the Geocoding Process

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

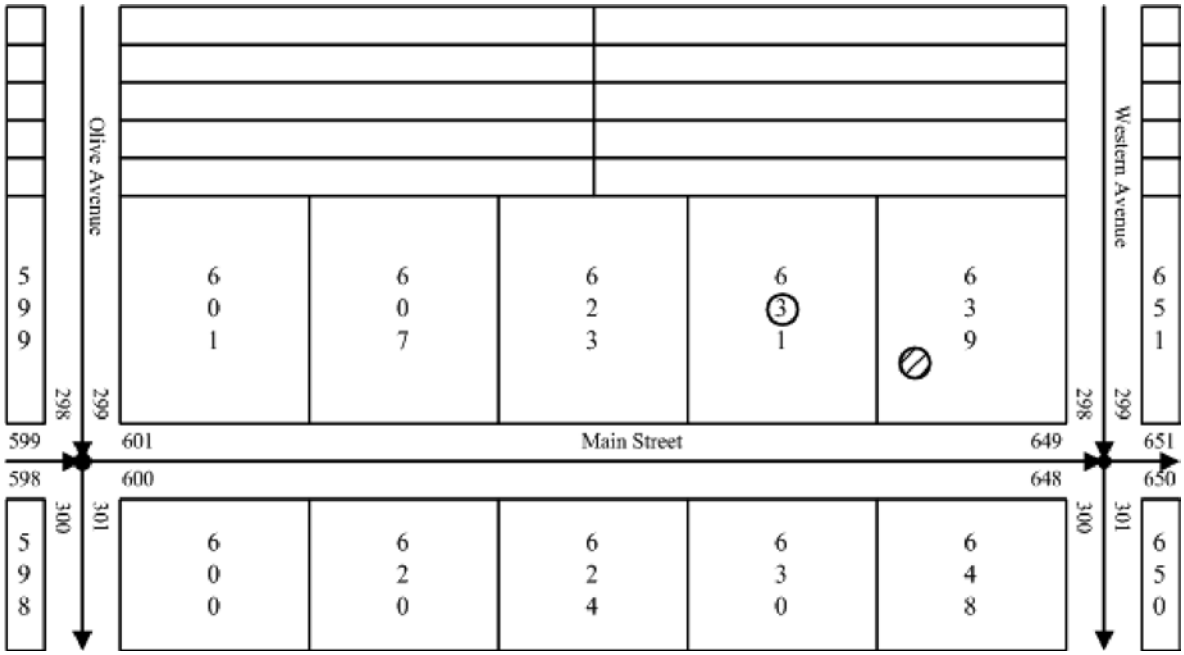

Figure 6. Sample address block with parcel homogeneity assumption using actual number of parcels showing erroneous geocoded point as ring and true geocoded point as shaded ring

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

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

Second, how accurate is the underlying data used as the reference dataset? Included in this discussion should be the concepts of spatial accuracy (how close are the features in the dataset to what is found on the ground [Karimi et al. 2004, Wu et al. 2005]?), temporal accuracy (how close are the features in this dataset to how they were at the time period of interest to me [McElroy et al. 2003, Han et al. 2005]?), original collection purpose (what were these data originally collected for [Boulos 2004]?), and lineage (what processes have been applied to this data [Veregin 1999]?). These aspects may be difficult to quantify because the accuracy measurements associated with datasets are estimates over the entire dataset, not on a per-feature basis. For example, while achieving an acceptable accuracy for short street segments in urban areas, the TIGER (U.S. Census Bureau 2006) datasets most commonly used for linear interpolation geocoding in the United States are known to be far less accurate for geocoding in rural areas with longer street segments (Drummond 1995, Vine et al. 1998, Cayo and Talbot 2003, Bonner et al. 2003, Wu et al. 2005). Assuming a consistent accuracy value for a dataset throughout the entire area of coverage is rarely discussed or noted as a point of contention in the determination of geocoding accuracy.

