Geocoding Quality and Implications for Spatial Analysis

Paul A. Zandbergen*
University of New Mexico

Abstract

Many spatial analysis techniques rely on the ability to geocode individual locations based on addresses or other descriptive information. The quality of geocoding and its effect on spatial analysis have received some attention in the literature, in particular in the field of health. This article reviews the foundation of geocoding and presents a framework for evaluating geocoding quality. Errors introduced by street geocoding include incompleteness, positional error, and incorrect assignment to geographic units. A review of empirical studies suggests that these errors are neither small nor random in nature and that substantial bias may be introduced in spatial analysis that employs the results of geocoding. Several alternatives have also emerged, including the use of address points and parcels, and these are gradually becoming more widely used. Several areas for future research on geocoding have been identified: (i) refinements of address data models to incorporate complex addressing situations; (ii) development of error propagation techniques to determine the level of geocoding quality required for a particular analysis scenario; (iii) development of measures of reliability for geocoding results; (iv) comparative analysis of geocoding quality across different jurisdictions; and (v) validation of online geocoding services and volunteered geographic information.

1 Introduction

Addresses are one of the fundamental means by which people conceptualize location in the modern world. In a Geographic Information System (GIS), addresses are converted to features on a map through the use of geocoding techniques. The general purpose of this article is fourfold: (i) to review the foundations of the geocoding process; (ii) to present a framework for evaluating geocoding quality; (iii) to summarize the current understanding of the various dimensions of geocoding quality based on existing empirical studies; and (iv) to review the implications of errors in geocoding for spatial analysis. The first review portion complements recent reviews of geocoding by Rushton et al. (2006), Rushton et al. (2007) and Goldberg et al. (2007). This article focuses more specifically on issues of geocoding quality and its implications for spatial analysis. Some emphasis is placed on applications in the fields of health and crime, but many of the findings have wider applicability.
Many types of geocoding exist and have their own set of methods and limitations. This article focuses on the geocoding of residential addresses down to individual locations as these are of most interest in several fields including criminal justice, epidemiology, public health, and socio-economic analysis in general. Several other types of geocoding exist but have been covered elsewhere, including events along networks, such as traffic accidents (Levine and Kim 1998), locality descriptions without addresses (Wieczorek et al. 2004), industrial facilities (García-Pérez et al. 2008) and broad textual descriptions of relative locations (Davis and Fonseca 2007).

2 Geocoding Foundations

2.1 Geocoding Process

Geocoding is the process of assigning an XY coordinate pair to the description of a place by comparing the descriptive location-specific elements to those in reference data. The geocoding process consists of translating an address entry, searching for the address in the reference data, and delivering the best candidate or candidates as a point feature on the map. This includes parsing the input address into address components (e.g., street name and street type), standardizing abbreviated values, assigning each address element to a category known as a match key, indexing the needed categories, searching the reference data, assigning a score to each potential candidate, filtering the list of candidates based on the minimum match score, and delivering the best match. Techniques involved in geocoding borrow from various academic fields, most notably, information theory, decision theory, probability theory, and phonetics.

While geocoding applications are diverse and span many different fields, there are several common problems associated with geocoding that have traditionally caused poor match rates, requiring excessive manual mapping by the user and potential inaccuracies and/or incompleteness in the resulting spatial datasets (e.g., Goldberg et al. 2007; Rushton et al. 2006).

Geocoding relies on record linkage that can be deterministic or probabilistic. In the deterministic approach, each address element being compared is evaluated and given a score based on how well it matched. Predetermined weights for each element are then used to determine an overall match score, which is then compared to a user-specified minimum. In the probabilistic approach, data files are matched under conditions of uncertainty and a form of fuzzy logic is used to score how well records match. Each address element participating in the linkage comparison is subject to error that is measured by the probability that the field agrees versus the probability of chance agreement of its values. The assignment of such probabilities is intended to mimic a human decision-making process. These different approaches have been addressed in some detail by other studies (e.g., Jaro
Major components in the linkage process are generally predefined in addresses, such as house number, prefix direction, street name and street type. However, there are significant regional variations in many parts of the world that may require the use of custom locator styles for differing datasets. For each record pair, a decision is made whether to classify the pair as a match (M), a non-match (U), or a tie (T), which can be followed up interactively by the user to manually specify the correct match.

### 2.2 Address Data Models for Geocoding

One of the main challenges to accurate geocoding is the availability of good reference data. This includes a set of geographic features to match against as well as robust address characteristics that enable matching address records to feature locations. This requires a sturdy address model to organize the reference data components in a logical, maintainable, and site-specific way. Several common address models exist, each with a particular set of supporting materials and characteristic errors. The first one can be characterized as the ‘geographic unit’ model. Geographic units can consist of postal codes (such as ZIP codes in the USA), counties, cities, census enumeration areas or any other geographic boundary considered meaningful. In the geocoding process, the location assigned to a particular address is the polygon (or the polygon centroid) representing the geographic unit. The utility of the results is obviously related to the size of the geographic units. For example, in the USA, five-digit ZIP codes tend to be quite large, making them less attractive when spatially detailed information is required. In other jurisdictions postal codes may be of sufficient spatial resolution. When geocoding at the level of geographic units is not sufficient several alternatives exist, including street networks, parcels, and address points. Each of these will be described in more detail below.

The most widely employed address data model is based on a street network represented as street line segments that hold street names and the range of house numbers on each side of the street. Address geocoding is accomplished by first matching the street name, then the segment that contains the house numbers and finally by placing a point along the segment based on linear interpolation within the range of house numbers. This approach to geocoding an address is referred to as ‘street geocoding’. Nearly all commercial firms providing geocoding services and most GIS software rely primarily on this type of geocoding. Figure 1 provides a conceptual diagram of this process.

Many different reference datasets are available for street geocoding, including national reference data, local street centerlines, and street network data from commercial providers. In the USA, one of the most common
sources is the Topologically Integrated Geographic Encoding and Referencing (TIGER) database from the US Census Bureau (2000). The TIGER data is available at no cost and provides a consistent data structure with coverage for the entire USA. Many low-cost geocoding systems rely on this data and commercial vendors make ‘enhanced’ versions available. While the limitations of the TIGER data have been recognized (e.g., Liadis 2000; O’Grady and Goodwin 2000), historically it has been the most widely used source of reference data for geocoding in the USA.

Parcels are traditionally the most spatially accurate data with address information available. Geocoding against parcels makes it possible to match against individual plots of land (or rather, the centroids of those polygons) rather than interpolating against a street centerline. Parcel geocoding typically results in much lower match rates, but is now becoming more widespread given the development of parcel level databases by many cities and counties in the USA (Rushton et al. 2006).

To overcome the limitations of parcels for geocoding, address points have emerged as an alternative address data model. Address points are commonly created from parcel centroids for all occupied parcels (or points can be placed elsewhere within the parcel, such as the location of the main structure or in front of the main structure). This is supplemented with address points for sub-addresses, such as individual apartment units, duplexes, etc. Field data collection or verification of building locations using digital aerial imagery can be used to further improve the address point file. Address point data are of great value to local government agencies,
in particular emergency services. Figure 2 shows an example of an address point data file in a GIS environment.

