Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computers, Environment and Urban Systems 32 (2008) 214–232

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

A comparison of address point, parcel and street geocoding techniques

Paul A. Zandbergen *

Department of Geography, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131, USA

Received 11 January 2007; received in revised form 15 November 2007; accepted 19 November 2007

Abstract

The widespread availability of powerful geocoding tools in commercial GIS software and the interest in spatial analysis at the individual level have made address geocoding a widely employed technique in many different fields. The most commonly used approach to geocoding employs a street network data model, in which addresses are placed along a street segment based on a linear interpolation of the location of the street number within an address range. Several alternatives have emerged, including the use of address points and parcels, but these have not received widespread attention in the literature. This paper reviews the foundation of geocoding and presents a framework for evaluating geocoding quality based on completeness, positional accuracy and repeatability. Geocoding quality was compared using three address data models: address points, parcels and street networks. The empirical evaluation employed a variety of different address databases for three different Counties in Florida. Results indicate that address point geocoding produces geocoding match rates similar to those observed for street network geocoding. Parcel geocoding generally produces much lower match rates, in particular for commercial and multi-family residential addresses. Variability in geocoding match rates between address databases and between geographic areas is substantial, reinforcing the need to strengthen the development of standards for address reference data and improved address data entry validation procedures.

Keywords: Geocoding; Reference data; Address models; Address points; Parcels

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 geocoding process. Much literature has been written on the topic of geocoding and the underlying algorithms that make it function. Questions such as, "What is an acceptable match rate?" (Ratcliffe, 2004) and "How do different algorithms affect the geocoding result?" (Karimi & Durcik, 2004) are readily encountered in the literature. Whatever the application, the primary concern generally relates to the accuracy of a geocoding technique. National datasets and local datasets face different challenges. While large national datasets must contend with a diversity of address formats, requiring more complex rules that define how an address is broken down for geocoding, local datasets require better positional accuracy since the geocoded data is often analyzed in relatively small geographic units. With the increasing power of GIS there has also been an increase in commercial vendors that provide custom geocoding tools and reference data (e.g. NavTech and TeleAtlas). Additionally, web-based address look-up engines such as Google Maps, MapQuest, and Yahoo Maps have become mainstream tools among the general public, making even greater the demand for accurate geocoding.

The general purpose of this paper is fourfold: (1) to review the foundations of the geocoding process; (2) to review address data models used in geocoding; (3) to present a framework for evaluating geocoding quality; and (4) to present the results of an empirical comparison of geocoding match rates using different address data models. The review portion of this paper complements recent reviews of geocoding by Rushton et al. (2006) and

* Tel.: +1 505 277 3105; fax: +1 505 277 3614.
E-mail address: zandberg@unm.edu

0198-9715/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2007.11.006

Goldberg, Wilson, and Knoclock (2007); this paper focuses more on address models and issues of geocoding quality. The empirical component in particular will examine the influence of geocoding techniques relative to the influence of variability in input address quality on geocoding match rates.

2. Geocoding foundations

While simple in concept, geocoding as a process is not as simple as just putting a dot on a map. Techniques involved in geocoding borrow from various disciplines, most notably, information theory, decision theory, probability theory, and phonetics. What follows is a brief review of the fundamental concepts of the geocoding process.

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 is defined as the steps involved in 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. Generally, these steps include parsing the input address into address components (such as street name, street type, etc.), 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.

While geocoding applications are diverse and span many types of applications, 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.

2.2. Probabilistic record linkage

Probabilistic record linkage is the process of matching two data files under conditions of uncertainty. The objective is to identify and link records which represent a common entity whether the entity is an individual, a family, an event, a business, an institution, or an address. Probabilistic record linkage systems use a form of fuzzy logic to score how well records do or do not match. The concept is in contrast to deterministic record linkage which assumes error-free identifying fields and links records that match exactly on these identifying fields. For example, the ability to join database records on matching primary and foreign keys is an example of deterministic linkage. When no error free identifier is shared by all of the data sources, a probabilistic record linkage technique can be used to join data sources (Gu, Baxter, Vickers, & Rainsford, 2003).