A third related issue arises when one considers multitiered geocoding approaches using multiple data sources. For example, in numerous instances, geocoding match rates in rural areas are far less than in urban areas (e.g., Gregorio et al. 1999, Kwok and Yankaskas 2001, Boscoe et al. 2002, Bonner et al. 2003, Cayo

and Talbot 2003). The typical approach to solving this problem involves a decision of whether to geocode to a less precise level or to include additional detail from other sources to determine the correct geocode. Choosing either case creates a resulting dataset with varying degrees of accuracy as a function of location, a condition recently defined as "cartographic confounding" (Oliver et al. 2005) that has been alluded to many times, yet remained undefined throughout the history of geocoding research (Block 1995, Ratcliffe 2001, Cayo and Talbot 2003, Nuckols et al. 2004, Ratcliffe 2004, Gregorio et al. 2005). A per-geocode accuracy is rarely maintained as a result of the geocoding process other than the level of geography matched to (i.e., census tract versus block group), and rarely do spatial models include variables to model this phenomena, although some researchers (Openshaw 1989, Arbia et al. 1998, Cressie and Kornak 2003, Gabrosek and Cressie 2002) have begun developing models to account for it. Despite this, information describing the varying degrees of accuracy of each individual geocode is not typically represented during subsequent spatial analysis.

Fourth, one needs to determine if the assumptions made by the geocoding algorithm are applicable to one's needs. As previously mentioned, the most common form of geocodin (linear interpolation–based) makes several key assumptions that can affect the level of accuracy of the results. First, it assumes that all addresses within an address range exist. Thus, when it determines the correct location for a particular address along a street segment by identifying the proportion along the segment where an address should fall, it will overestimate the number of addresses placing it at the wrong location. Second, it assumes a homogeneous distribution of addresses in terms of lot placement and size, known as the parcel homogeneity assumption (Dearwent et al. 2001, 332). This means that each lot on the street is assumed to have the same dimensions, and be oriented in the same direction, which is typically not a realistic assumption. Furthermore, it does not take into account that the corner lot on a segment may belong to the segment in question, or to the segment that forms the corner (Bakshi et al. 2004). While the magnitude of error introduced by these assumptions is small (on the order of half the length of the street segment [Wu et al. 2005, 596]), it can have dramatic effects when the variable and/or relationships of interest (e.g., environmental exposure doses to pesticide [Brody et al. 2002, Kennedy et al. 2003], air pollution [Wu et al. 2005], or proximity to voting precincts [Haspel and Knotts 2005]) vary over tens or hundreds of meters, and becomes amplified as the landscape becomes more rural. Additionally, it has been shown that when geocodes are used for point-in-polygon operations to derive attributes from other datasets, small spatial errors in geocodes that lie along borders between the larger level features can cause serious misclassifications in combined data (Ratcliffe 2001, Schootman et al. 2004).

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

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

PERSISTENT GEOCODING DIFFICULTIES

For all the technological advances and improvements that have been made to the geocoding process and the underlying reference datasets, the geocoding difficulties identified early on still exist. In developing countries with little GIS data infrastructure, the main roadblock to accurate geocoding is the simple nonexistence of reference datasets or GIS data infrastructure (Croner 2003, United Nations Economic Commission 2005). The development of basic GIS reference datasets is hindered by the existence of slum-like areas that change frequently, contain geographic features

that are not street addressable, and where many areas lack a consistent addressing scheme (Davis 1993, Oppong 1999, Davis et al. 2003, United Nations Economic Commission 2005). Efforts are under way to remedy these situations by developing standardized addressing systems that include facets for encouraging public participation aimed at promoting acceptance and eventual adoption, but these are costly endeavors being undertaken in areas with few economic resources to dedicate to the task (United Nations Economic Commission 2005).

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

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

As people move away from traditional land-line phones with the adoption of cell phone technology, some may argue that the promise of E911 solving addressing issues will begin to disappear. However, while it is true that in the future more calls will undoubtedly be made from cell phones, this is irrelevant for most municipalities still assume that structures will have phones and legislation is often in place that requires the E911 system to be kept up-to-date and accurate. As such, when official addresses are requested for new construction, the department responsible for maintaining the E911 system will most likely be required to visit the property and assign the E911-based geocode for the address.