Australia, Canada, The Netherlands, and the UK have already developed national address point databases and research in these jurisdictions has started to use address points instead of street geocoding (e.g., Brimicombe et al. 2007; Watkins et al. 2007). These efforts have set the stage for other jurisdictions to develop similarly detailed and comprehensive databases, but there has been limited published research on the quality of the geocoding based on address points. In the USA, address point geocoding is not in very widespread use. However, many local governments have started to create address point databases and several commercial geocoding firms provide address point geocoding. Commercial firms claim that around 80 million address points are available for the USA (TeleAtlas 2008), covering a selected number of metropolitan areas.
2.3 A FRAMEWORK FOR GEOCODING QUALITY

Many issues can affect the quality of geocoding results, including the choice of address data model, the linear interpolation process (when using street geocoding), the quality of the input address data, the accuracy of the reference data, and the uncertainty introduced by the matching algorithm. A substantial body of literature has emerged on these subjects.

For the results of geocoding to be meaningful, the process needs to meet certain quality expectations. The overall quality of any geocoding result can be characterized by the following components: completeness, positional accuracy, concordance with geographic units, and repeatability. Completeness is the percentage of records that can reliably be geocoded, also referred to as the match rate. Positional accuracy indicates how close each geocoded point is to the ‘true’ location of the address. Concordance is the degree to which geocoded locations are assigned to the correct geographic unit of interest. Repeatability indicates how sensitive the geocoding results are to variations in the reference data input, the matching algorithms of the geocoding software, and the skills and interpretation of the analyst. Geocoding results of high quality are complete, spatially accurate, and placed within the correct geographic unit. Geocoding as a process is reliable if it provides high quality results and if it is repeatable.

3 Match Rates

The simplest measure of geocoding quality is the match rate, or the percentage of records that produce a reliable match. Match rates vary greatly as they depend on many factors and there is no consensus on a universal standard for an acceptable match rate. The match rate typically increases if efforts are made to improve the quality of the address file and/or the geographic reference file. Interpreting match rates, however, is very subjective, as much depends on the criteria used to characterize a ‘match’. For example, lowering the minimum match score will increase the overall match rate, but may inadvertently introduce false positives. For a given real-world set of addresses there is thus a trade-off: increasing the match rate by lowering the standards results in a decrease in accuracy.

An obvious question that emerges is: What is an acceptable match rate? Surprisingly, this question has received limited attention in the literature. In one of the few studies on the subject, Ratcliffe (2004) employed Monte Carlo simulation of geocoded crime incidents aggregated at the census block level to determine the minimum match rate needed to obtain a reliable pattern of crime incidents. Results indicated that a match rate of 85% was necessary. This threshold was determined using data aggregated to census unit and may not apply to analysis techniques employing individual geocoded locations.
While match rates vary greatly between different studies, most have found that match rates are much lower in rural areas compared to urban areas (e.g., Cayo and Talbot 2003; Dearwent et al. 2001; Drummond 1995; Kravets and Hadden 2007; Kwok and Yankaskas 2001). In rural areas, the use of rural routes and PO Boxes is common and these are not suitable for reliable geocoding. In most studies that report specific results for rural areas PO Boxes account for the majority of ungeocodable addresses. Hurley et al. (2003) demonstrated that residential addresses associated with PO Boxes can be obtained from post office rental records, but that this approach is unlikely to yield substantial improvements in overall geocoding success, particularly for older addresses. It is common practice to assign coordinates to PO Boxes based on the centroid of the postal code, typically weighted by delivery as determined by the postal service to account for the uneven distribution of residences within a postal code. This practice, however, may introduce substantial error, in particular given the fact that a single postal code in a rural area can be quite large. On the other hand, excluding PO Box holders from further analysis may introduce selection bias, as they are not necessarily representative of the whole population within a postal code (Hurley et al. 2003). As a result, PO Boxes remain a persistent challenge for geocoding in rural areas. The use of rural routes and PO Boxes is on the decline, however, as more rural residents receive street addresses. For example, Oliver et al. (2005) documented a sharp decline of the use of rural routes and PO Boxes in rural counties in Virginia.

Incomplete geocoding is a potential source of bias. Gilboa et al. (2006) found that for a sample of 9,912 residential addresses with a match rate of 83.2%, geocoded and non-geocoded populations differed substantially in terms of ethnicity, and that incomplete geocoding may have resulted in a selection bias in the analysis of air quality and birth effects. Oliver et al. (2005) found a strong pattern of differential match rates across urban-rural gradients in the geocoding of 26,338 cases of prostate cancer cases in Virginia. This confounded the analysis of the association between area-level socio-demographic factors and disease incidence. Skelly et al. (2002) determined systematic underestimates of rural rates of disease in disease surveillance systems due to disproportional poor match rates in rural areas.

Different types of addresses also result in different match rates. For example, higher match rates are typically obtained for residential addresses relative to commercial addresses (Zandbergen 2008a). Zandbergen (2008a) also found that geocoding match rates vary substantially by type of address database and by geographic area, suggesting that determining an ‘acceptable’ match rate requires very context specific considerations. The lack of consistency in match rates also suggests that geocoding quality is very much a function of the quality and consistency of local reference data.

Match rates also vary greatly with the address data model employed. Street geocoding is by far the most widely used method and match rates typically vary between 70 and 95%, although both lower and higher
match rates are sometimes reported. Geocoding against parcel boundaries is more spatially accurate, but results in much lower match rates, typically between 40 and 75% (Dearwent et al. 2001; Zandbergen 2008a). These can be attributed to number of factors. First, parcel databases are not created for the purpose of geocoding, so its structure and attributes often do not lend themselves for use in geocoding tools. Second, parcel databases are limited to describing legal properties, not necessarily the street addresses associated with them. Multi-unit properties, such as apartment complexes, mobile home parks, shopping plazas and college campuses, may consist of a single parcel but contain many street addresses. Unless a conscious effort has been made to capture these multiple addresses for each parcel, such locations cannot be geocoded using parcel data alone. Finally, as each parcel is typically associated with only a single street number, non-existent street numbers will not produce a match. In contrast, a non-existent street number will produce a match in street geocoding as long as the number falls within the address range coded for the street segment. In other words, street geocoding may include a number of false positives and this is much less likely to occur when using parcels. Geocoding against address points, which has only emerged in the last couple of years, has received much less attention, but the limited research so far indicates that match rates are very similar to those of street geocoding for the same address input data with superior positional accuracy (Zandbergen 2008a). This can largely be attributed to the fact that address points are typically created with the specific purpose of geocoding.

4 Positional Accuracy

Positional accuracy of geocoded locations is defined as the Euclidean distance between the geocoded point location and the actual location of the residence associated with the address. Different components contribute to the error:

1. Incorrect street segment. Due to errors in the input addresses and/or the street reference data, the address may be matched to the incorrect street segment. This typically results in very large positional errors, up to many kilometers.
2. Incorrect placement along the street segment. The linear interpolation algorithm assumes that the address range in the street reference data is reliable, that street numbers are evenly spaced along the segment, and that there are no irregularities in the sequence of street numbers. As these assumptions rarely apply to any real-world street segment, this results in positional error, which is obviously bound by the length of the segment.
3. Incorrect offset from the street segment. Most geocoding techniques employ a uniform perpendicular offset of around 10 to 15 m, which
may not be a good reflection of the actual distance of a residence from
the street centerline.