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

CONCLUSION

This article has explored the state of the art in geocoding through a discussion of the path geocoding and its reference datasets have taken over the years. This work should serve as a starting point from which potential geocoding projects can be undertaken with regard to identifying the potential pitfalls and challenges that are

commonly encountered. Each particular geocoding project will have its own requirements in terms of input and output data structure and format, confidentiality, cost, available tools, and technical know-how, but the survey presented here should allow a more thorough understanding of the ramifications of particular choices made during the process.

Acknowledgments

This research is based on work supported in part by the National Science Foundation under Award Number IIS-0324955 and in part by the University of Southern California Libraries. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of these organizations or any person connected with them.

About The Authors

Dan W. Goldberg is a third-year computer science Ph.D. student working in the GIS Research Laboratory at the University of Southern California. He is a recent recipient of the 2005–2006 U.S. Geospatial Intelligence Foundation's Graduate Student Scholarship, and his research interests include geographic information extraction and integration, automated approaches to building highly detailed and accurate gazetteers, and developing new methods for geocoding textual locational descriptions.

Corresponding Address:
GIS Research Laboratory
University of Southern California
Los Angeles, CA 90089-0255
Phone: (213) 740-8263
E-mail: daniel.goldberg@usc.edu

John P. Wilson is a professor of geography and Director of the GIS Research Laboratory at the University of Southern California. He is the founding editor of *Transactions in GIS*, an active participant in the UNIGIS International Network, and Past-President of the University Consortium for Geographic Information Science. His major publications include two books *(Terrain Analysis: Principles and Applications; Handbook of Geographic Information Science)* along with numerous book chapters and journal articles on topics ranging from soil erosion and groundwater pollution problems to urban growth modeling and the environmental and social characteristics of place and their impacts on selected health outcomes.

Corresponding Address:
Department of Geography
University of Southern California
Los Angeles, CA 90089-0255
Phone: (213) 740-1908
E-mail: jpwilson@usc.edu.

Craig A. Knoblock is a senior project leader at the Information Sciences Institute and a research professor in computer science at the University of Southern California. He received his Ph.D. in computer science from Carnegie Mellon University. His current research interests include information integration, automated planning, machine learning, constraint reasoning, and the application of these technologies to geospatial data integration. He is currently President of the International Conference on Automated Planning and Scheduling and a fellow of the American Association of Artificial Intelligence.

References

Arbia, G., D. Griffith, and R. Haining. 1998. Error propagation modeling in raster GIS: overlay operations. Int. Journal of Geographical Information Science 12(2): 145-67.

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

Bakshi, R., C. A. Knoblock, and S. Thakkar. 2004. Exploiting online sources to accurately geocode addresses. In D. Pfoser, I. F. Cruz, and M. Ronthaler, eds., ACM-GIS '04: Proceedings of the 12th ACM International Symposium on Advances in Geographic Information Systems, Washington D.C., November 2004, 194-203.

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

Block, R. 1995. Geocoding of crime incidents using the 1990 TIGER file: the Chicago example. In C. R. Block, M. Dabdoub, and S. Fregly, eds., Crime analysis through computer mapping. Washington, D.C.: Police Executive Research Forum, 15.

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): 408-11.

Boscoe, F. P., C. L. Kielb, M. J. Schymura, and T. M. Bolani. 2002. Assessing and improving census track completeness. Journal of Registry Management 29(4): 117-20.

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. Int. Journal of Health Geographics 3(28).

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

Brody, J. G., D. J. Vorhees, S. J. Melly, S. R. Swedis, P. J. Drivas, and R. A. Rudel. 2002. Using GIS and historical records to reconstruct residential exposure to large-scale pesticide ap-

plication. Journal of Exposure Analysis and Environmental Epidemiology 12(1): 64-80.

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

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

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

Christen, P., and T. Churches. 2005. A probabilistic reduplication, record linkage and geocoding system. In Proceedings of the Australian Research Council Health Data Mining Workshop (HDM05), Canberra, AU, April 2005. In press, acrc.unisa.edu.au/groups/healthhdw2005Christen.pdf.