4. Positional error in the street segment. Any positional error of the street
segment itself may contribute further to the displacement of the geo-
coded location relative to the actual location.

In most empirical studies, these components are not addressed sepa-
rately and the measured error is therefore the aggregate effect of all four
components. Distinguishing between the various components, however,
is useful, as it provides insight into how street geocoding can be improved.
For example, excellent positional accuracy of street reference data is no
guarantee geocoding results will be accurate. Improved attribute accuracy
of street reference data will reduce the number of matches to incorrect
street segments, and improved address ranges will reduce the incorrect
placement along the street segment. However, the linear interpolation
algorithm employed in street geocoding has inherent limitations. An
example will serve to illustrate this.

Figure 3A shows a relatively short street segment in a typical medium
density single family residential area. Figure 3B shows the property bound-
aries overlaid on aerial photography. Figure 3C shows the street number
associated with each parcel as well as the address range coded for this street
segment in the attributes. The address range is a relatively close match
with the actual addresses and is correct in terms of polarity and direction,
and therefore presents a ‘near perfect’ scenario in terms of the address
range. The positional accuracy of the street segment is also reasonably
good as can be seen in the relatively close match with the imagery.
Figure 3D shows the result of geocoding all the addresses from the actual
properties along the segment. The lines connect the geocoded locations
with the centerpoint of the front parcel boundary of the associated property.
As can be seen in Figure 3D, the positional error is quite small, in the
order of tens of meters. The use of a side offset would further reduce the
error for most addresses.

The reality of most street segments, however, does not always follow
this scenario. It is very common, for example, for street segments to be
automatically coded as having a complete 100-block of street numbers,
for example, from 401 to 499 on the left side and from 400 to 498 on
the right side. Figure 3E shows the location of street number 441, while
Figure 3F shows the result of geocoding this address using a widely used
online geocoding tool. In this case, the street number is geocoded at
approximately 40% from the start of the street segment, while the actual
location is much further down the street. The placement of geocoded loca-
tions towards one side of the street as a result of incorrect address ranges is
known as the ‘squeeze’ effect, and is illustrated in more detail in Figure 4.

For longer street segments, varying parcel sizes, and irregular street
number sequencing, the positional errors can become much larger, even
Fig. 3. Street geocoding results relative to actual locations of residences. Figure 3A shows a street segment obtained from local street centerlines overlaid on high resolution digital imagery. Figure 3B shows the parcel boundaries associated with each individual property. Figure 3C shows the address range as coded in the attributes of the street segment as well as the street numbers associated with each individual parcel. The comparison shows that the address range is a close approximation of the actual street numbers. Figure 3D illustrates where the parcel addresses are placed using street geocoding using a small side offset. The connecting lines show which location corresponds to which parcel and the results suggest small positional error. Figure 3E highlights the location of one particular address that was also geocoded using several alternative techniques. Figure 3F shows the geocoded location for number 441 using a common online geocoding service. Results reveal a substantial displacement along the street segment due to an incorrect address range, that is, 401–499 instead of 401–453.
when the address ranges coded in the reference data are correct. Additional data sources can be used to obtain knowledge about the number of parcels and their distribution to overcome the parcel homogeneity assumption (Bakshi et al. 2004). However, when reliable parcel data are available, it may be more effective to geocode using the actual parcels instead of the street segments.

Fig. 4. The ‘squeeze’ effect of street geocoding. As a result of incorrect address ranges, all the street geocoded locations are displaced towards one side of the street relative to the location of the actual residence.
In addition to displacement along street segments, positional error in the street reference data can introduce additional error. The widely employed TIGER data from the US Census Bureau, for example, has been recognized as having relatively poor positional accuracy. The original TIGER data were created from a variety of sources, including the US Geological Survey’s 1:100,000 scale map series, and the occurrence of relatively large positional errors in TIGER data has been widely recognized (e.g. O’Grady and Goodwin 2000; Trainor 2003). Figure 5 shows a typical example. A major effort to enhance the quality of TIGER data was initiated in 2002. The Master Address File (MAF)/TIGER Enhancement Program is expected to result in substantial improvements in the positional accuracy of the TIGER data (US Census Bureau 2006).

Several empirical studies in recent years have determined the positional accuracy of geocoding. Table 1 provides a summary of the studies that have been identified as part of this review. What follows is a discussion of
Table 1. Summarized findings from studies on the positional error of street geocoding.