Christen, P., T. Churches, and A. Willmore. 2004. A probabilistic geocoding system based on a national address file. Australasian Data Mining Conference, Cairns, AU, December 2004. http://datamining.anu.edu.au/publications/2004/ausdm2004.pdf.

Chung, K., D.-H. Yang, and R. Bell. 2004. Health and GIS: toward spatial statistical analyses. Journal of Medical Systems 28(4): 349-60.

Churches, T., P. Christen, K. Lim, and J. X. Zhu. 2002. Preparation of name and address data for record linkage using hidden Markov models. Medical Informatics and Decision Making 2(9). http://www.biomedcentral.com/content/pdf/1472-6947-2-9.pdf.

Collins, S. E., R. P. Haining, I. R. Bowns, D. J. Crofts, T. S. Williams, A. S. Rigby, and D. M. Hall. 1998. Errors in postcode to enumeration district mapping and their effect on small area analyses of health data. Journal of Public Health Medicine 20(3): 325-30.

Cressie, N., and J. Kornak. 2003. Spatial statistics in the presence of location error with an application to remote sensing of the environment. Statistical Science 18(4): 436-56.

Croner, C. M. 2003. Public health GIS and the Internet. Annual Review of Public Health 24: 57-82.

Davis Jr., C. A. 1993. Address base creation using raster/vector integration. Proceedings of the URISA 1993 Annual Conference, Atlanta, GA, 45-54.

Davis Jr., C. A., F. T. Fonseca, and K. A. De Vasconcelos Borges. 2003. A flexible addressing system for approximate geocoding, GeoInfo 2003: Proceedings of the Fifth Brazilian Symposium on GeoInformatics, Campos do Jordão, São Paulo, Brazil, October 2003.

Dearwent, S. M., R. R. Jacobs, and J. B. Halbert. 2001. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis Environmental Epidemiology 11(4): 329-34.

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

Drummond, W. J. 1995. Address matching: GIS technology for mapping human activity patterns. Journal of the American Planning Association 61(2): 240-51.

Dueker, K. J. 1974. Urban geocoding. Annals of the Association of American Geographers 64(2): 318-25.

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

Eichelberger, P. 1993. The importance of addresses: the locus of GIS. Proceedings of the URISA 1993 Annual Conference, Atlanta, GA, 212-13.

Fonda-Bonardi, P. 1994. House numbering systems in Los Angeles. Proceedings of the GIS/LIS '94 Annual Conference and Exposition, Phoenix, AZ, October 1994, 322-31.

Frew, J., M. Freeston, N. Freitas, L. L. Hill, G. Janee, K. Lovette, R. Nideffer, T. R. Smith, and Q. Zheng. 1998. The Alexandria digital library architecture. In C. Nikalaou and C. Stephanidis, eds., ECDL '98: Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, Greece, September 1998. Lecture notes in computer science, Vol. 1513. London, UK: Springer, 61-73.

Fulcomer, M. C., M. M. Bastardi, H. Raza, M. Duffy, E. Dufficy, and M. M. Sass. 1998. Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996. In R. C. Williams, M. M. Howie, C. V. Lee, and W. D. Henriques, eds., Proceedings of the 1998 Geographic Information Systems in Public Health Conference, San Diego, CA, August 1998, 547-60. http://www.atsdr.cdc.gov/GIS/conference98/proceedings/pdf/gisbook.pdf.

Gabrosek, J., and N. Cressie. 2002. The effect on attribute prediction on location uncertainty in spatial data. Geographical Analysis 34: 262-85.

Gaffney, S. H., F. C. Curriero, P. T. Strickland, G. E. Glass, K. J. Helzlsouer, and P. N. Breysse. 2005. Influence of geographic location in modeling blood pesticide levels in a community surrounding a U.S. Environmental Protection Agency Superfund Site. Environmental Health Perspectives 113(12): 1712-16.

Gatrell, A. C. 1989. On the spatial representation and accuracy of address-based data in the United Kingdom. Int. Journal of Geographical Information Systems 3(4): 335-48.