<table>
<thead>
<tr>
<th>Study</th>
<th>Geographic area</th>
<th>Sample size</th>
<th>Geocoding method</th>
<th>Side offset distance</th>
<th>Error determination</th>
<th>Error statistics</th>
<th>Other notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bonner et al. (2003)</td>
<td>Erie and Niagara Counties, NY</td>
<td>200 total,</td>
<td>ArcView 3.2, data not specified</td>
<td>Not reported</td>
<td>GPS, street in front of residence</td>
<td>Median: 32 m for urban, 52 for non-urban</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 urban and 100 rural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cayo and Talbot (2003)</td>
<td>Five counties in upstate NY</td>
<td>3,000 total, 1,000 urban, 1,000 suburban and 1,000 rural</td>
<td>MapMarker Plus 6.0 using GDT data</td>
<td>Various, in 5 m increments</td>
<td>Aerial photography, center of residence</td>
<td>Median: 38 m (urban), 78 m (suburban) and 201 m (rural); 75th percentile: 62 m (urban), 158 m (sub-urban) and 498 m (rural); 95th percentile: 152 m (urban), 421 m (suburban) and 2,872 m (rural).</td>
<td>Direction of the error was found to be not significantly different from uniform distribution</td>
</tr>
<tr>
<td>Dearwent et al. (2001)</td>
<td>Jefferson County, AL</td>
<td>10,026</td>
<td>Software not specified, local street centerlines</td>
<td>Not reported</td>
<td>Parcel centroids</td>
<td>Mean: 75 m</td>
<td></td>
</tr>
<tr>
<td>Karimi and Durcik (2004)</td>
<td>Allegheny County, PA</td>
<td>93</td>
<td>LocMatch (technique 1), ArcView 3.2 (technique 2) and a commercial firm (technique 3) all using 2000 TIGER/Line files</td>
<td>0 m</td>
<td>GPS, center of the street in front of building</td>
<td>Median: 57 m (technique 1), 57 (technique 2) and 42 (technique 3)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>Geographic area</th>
<th>Sample size</th>
<th>Geocoding method</th>
<th>Side offset distance</th>
<th>Error determination</th>
<th>Error statistics</th>
<th>Other notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratcliffe (2001)</td>
<td>Sydney, Australia</td>
<td>21,890</td>
<td>MapInfo 5.5 using StreetWorks 5.0 data</td>
<td>10 m</td>
<td>Parcel centroids</td>
<td>Mean: 47 m, 31 m after trimming 5% outliers</td>
<td>Median error not reported</td>
</tr>
<tr>
<td>Schootman et al. (2007)</td>
<td>St. Louis, MO</td>
<td>261</td>
<td>ArcGIS 8.3 using 2000 TIGER/Line files (technique 1) and commercial firm (technique 2)</td>
<td>17 m</td>
<td>Aerial photography, front door</td>
<td>Median: 31 m (technique 1) and 26 m (technique 2); 75th percentile 55 m (technique 1) and 49 (technique 2)</td>
<td>Direction of the error was found to be significantly different from uniform distribution, but no explanation was found</td>
</tr>
<tr>
<td>Strickland et al. (2007)</td>
<td>Atlanta, GA</td>
<td>599</td>
<td>Agency in-house (technique 1) and commercial firm (technique 2)</td>
<td>Not reported</td>
<td>Aerial photography, center of residence</td>
<td>Median: 91 m (technique 1) and 71 m (technique 2); 75th percentile: 155 m (technique 2) and 147 m (technique 2); 95th percentile: 369 m (technique 1) and 352 (technique 2)</td>
<td>Direction of the error was found to be not significantly different from uniform distribution</td>
</tr>
<tr>
<td>Ward et al. (2005)</td>
<td>Iowa, IA</td>
<td>234</td>
<td>ArcView 3.2 using 2000 TIGER/Line files (technique 1) and commercial firm (technique 2)</td>
<td>13 m</td>
<td>GPS, outside of home</td>
<td>Median: 56 m (urban) and 88 m (rural) (technique 1); 50 m (urban) and 212 m (rural) (technique 2)</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Geographic area</td>
<td>Sample size</td>
<td>Geocoding method</td>
<td>Side offset distance</td>
<td>Error determination</td>
<td>Error statistics</td>
<td>Other notes</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------</td>
<td>-------------</td>
<td>-------------------------------------------------------</td>
<td>----------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Whitsel et al.</td>
<td>48 US States</td>
<td>3,615</td>
<td>Four commercial firms</td>
<td>0, 2, 12 and 15 m</td>
<td>Varies with types of address; parcel centroids for residential addresses</td>
<td>Median: 163 m (technique 1), 117 m (technique 2), 121 m (technique 3) and 115 m (technique 4)</td>
<td>70% of the original address consisted of EPA air quality system monitors, which may not be representative of residential addresses; 1,050 residential addresses were located in five counties in NC</td>
</tr>
<tr>
<td>Zandbergen</td>
<td>Orange County, FL</td>
<td>104,865</td>
<td>ArcGIS 9.1 using local street centerlines</td>
<td>8 m</td>
<td>Parcel centroids</td>
<td>Median: 41 m; 95th percentile: 137 m</td>
<td></td>
</tr>
<tr>
<td>Zhan et al.</td>
<td>Austin and San Antonio, TX</td>
<td>200</td>
<td>ArcGIS 9.1 using StreetMap USA (technique 1) and Centrus Geocoder for ArcGIS (technique 2)</td>
<td>8 m</td>
<td>GPS, edge of street in front of door</td>
<td>Median: 69 m (technique 1) and 48 m (technique 2); 75th percentile: 110 m (technique 2) and 73 m (technique 2)</td>
<td></td>
</tr>
<tr>
<td>Zimmerman et al.</td>
<td>Carroll County, IA</td>
<td>1,423 (rural only)</td>
<td>ArcGIS 9.1 using 2000 TIGER/Line files</td>
<td>0 m</td>
<td>Aerial photography, center of residence</td>
<td>Median: 168 m</td>
<td>Direction of positional error was found to be primarily in the main directions of streets</td>
</tr>
</tbody>
</table>
the major points. Studies have mostly been limited to the USA; the study by Ratcliffe (2001) of Sydney, Australia, is a notable exception. Most studies have also been limited to relatively small geographic areas, typically one or several counties. While these two observations suggest that the generalizability of the results may be limited, a number of findings are very similar across different geographic regions within the USA.

Geocoding techniques vary among the studies, with some employing multiple methods for comparison. Both in-house geocoding and the use of commercial vendors are common. Different measures are employed to determine the ‘true’ location of the residential address. The most accurate methods rely on direct observation of the location of the actual building that corresponds to the address. Several studies have accomplished this using field measurements with a GPS unit (Bonner et al. 2003; Karimi and Durcik 2004; Ward et al. 2005; Zhan et al. 2006). Other studies using direct observations have employed high resolution aerial photography, typically in the form of color imagery at 1-m resolution or better (Cayo and Talbot 2003; Schootman et al. 2007; Strickland et al. 2007; Zimmerman et al. 2007). The use of GPS and aerial photography is time-consuming and has therefore been used in studies with relatively small sample sizes. Studies using larger sample sizes have relied on more automated techniques, in particular the use of parcel centroids (Dearwent et al. 2001; Ratcliffe 2001; Zandbergen 2007; Whitsel et al. 2006).

All of the three techniques (GPS, aerial photography, and parcel centroids) have some inherent positional error, but the use of parcel centroids is conceptually the least reliable, in particular for larger parcels, because the actual residence may not be located at the center of the parcel. Cayo and Talbot (2003) compared parcel centroids and aerial photography and found relatively small differences in urban areas [root mean square error (RMSE) = 18 m] but very large differences in rural areas (RMSE = 211 m). This suggests that where possible aerial photography or methods with similar accuracy should be employed to determine the positional error in geocoding.

In terms of the magnitude of the positional error, studies vary greatly in how the error values are reported. Median is most commonly used, but mean, standard deviation, inter-quartile ranges, percentiles of the distribution, RMSE and others are also used. This presents complications when trying to compare findings from different studies, as it is firmly established that the distribution of the positional error of street geocoding does not follow a normal distribution (Cayo and Talbot 2003; Karimi and Durcik 2004; Whitsel et al. 2006; Zandbergen 2007). Figure 6 shows a typical distribution of the positional accuracy of street geocoding. Formal testing by Zandbergen (2008b) has shown that the distribution of the magnitude of positional error approximates a log-normal distribution when the distribution of the direction of the error is uniform. Zimmerman et al. (2007) have shown that mixtures of bi-variate distributions with two or three components are required to characterize the distribution of the
The magnitude of positional error when the distribution of the direction is strongly influenced by the gridded nature of the street network.

The non-normal nature of the error distribution limits the usefulness of commonly used statistics, such as mean, standard deviation and RMSE. Table 1 therefore reports the median values and any percentiles where available. Based on the median value estimates, the 'typical' positional error for residential addresses ranges from 26 to 201 m. This is a very broad range and much of this can be attributed to differences across urban-rural gradients, as results in urban areas are more accurate than in rural areas. For example, Cayo and Talbot (2003) found a median error of 38 m for urban areas, 78 m for suburban areas, and 201 m for rural areas. Bonner et al. (2003) and Ward et al. (2005) found similar differences, confirming a clear general trend that geocoding is much more accurate in urban areas compared to rural areas. This is a result of longer street segments, larger parcels, and greater variability in parcel size along a street segment in rural areas.

In addition to the magnitude of the positional error, several studies have examined the direction of the error, that is, the angle in degrees of the line connecting the actual residence with the geocoded location. Studies reveal mixed results with several finding no significant difference from a uniform distribution (Cayo and Talbot 2003; Strickland et al. 2007), while...
others did find a significant difference (Schootman et al. 2007; Zimmerman et al. 2007). Few studies, however, present reasoning as to why the distribution would be uniform or not. As displacement along the street segment is typically a major component in the positional error of street geocoding, logic suggests that any direction in the error will be driven by the orientation of street segments within the study area. For large study areas, this distribution is likely to be uniform, but for specific regions this may not be the case. A good example of this can be seen in the results of Zimmerman et al. (2007), which demonstrate that many of the larger positional errors are in either the X direction (90 or 270 degrees from North) or the Y direction (0 or 180 degrees) and much less in any other direction. This is not very surprising given that the sample consisted of rural addresses in Carroll County, Iowa, where the road network forms a near perfect grid of segments running East–West and North–South.

Because the placement of geocoded locations is logically constrained by the street reference network, at fine spatial scales the direction of positional error will be strongly influenced by the direction of the street segment. This is most strongly observed in the ‘squeeze effect’ discussed previously in which geocoded locations are displaced to one side of the street due to incorrect address ranges. In combination with the pattern in the orientation of street segments, this results in non-random positional errors (Zimmerman et al. 2007), which has been shown to introduce substantial analysis bias (Zandbergen 2007). Despite the evidence of non-random behavior, error propagation models for geocoding have employed uniform distributions for direction (Strickland et al. 2007; Whitsel et al. 2006).

One point of contention in the literature is the effect of offsets on positional accuracy. Two types of offsets can be used in street geocoding: (i) side offsets, which place the geocoded point at some specified distance from and perpendicular to the street centerline, and (ii) end offsets, which squeeze the geocoded points away from the endpoints of the segment. Side offsets are meaningful as they avoid the placement of geocoded locations in the middle of the street and may place them closer to the actual residences. Figure 7 illustrates the effect of a perpendicular side offset on the positional accuracy of geocoding. End offsets are meaningful as they avoid the placement of geocoded locations at intersections and may place them within a part of the street segment that is more likely to be in front of actual residences. Not all studies, however, report the offsets used, and some only employ one or the other. Some software also do not allow the offsets to be modified and may have a hard-coded value for these parameters. Commercial vendors also typically do not allow for modification of offsets and employ a single value, if any.

Among the studies in Table 1 that reported side offset parameters, values range from 0 to 17 m, often without specific justification for their selection. Cayo and Talbot (2003) varied side offset values in 5 m increments and determined that 15 m was the optimal value for their sample. However,
Fig. 7. Illustrative examples of the effect of perpendicular side offset values on the positional accuracy of street geocoding. Figure 7a shows the scenario where the street geocoded location is directly in front of the correct property, and the use of a properly selected offset will place the geocoded location in closer proximity (or even within), the correct parcel and closer to the actual residence. When the street geocoded location is slightly displaced along the street segment as in Figure 7b, the use of an offset still results in improved positional accuracy, but the gain is relatively minor. When displacement along the street segment is substantial (Figure 7c), the use of an offset has very limited benefit and may in fact result in decreased positional accuracy. When geocoded locations are placed on the wrong side of the street segment (Figure 7d), the use of an offset typically decreases positional accuracy, even if the location is in very close proximity to the correct parcel.
the effect on the positional error distribution was found to be minimal. Zandbergen (2007) varied offset values in 10 m increments and offset values of 10, 20, and 30 m were found to have nearly identical error distributions. These results suggest that the effect of using a side offset perpendicular to the street segment is very small relative to the error resulting from the incorrect placement along the street segment. Very few studies report the use of end offsets, making it more difficult to discern its possible effects on positional accuracy. A further complicating factor is that end offsets can be specified as an actual distance (e.g., 20 m) or a relative portion of the street segment (e.g. % of the length). Cayo and Talbot (2003) varied end offset values in 5 m increments and determined that 50 m was the optimal value for their sample. However, the effect on the positional error distribution was found to be minimal.

A final consideration is the potential trade-off between match rate and positional accuracy. Logically, it can be expected that a less-than-perfect match has a high probability of being placed inaccurately and it has therefore been argued that increasing the match rates by loosening matching standards may result in higher positional error (Rushton et al. 2006). However, there has been surprisingly little empirical evidence to support this and this clearly remains a topic where further research is warranted.

While the positional error of street geocoding is substantial, it is typically much more accurate than the alternative of assigning an address to a geographic unit, such as a postal code. In the USA, for example, five-digit ZIP codes typically contain several thousand addresses and the centroid of this polygon may be several hundred meters, or even a few kilometers, away from the actual residence. In several other jurisdictions, the postal code system is much finer grained and can provide a fairly accurate location. For example, Canada uses a 6-character postal code that typically corresponds to a single block-face in urban areas (Statistics Canada 2002). An empirical validation study by Bow et al. (2004) determined that for a sample of addresses in the City of Calgary, 87.9% of postal code locations were within 200 m of the true address location and 96.5% were within 500 m. While not as accurate as typical street geocoding in urban areas, postal codes tend to be much more reliable than other parts of the address information, resulting in very high match rates. In some cases, a slightly higher positional error may be acceptable if very high (and less biased) match rates are desired.

5 Concordance with Geographic Units

For many applications, the purpose of geocoding addresses is to associate the individual location with demographic and other socio-economic variables. This is most commonly accomplished by establishing the census enumeration unit in which the address is located and deriving variables from the census data. This can be achieved using two methods: (i) lookup
tables from the census agencies; and (ii) point-in-polygon overlays comparing geocoded point locations and polygon boundaries of census enumeration areas. In the USA, the Census Bureau itself employs lookup tables and suggests that point-in-polygon overlays may not provide accurate results, in part due to inaccuracies in the cartographic boundary files, in part due to errors in geocoding. The TIGER data do not provide sufficient accuracy to be able to rely on point-in-polygon overlays (Rushton et al. 2006). As a result, individual addresses may be assigned the incorrect census enumeration unit if the point-in-polygon method is used. The ‘gold standard’ in this case is the enumeration unit assigned by the lookup tables provided by the US Census Bureau. Unfortunately, the use of point-in-polygon overlays in common in many commercial geocoding programs and has in fact been used (incorrectly) in many studies.

In an evaluation of commercial firms, Krieger et al. (2001) found that 4% of geocoded locations were assigned the incorrect census block group. Schootman et al. (2007) found that less than 1% of geocoded locations were assigned the incorrect block group or track when using TIGER roads as the reference network. When addresses were geocoded to the actual location of the residence using aerial photography, however, 16% were assigned the incorrect block group and 7% the incorrect tract. This confirms that data from the US Census Bureau are internally consistent, that is, lookup tables and geocoding based on TIGER line files provide very similar results. However, combining TIGER data with other data of very high positional accuracy can lead to substantial errors. Ratcliffe (2001) determined that 7.5% of geocoded locations fell into incorrect census enumeration units based on point-in-polygon comparisons for parcel and street geocoded locations. Strickland et al. (2007) employed an error propagation technique in which random address locations were displaced based on empirically derived positional error distributions. Using point-in-polygon comparisons, approximately 5% of locations were placed in the wrong census tract. Kravets and Hadden (2007) determined that approximately 5% of geocoded locations fell into incorrect census enumeration units using a comparison of census lookup tables and point-in-polygon analysis. The percentage of incorrectly placed locations was substantially higher in rural areas, confirming that geocoding quality also varies along urban-rural gradients in terms of concordance with geographic units.