Geronimus, A. T., and J. Bound. 1998. Use of census-based aggregate variables to proxy for socioeconomic group: evidence

from national samples. American Journal of Epidemiology 148(5): 475-86.

Geronimus, A. T., and J. Bound. 1999a. Re: Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. American Journal of Epidemiology 150(8): 894-6. Letter.

Geronimus, A. T., and J. Bound. 1999b. Re: Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. American Journal of Epidemiology 150(9): 997-9. Letter.

Geronimus, A. T., J. Bound, and L. J. Neidert. 1995. On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. Technical Working Paper 189. Cambridge, MA: National Bureau of Economic Research.

Gottesfeld Brown, L. 1992. A survey of image registration techniques. ACM Computing Surveys 24(4): 325-76.

Gregorio, D. I., E. Cromley, R. Mrozinski, and S. J. Walsh. 1999. Subject loss in spatial analysis of breast cancer. Health & Place 5(2): 173-7.

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

Griffin, D. H., J. M. Pausche, E. B. Rivers, A. L. Tillman, and J. B. Treat. 1990. Improving the coverage of addresses in the 1990 census: preliminary results. Proceedings of the American Statistical Association Survey Research Methods Section, Anaheim, CA, August 1990, 541-6. http://www.amstat.org/sections/srms/Proceedings/papers/1990+091.pdf.

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

Haspel, M., and H. G. Knotts. 2005. Location, location, location: precinct placement and the costs of voting. The Journal of Politics 67(2): 560-73.

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

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

Hill, L. L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In J. L. Borbinha and T. Baker, eds., ECDL '00: research and advanced technology for digital libraries. 4th European Conference, Lisbon, Portugal, September 2000. Lecture notes in computer science, Vol. 1923. London, UK: Springer, 280-90.

Hill, L. L., J. Frew, and Q. Zheng. 1999. Geographic names: the implementation of a gazetteer in a georeferenced digital library. D-Lib Magazine 5(1). http:// www.dlib.org/dlib/january99/hill/01hill.html.

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

Hurley, S. E., T. M. Saunders, R. Nivas, A. Hertz, and P. Reynolds. 2003. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 14(4): 386-91.

Hutchinson, M., and B. Veenendall. 2005a. Towards a framework for intelligent geocoding. In SSC 2005 spatial intelligence, innovation and praxis. The National Biennial Conference of the Spatial Sciences Institute, Melbourne, AU, September 2005.

Hutchinson, M., and B. Veenendall. 2005b. Towards using intelligence to move from geocoding to geolocating. Proceedings of the 7th Annual URISA GIS in Addressing Conference, Austin, TX, August 2005.

Jaro, M. 1984. Record linkage research and the calibration of record linkage algorithms. Statistical Research Division Report Series SRD Report No. Census/SRD/RR-84/27. Washington, D.C.: U.S. Census Bureau. http://www.census.gov/srd/papers/pdf/rr84-27.pdf.

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

Johnson, S. D. 1998b. Address matching with stand-alone geocoding engines: part 2. Business Geographics, April, 30-36.

Karimi, H. A., M. Durcik, and W. Rasdorf. 2004. Evaluation of uncertainties associated with geocoding techniques. Journal of Computer-Aided Civil and Infrastructure Engineering 19(3): 170-85.

Kennedy, T. C., J. G. Brody, and J. N. Gardner. 2003. Modeling historical environmental exposures using GIS: implications for disease surveillance. Proceedings of the 2003 ESRI Health GIS Conference, Arlington, Virginia, May 2003. http://gis.esri.com/library/userconf/health03/papers/pap3020/p3020.htm.

Krieger, N. 1992. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. American Journal of Public Health 82(5): 703-10.

Krieger, N. 2003. Place, space, and health: GIS and epidemiology. Epidemiology 14(4): 384-85.

Krieger, N., J. T. Chen, P. D. Waterman, D. H. Rehkopf, and S. V. Subramanian. 2005. Painting a truer picture of U.S. socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. American Journal of Public Health 95(2): 312-23.