6 Repeatability

The repeatability of geocoding has not received as much attention as match rates and positional accuracy. In one recent study by Whitsel et al. (2006) using a large sample \(n = 3,615\) of addresses in 49 US states, substantial differences were found between four commercial vendors in address match rate (30 to 90%), concordance between established and vendor-assigned census tracts (85 to 98%), and distance between established

© 2009 The Author
Journal Compilation © 2009 Blackwell Publishing Ltd
and vendor-assigned coordinates (mean of 228 to 1,809 m). This confirmed earlier findings by Whitsel et al. (2004) for a smaller sample that the repeatability of commercial geocoding is not very good.

Zhan et al. (2006) compared two geocoding methods (ArcGIS using StreetMap USA versus Centrus Geocoder for ArcGIS) for a small sample (n = 200) and found statistically significant differences in both match rates (79 versus 89%) and positional accuracy (median of 69 m versus 48 m). Ward et al. (2005) compared two geocoding methods (ArcView 3.2 using TIGER 2000 data and commercial firm) and found no significant difference. In a comparison of three geocoding algorithms (LocMatch, ArcView 3.2 and a commercial firm) using the same TIGER 2000 reference data, Karimi and Durcik (2004) found that the differences between the results were not significant. Yang et al. (2004) compared a deterministic matching algorithm (ArcView 8.3 StreetMap USA using TIGER 2002 data) and a probabilistic algorithm (Automatch using TIGER 1998 data) and found substantial disagreement in terms of the sample of addresses that produced a match and the census enumeration units the geocoded locations were placed in (based on point-in-polygon overlay). Schootman et al. (2007) found very similar match rates and positional accuracy in a comparison of in-house geocoding using GIS software and a commercial vendor, both using TIGER 2000 data. While the positional error for each set of geocoded locations was found to be substantial, the difference in location between the two sets was much smaller, suggesting good repeatability when using the same reference data.

Bichler and Balchak (2007) have documented several specific limitations to the repeatability of geocoding, including the accuracy of the reference database, choice of GIS software, matching algorithms employed, and user-selected settings in the geocoding process. Studies by Karimi and Durcik (2004) and Schootman et al. (2007), however, suggest that the nature and accuracy of the reference data is the critical factor that explains a lack of repeatability. Results from studies that have employed commercial firms are much more difficult to interpret in this regard, as the source of the reference data and/or the geocoding algorithms is not revealed in sufficient detail.

7 Effects of Geocoding Quality on Spatial Analysis

The research on the quality of geocoding suggests that the errors in geocoding can be very substantial and need to be characterized in a meaningful manner relevant to the use of the geocoding results. Strong evidence has been found that the errors in geocoding are not random in nature and may introduce bias in terms of both completeness and positional accuracy.

Contrary to other forms of digital spatial data (e.g., land use, roads, and census boundaries), geocoding results do not have an implicit scale, and
hence its spatial resolution is not known without some degree of testing. Certainly, the scale (i.e., spatial resolution) of geocoded locations is not the same as the scale of the street reference data employed in geocoding. Based on a ‘typical’ positional error distributions for large sample sizes, such as those reported by Cayo and Talbot (2003) and Zandbergen (2007), the 90th percentile of the error is approximately 100 m for urban areas and much larger for rural areas. Based on this 90th percentile, typical street geocoding in urban areas would not meet the accuracy standards for a 1:100,000 scale map based on the National Map Accuracy Standards (US Bureau of the Budget 1947). The use of street reference data of very high positional accuracy does not result in geocoded locations of similar accuracy. This has received very limited attention in the literature and may come as a surprise to analysts who commonly employ geocoding. Geocoding is notably absent from any standard on the positional accuracy of spatial data, including the National Standard for Spatial Data Accuracy (FGDC 1998). The Street Address Data Standard currently under development (FGDC 2005) includes only a very small section on positional accuracy. Logically, any effort that improves the logical consistency and attributes (in particular address ranges) of street reference data will contribute to improved positional accuracy, but the unique nature of the errors of geocoding is notably absent from any current or proposed standard.

The errors in geocoded locations may adversely affect spatial analytic methods and this has started to receive attention in the literature. Specific effects include inflation of standard errors of parameter estimates and a reduction in statistical power to detect such spatial features as clusters and trends (Jacquez and Waller 2000; Waller 1996; Zimmerman 2008). Most empirical studies have quantified these effects by examining the degree to which geocoded locations are misclassified in subsequent spatial analysis, although error propagation modeling through simulation has also been employed.

Research on this topic has been mostly confined to the health field. For example, typical street geocoding was not sufficiently accurate for the analysis of exposure to traffic-related air pollution of children at short distances of 250 to 500 m (Zandbergen 2007; Zandbergen and Green 2007) in Orange County, Florida. Positional errors in street geocoding were found to be non-random in nature and introduced substantial bias and error in exposure classification. Street geocoding was found to consistently overestimate the number of potentially exposed children at small distances up to 250 m as a direct result of the ‘squeeze’ effect. False positives and negatives were also found to be very common at these small distances. A similar study by Whitsel et al. (2006) on traffic-related air pollution also found exposure misclassifications, but to a lesser extent and without any sign of bias towards overestimation. This can partly be attributed to the fact that the error propagation modeling employed by Whitsel et al. (2006) utilized a uniform distribution for the direction of positional
error that may have underestimated the effects of geocoding errors. Wu et al. (2005) found that the use of road networks of poor positional accuracy also lead to substantial misclassification in an assessment of exposure to vehicle emissions. The scenario of exposure to vehicle emissions, however, is inherently very sensitive to geocoding errors, because the distances of concern are very short and the source itself (major roads with high traffic counts) is part of the geocoding process. Other exposure scenarios considering larger distances and/or different types of scenarios are less likely to result in misclassification. For example, Ward et al. (2005) found that geocoding errors affected exposure classification based on distance to crops field at 100 m, but not at greater distances. Zhan et al. (2006) found that difference in match rate and positional accuracy of two geocoding methods did not alter exposure classification using a 1500-m buffer around Toxic Release Inventory facilities. Burra et al. (2002) found that relatively small errors in geocoding resulted in significantly different mortality clusters using local indicators of spatial autocorrelation. Mazumdar et al. (2008) determined the influence of geocoding error on the statistical power of the relationship between environmental exposure and health. Power analyses showed that the quality associated with different geocoding processes affected the ability to recover the relationships. Griffith et al. (2007) found a noticeable but not shocking large effect of the positional error of geocoding on spatial regression analysis. Combined, these findings suggest that the effect of the positional error in geocoding is primarily a function of the nature of the spatial analysis that employs the geocoding results, in particular: (i) magnitude of the distances considered and (ii) type of spatial analysis technique employed (i.e., spatial regression, global clustering, and local clustering).