Krieger, N., J. T. Chen, P. D. Waterman, M.-J. Soobader, S. V. Subramanian, and R. Carson. 2002a. Geocoding and monitoring of U.S. socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? American Journal of Epidemiology 156(5): 471-82.

Krieger, N., and D. Gordon. 1999. Re: Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. American Journal of Epidemiology 150(8): 894-6.

Krieger, N., P. Waterman, J. T. Chen, M.-J. Soobader, S. V. Subramanian, and R. Carson. 2002b. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and U.S. census-defined areas: The Public Health Disparities Geocoding Project. American Journal of Public Health 92(7): 1100-2.

Krieger, N., P. D. Waterman, J. T. Chen, M.-J. Soobader, and S. V. Subramanian. 2003. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures. Public Health Reports 118(3): 240-60.

Kwok, R. K., and B. C. Yankaskas. 2001. The use of census data for determining race and education as SES indicators: a validation study. Annals of Epidemiology 11(3): 171-7.

Laender, A. H. F., K. A. V. Borges, J. C. P. Carvalho, C. B. Medeiros, A. S. da Silva, and C. A. Davis, Jr. 2005. Integrating Web data and geographic knowledge into spatial databases. In Y. Manalopoulos and A. N. Papadapoulos, eds., Spatial databases: technologies, techniques and trends, Chap. 2. (Hershey, PA: Idea Group Inc.), 23-47.

Lee, J. 2004. 3D GIS for geo-coding human activity in micro-scale urban environments. In M. J. Egenhofer, C. Freksa, and H. J. Miller, eds., Geographic information science. Third International Conference, GIScience 2004, College Park, MD, October 2004, 162-78.

Levesque, M. 2003. West Virginia Statewide Addressing and Mapping Project. Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference, Providence, RI, August 2003. http://www.urisa.org/Street Smart Conference/2003/LevesqueM.pdf.

Levine, N., and K. E. Kim. 1998. The spatial location of motor vehicle accidents: a methodology for geocoding intersections. Computers, Environment, and Urban Systems 22(6): 557-76.

Lind, M. 2001. Developing a system of public addresses as a language for location dependent information. In Proceedings of the 2001 URISA Annual Conference, Long Beach, CA, October 2001. http://www.adresseprojekt.dk/files/Develop-PublicAddress urisa2001e.pdf.

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

Locative Technologies. 2006. Geocoder.us: a free U.S. geocoder. http://geocoder.us.

Lockyer, B. Office of the Attorney General of the State of California Legal Opinion 04-1105. http://ag.ca.gov/opinions/pdfs/04-1105.pdf.

Martin, D. J. 1998. Optimizing census geography: the separation of collection and output geographies. Int. Journal of Geographical Information Science 12(7): 673-85.

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

Martin, D. J. 1999. Spatial representation: the social scientist's perspective. In P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, eds., Geographical information systems, Vol. 1, 2nd Ed. New York: Wiley, 6: 71-80.

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.

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

Nicoara, G. 2005. Exploring the geocoding process: a municipal case study using crime data. Master's thesis, University of Texas at Dallas, Dallas, TX.

Nuckols, J. R., M. H. Ward, and L. Jarup. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environmental Health Perspectives 112(9): 1007-15.

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

Olligschlaeger, A. M. 1998. Artificial neural networks and crime mapping. In D. Weisburd and T. McEwen, eds., Crime mapping and crime prevention. Crime prevention studies, Vol. 8. Monsey, NY: Criminal Justice Press, 313-47.

Openshaw, S. 1989. Learning to live with errors in spatial databases. In M. F. Goodchild and S. Gopal, eds., Accuracy of spatial databases. Bristol, PA: Taylor and Francis, 23: 263-76.

Oppong, J. R. 1999. Data problems in GIS and health. Proceedings of Health and Environment Workshop 4: Health Research Methods and Data, Turku, Finland, July 1999. http://geog.queensu.ca/hande/healthandenvir/FinlandWorkshop Papers/OPPONG.DOC.

Ordnance Survey. 2006. ADDRESS-POINT: Ordnance Survey's map dataset of all postal addresses in Great Britain. http://www.ordnancesurvey.co.uk/oswebsite/products/address-point.