Incomplete geocoding can also limit the utility of geocoded locations in subsequent analysis. Most studies implicitly assume that the ungeocoded records represent an unbiased sample of the original records, but selection bias in terms of ethnicity (Gilboa et al. 2006) and across urban–rural gradients (Oliver et al. 2005) has been documented. This bias confounds the analysis of the association between the characteristics of the individuals or incidents that have been geocoded and area-level socio-demographic factors.

Traditional methods to handle missing data are not very appropriate to account for incomplete geocoding due to the selection bias across urban–rural and socio-economic gradients. However, addresses that cannot be geocoded to the street level can often be geocoded to a coarser scale, in particular to the postal code level. For example, an epidemiological study may utilize the postal code centroid of an address that could not be geocoded to the street level to determine an association with socio-economic variables based on the census tract in which the postal code centroid is located. Such methods are referred to as geographical imputation methods. Henry and Boscoe (2008) have shown how geographical imputation can
be improved by using the distribution of socio-economic characteristics within a postal code. Results from the study by Henry and Boscoe (2008) revealed no serious drawbacks to the use of geographic imputation, although they cautioned for the possibility that it might introduce geographic bias. Shi (2007) on the other hand found that geographic imputation in rural areas with a high incidence of PO Boxes greatly reduced the ability to reliably detect disease clusters. Zimmerman (2008) has demonstrated how the additional information at a coarser scale can be used to improve the estimation of local clusters resulting from reduced selection bias caused by incomplete geocoding, although the methodology is limited to first-order properties of a point pattern. Further refinements to methods that employ both fine and coarse scale data would be a welcome addition to the set of analytical tools available to researchers, as are spatial-analytical techniques that recognize the inherent uncertain nature of point events (e.g., Grubesic 2006).

There have also been a number of studies in the crime literature on the effects of geocoding quality on the results of spatial analysis. Crime hotspot detection in particular is very sensitive to errors in geocoding due to its heavy reliance on individual locations and sensitivity to sample size (Bichler and Balchak 2007). Aggregation to small administrative units has therefore been recommended (Bichler and Balchak 2007), but this severely limits the applicability and statistical power of several analysis techniques. Brimicombe et al. (2007) demonstrated for a large metropolitan area in the UK that different geocoding match rates for the same crime incident database revealed distinct kernel density hotspots, although no statistical comparisons were made. Harada and Shimada (2006) compared kernel density surfaces derived from geocoded crime locations of different positional accuracy for Tokyo, Japan. Hotspots appeared relatively robust to geocoding errors, although this can partially be attributed to the large bandwidth used (500 m). Zandbergen and Hart (2009) determined that typical street geocoding is not sufficiently accurate to reliably determine residency restrictions for registered sex offenders around schools and daycares.

The degree to which positional errors in geocoding affect the reliability of subsequent spatial analysis will obviously depend on the nature of this analysis. The first dimension is the scale or spatial resolution of the analysis. For example, analysis of exposure to particulate matter from vehicle exhaust or exposure to pesticides applied to cropland may consider very short distances, in the order of several hundred meters, while other types of analysis may be concerned with distances of several kilometers. Several studies have demonstrated a lack of reliability at short distances (Mazumdar et al. 2008; Zandbergen 2007; Zandbergen and Green 2007). The second dimension is the nature of the specific spatial analysis technique. For example, cluster analysis using kernel density estimation has shown to be relatively robust to geocoding errors, while other types of cluster analysis are very sensitive to these errors. As with any spatial-analytical technique, a
good understanding of both the input data quality and knowledge of the robustness of the technique itself is required to determine the reliability of the analysis result. Some progress has been made in terms of the development of error propagation models to test the sensitivity of spatial-analytical techniques to geocoding errors, but this remains an area of very active research.

8 International Comparisons

Much of the published literature on geocoding is limited to the USA, and to a lesser extent Australia, Canada, and the UK. The most widely employed address model in the USA uses street networks as reference data. This is also reflected in commercial GIS products that are often limited to using address formats used in the USA or Western Europe. This is changing gradually as the market for commercial GIS products has become a global market, but many Asian, African, and Latin American countries are notably absent from the list of countries supported by geocoding software. The findings from existing studies, therefore, may not be generalizable to other jurisdictions and other addressing systems.

Addressing systems vary greatly among jurisdictions, including irregular or non-metric numbering, different criteria for naming streets, association of address numbering with regions instead of streets, and others. For example, in Japan address numbers along a street are not assigned consecutive numbers, but instead based on their date of construction. Many streets in Japan in fact have no name at all and instead a hierarchical system of areas is employed consisting of districts, subdistricts (chome), blocks (ban), and entrance locations (gou). A similar hierarchical system is used in Korea where numbers are assigned inside neighborhoods (dong) within urban sectors (gu). A detailed characterization of many international addressing systems can be found in Rhind (2004). These international differences present challenges to the development of geocoding software and the evaluation of geocoding quality across jurisdictions. No comparative studies on geocoding quality in different (international) jurisdictions could be identified as part of this review, and this appears to be an obvious gap in the existing knowledge on geocoding.

9 Improving Geocoding Quality

Many different strategies have been employed to improve geocoding quality. Efforts to improve the algorithms used in the matching process include the use of multi-stage geocoding methods that employ both national and local reference data (Lovasi et al. 2007) or iterative cleaning procedures (Brimicombe et al. 2007; McElroy et al. 2003). Progress has also been made in terms of trying to improve the record linkage process that underlies the matching algorithms in geocoding by refining probabilistic approaches.
For example, Churches et al. (2002) present a probabilistic method for standardization using Hidden Markov Models that has been implemented in the Geocoded National Address File for Australia (Christen et al. 2004). Improvement to the widely used Soundex method for indexing information based on how word sounds have also been proposed (Christian 1998).

Improving the positional accuracy of geocoding can be accomplished by developing more accurate reference data and by employing higher resolution address data models. Reference data can be improved in terms of both positional accuracy and attribute accuracy, in particular more reliable address ranges. The TIGER data maintained and distributed by the US Census Bureau are the most widely employed reference data for geocoding in the USA. In 2002, the MAF/TIGER Accuracy Improvement Project was initiated (US Census Bureau 2006). MAF is designed to be a complete and current list of all addresses and locations where people live or work, covering an estimated 115 million residences, as well as 60 million businesses and other structures in the USA. While the public will not have access to this data, the project is expected to result in substantial improvements to the TIGER line files that are used in street geocoding. Improved TIGER data for many Counties have become available in the last several years, but limited empirical validations have been published so far.

The positional accuracy of geocoding can also be improved by employing address point and parcel geocoding instead of street geocoding. Current limitations in this regard are limited awareness of and experience with these methods, varying data availability, and relatively poor performance in terms of match rates in the case of parcels. If historic trends in street geocoding can be used as a guide, address point geocoding will likely become more widespread and more affordable.

Another area where additional research efforts are needed is the development of measures that characterize the degree of reliability of geocoded locations, similar to those employed in the creation of data using photogrammetric techniques (e.g., ASPRS 1990) or elevation models from various sources (ASPRS 2004; NDEP 2004). The Geocoding Certainty Indicator proposed by Davis and Fonseca (2007) provides a promising conceptual framework for such a standard and includes measures for parsing, matching, and location certainty. Such a measure of reliability could be used to filter geocoding results based on what level of reliability is required for a particular application or the measure of reliability could be used as a weight in spatial-analytical techniques. However, the exact design and implementation of this indicator of reliability has not been realized yet. The point-radius method developed by Wieczorek et al. (2004) for georeferencing natural history collections can serve as a good example of how to quantify the degree of reliability of geocoded locations. While Wieczorek et al. (2004) dealt primarily with descriptive information that
does not contain actual street addresses, some of the same principles can be applied to street geocoding.