O'Reagan, R. T., and A. Saalfeld. 1987. Geocoding theory and practice at the Bureau of the Census. Statistical Research Report Census/SRD/RR-87/29. Washington, D.C.: U.S. Census Bureau.

Paull, D. 2003. A geocoded national address file for Australia: the G-NAF what, why, who and when? http://www.addressonline.com.au/addressonline/home/GNAF What Why Who When.pdf.

Pohl, C., and J. L. Van Genderen. 1998. Review article: multisensor image fusion. In Remote sensing: concepts, methods and applications. Int. Journal of Remote Sensing 19(5): 823-54.

Ratcliffe, J. H. 2001. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. Int. Journal of Geographical Information Science 15(5): 473-85.

Ratcliffe, J. H. 2004. Geocoding crime and a first estimate of a minimum acceptable hit rate. Int. Journal of Geographical Information Science 18(1): 61-72.

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

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

Schootman, M., D. Jeffe, E. Kinman, G. Higgs, and J. Jackson-Thompson. 2004. Evaluating the utility and accuracy of a reverse telephone directory to identify the location of survey respondents. Annals of Epidemiology 15(2): 160-6.

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

Smith, G. D., Y. Ben-Shlomo, and C. Hart. 1999. Re: Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. American Journal of Epidemiology 150(9): 996-7.

Soobader, M., F. B. LeClere, W. Hadden, and B. Maury. 2001. Using aggregate geographic data to proxy individual socioeconomic status: does size matter? American Journal of Public Health 91(4): 632-6.

Stage, D., and N. von Meyer. 2005. An assessment of parcel data in the United States 2005 Survey results. Federal Geographic Data Committee Subcommittee on Cadastral Data Report. http://www.nationalcad.org/showdocs.asp?docid=170.

Stevenson, M. A., J. Wilesmith, J. Ryan, R. Morris, A. Lawson, D. Pfeiffer, and D. Lin. 2000. Descriptive spatial analysis of the epidemic of bovine spongiform encephalopathy in Great Britain to June 1997. The Veterinary Record 147(14): 379-84.

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

Tobler, W. 1972. Geocoding theory. In Proceedings of the National Geocoding Conference, Washington D.C. Washington, D.C.: Department of Transportation, IV.1.

United Nations Economic Commission. 2005. A functional addressing system for Africa: a discussion paper. http://geoinfo.uneca.org/Docs/Situs Addressing background paper-Draft.pdf.

U.S. Census Bureau. 2006. Topologically integrated geographic encoding and referencing system. Washington, D.C.: U.S. Census Bureau. http://www.census.gov/geo/www/tiger.

U.S. Department of Health and Human Services. 2000. Healthy people 2010: understanding and improving health, 2nd Ed. Washington, D.C.: U.S. Government Printing Office. http://www.healthypeople.gov/Document/html/uih/uih 2.htm.

Veregin, H. 1999. Data quality parameters. In P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, eds., Geographical information systems, 2nd Ed., Vol. 1, Chap. 12. New York, NY: Wiley, 177-89.

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

Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, J. S. Colt, and P. Hartge. 2005. Positional accuracy of two methods of geocoding. Epidemiology 16(4): 542-7.

Walls, M. D. 2003. Is consistency in address assignment still needed? Proceedings of the Fifth Annual URISA Street Smart and Address Savvy Conference, Providence, RI, August 2003. http://www.urisa.org/Street Smart Conference/2003/WallsM.pdf.

Werner, P. A. 1974. National geocoding. Annals of the Association of American Geographers 64(2): 310-7.

Wieczorek, J., Q. Guo, and R. J. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. Int. Journal of Geographical Information Science 18(8): 745-67.

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

Yahoo! Inc., 2006, Yahoo! Maps Web Services - Geocoding API. http://developer.yahoo.com/maps/rest/V1/geocode.html.

Yang, D.-H., L. M. Bilaver, O. Hayes, and R. Goerge. 2004. Improving geocoding practices: evaluation of geocoding tools. Journal of Medical Systems 28(4): 361-70.