Characterizing the quality of geocoding using standards for completeness, positional accuracy, and/or concordance with geographic units is no easy task, but perhaps a necessary one. Based on the most rigorous studies of positional accuracy to date, typical street geocoding of residential address in urban areas does not meet the accuracy standards for a 1:100,000 scale map. This is much less accurate than is typically assumed and it may defeat the purpose of developing spatial analysis models of very high resolution that use geocoded locations as input. Unfortunately, the non-random nature of geocoding error (selection bias in unmatched records, positional bias, and non-normal distributions of positional error) complicates the development of such standards and alternative statistical frameworks will need to be developed in this regard.

More complex data models may also be required to effectively address the challenges of very high density urban areas, including apartment buildings and multiple-use complexes. While conceptual models and prototypes for 3D geocoding have been proposed (e.g., Lee 2004), these have yet to see widespread implementation.

10 Changing Paradigm

Traditionally, address geocoding has been the domain of researchers and GIS professionals. However, this paradigm is rapidly changing due to advances in the availability of digital data and online geocoding tools. Arguably, geocoding has become one of the most successful commercial applications of GIS as evidenced by the widespread use of online mapping tools, such as MapQuest, Yahoo Maps, Google Maps, and Microsoft Live Earth. While most of these online geocoding services typically only allow for the geocoding of a single address at one time, their functionality can often be accessed using an application programming interface that allows for automation. Very little has been published on the quality of the geocoding results obtained from online geocoding services. The reference data underlying most of the well-established services are obtained from commercial data providers (e.g., TeleAtlas or Navteq), while others rely on freely available TIGER data. The specific geocoding algorithms, however, are often proprietary and remain largely unknown. While there is no published evidence to suggest that geocoding quality from online services is any better or worse than geocoding using commercial GIS software, empirical testing of online geocoding services seems appropriate to confirm whether the substantial body of knowledge on geocoding can be extended to online geocoding tools.

A further challenge to the current paradigm is presented by the emergence of volunteered geographic information (Elwood 2008; Flanagin and Metzger 2008; Goodchild 2008). At present, few of the efforts in this area...
specifically provide geocoding services, but this is likely to change, particularly in jurisdictions where access to low-cost digital data has typically been poor. Perhaps more importantly, increasing amounts of information of varying quality are likely to be distributed in geocoded format, presenting new challenges to metadata management and QA/QC procedures. This will present additional difficulties for spatial analysis as the input data may be compiled from different sources of varying and sometimes unknown quality. This further demonstrates the need to better understand the robustness of spatial-analytic methods to error and uncertainty in the input data.

The widespread availability of geocoding of individual addresses has also lead to concerns over privacy when the locations of individuals and/or households are made public as published maps, either in digital or paper format (Armstrong and Ruggles 2005; Curtis et al. 2006). A technique known as ‘reverse geocoding’ can be used to identify the addresses associated with the published locations. Individual-level data can therefore often not be released due to privacy concerns and/or legal requirements. Techniques to preserve geospatial privacy exist and are collectively referred to as ‘geographic masking’ (Armstrong et al. 1999; Kwan et al. 2004). While geographic masking methods are relatively well-understood, very little research has been done on how good reverse geocoding actually is, and as a result there is limited guidance on the nature and magnitude of geographic masking necessary to effectively protect geospatial privacy.

11 Conclusions

Street geocoding is currently the most widely employed technique to assign locations to individual addresses in the USA. Errors introduced by street geocoding include incompleteness, positional error, and incorrect assignment to geographic units. A review of empirical studies suggests that these errors are neither small nor random in nature, and that substantial bias may be introduced in spatial analysis that employs the results of geocoding. Specific effects include inflation of standard errors of parameter estimates and a reduction in statistical power to detect such spatial features as clusters and trends.

Findings from empirical studies suggest that the effect of geocoding quality on spatial analysis depends strongly on the specific nature of the analysis technique, including the scale or spatial resolution. The ‘typical’ positional error of street geocoding in urban areas is 100 m based on the 90th percentile of the error distribution and this can have a strong effect on spatial analysis at short distances. Local cluster detection techniques, for example, have been shown to be fairly sensitive to such errors. Broader scale analyses on the other hand, such as spatial regression and global clustering, are less affected. As with any spatial-analytical technique, a good understanding of both the input data quality and knowledge of the
robustness of the technique itself is required to determine the reliability of the analysis result. Some progress has been made in terms of the development of error propagation models to test the sensitivity of spatial-analytical techniques to geocoding errors, but this remains an area of very active research. Further refinements to methods that recognize the inherent uncertain nature of point events would also be a welcome addition to the set of analytical tools available to researchers.

One of the more persistent problems is incomplete geocoding in rural areas resulting from PO Boxes. Traditional methods to handle missing data are not very appropriate to account for this due to the selection bias across urban–rural and socio-economic gradients. Refined geographic imputation measures have shown promise in this regard.

The empirical studies reviewed as part of this article are mostly limited to the USA, reflecting a strong bias in the literature. While a handful of examples have been published on geocoding in other jurisdictions, no comparative studies on geocoding quality in different (international) jurisdictions could be identified, and this appears to be an obvious gap in the existing knowledge on geocoding.

Alternative address models have emerged and are gradually becoming more widely used. Address points in particular appear very promising as an address data model for geocoding, as they represent excellent positional accuracy and produce match rates similar to those for street geocoding.

Online geocoding services and volunteered geographic information present new challenges and the view that geocoding is mostly the domain of researchers and GIS professionals is no longer appropriate. A better understanding of the quality of the results from such services would be beneficial, as well as the recognition that a multi-method approach to geocoding may be required to achieve the best result.

Several areas for future research on geocoding have been identified: (i) refinements of address data models to incorporate complex addressing situations; (ii) development of error propagation techniques to determine the level of geocoding quality required for a particular analysis scenario; (iii) development of measures of reliability for geocoding results; (iv) comparative analysis of geocoding quality across different jurisdictions; and (v) validation of online geocoding services and volunteered geographic information.

Short Biography

Paul Zandbergen is Associate Professor in the Department of Geography at the University of New Mexico. He obtained his PhD in Resource Management and Environmental Studies at the University of British Columbia in Vancouver, Canada. Prior to coming to the University of New Mexico, he was Assistant Professor in the Department of Geography at the University of South Florida. He is a Geographic Information scientist.
with interests in both the fundamentals of GI Science, as well as the applications of geospatial technologies to several fields, including water resources, spatial ecology, environmental health, and criminal justice. His current research has focused on issues of scale, error, and uncertainty in spatial analysis, as well as on the robustness of spatial analytical techniques.

Note

* Correspondence address: Paul A. Zandbergen, Department of Geography, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131, USA. E-mail: zandberg@unm.edu.

References