Within this probabilistic system, each field participating in the linkage comparison is subject to error which 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.

The general information flow in a probabilistic record linkage system can be grouped into seven main categories (Gu et al., 2003): data, standardization, searching/blocking, selection of attributes for matching/comparison, weights, the decision model, and performance measurement. Data includes datasets from different sources that need to be linked. Standardization is used next to replace spelling variations of commonly occurring words with a standard spelling. Without standardization many true matches could be wrongly designated as non-matches because the common identifying attributes do not have sufficient similarity. After standardization, searching/blocking is used to reduce the number of comparisons of record pairs by bringing only the linkable pairs together. A good attribute variable for blocking should contain a large number of attribute values that are fairly uniformly distributed. Such an attribute must have a low probability of reporting errors. The ideal blocking component would be one which nearly always agrees in “true match” record pairs but nearly always disagrees between pairs which are not valid matches (Jaro, 1984). Due to their key role in defining location, street names are generally used as the blocking mechanism in address geocoding tools. Soundex, described in a following section, is then used to create an index on the street name attribute to reduce the possibility of a false mismatch due to misspellings. Therefore, if the street name (or rather, the Soundex index value generated from the street name) is not found, no possible matches are suggested. A common geocoding error is the incorrect standardization of an address like “1300 North Star Rd” to “1300 N Star Rd”. Due to pre-defined look-up tables that define “North” as a directional prefix, through the standardization process “North” is stripped from the street name and a Soundex code is generated for “Star” rather than “North Star”, and consequently a match is not found, resulting in a false negative result. The next step involves the selection of attributes for matching/comparison. Common attributes should be selected for use in the comparison function. Major components 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 United States and the world (which might incorporate additional components, such as zone, street suffix, prefecture, etc.) which requires the use of custom locator styles for differing datasets. A comparison vector (a weight) is then used for each pair based on assigned weights. The discriminating power of a component (such as an address element) is a measure of how useful that component is in predicting a match. If the components are assumed to be statistically independent, then the composite weight is equal to the sum of the individual component weights (Jaro, 1984). For each record pair, a decision is then made whether to classify the pair as a match (M), a non-match (U), or tie (T) which must be followed-up interactively by the user to manually specify the correct match.

2.3. Standardization

Standardization is a part of the probabilistic record linkage process. However, it is such a specialized component in GIS that it deserves individual attention. The most common approach for name and address standardization is the manual specification of parsing and transformation rules. An input string is first parsed into individual words. Each word is then mapped to a token of a particular class (Churches, Christen, Lim, & Zhu, 2002). The choice of class is determined by the presence of that word in user-supplied, class specific lexicons, or by the type of characters found in the word such as all numeric, alphanumeric or alphabetical. The process described above is a deterministic approach, meaning a one-to-one match must be found for certain address components (such as house number or street type). Churches et al. (2002) present a probabilistic method for standardization using Hidden Markov Models (HMMs) as an alternative. At present few practical implementations of HMMs have emerged for geocoding with the notable exception of the Geocoded National Address File (G-NAF) for Australia (Christen, Churches, & Willmore, 2004).

2.4. Soundex

Soundex is a way of indexing information based on how the word sounds rather than how it is spelled. It is a phonetic indexing system, blocking together many of the common types of spelling errors and abbreviations. Most versions of Soundex convert a text string into a code consisting of the first (leftmost) letter of the string, followed by 3 or more digits (Patman & Shaefer, 2001). The method is based on the phonetic classifications of human speech sounds, which in turn are based on where you put your lips and tongue to make the sounds. The key concept behind Soundex relies on the assumption that a constant relationship between letters and sounds should assure that similar-sounding names are assigned the same code. Soundex also functions as a compression scheme since the code contains one half to two thirds the information contents of the full name (Winkler, 1999). Within the geocoding process a Soundex index is commonly applied for the street name component of the standardized address.

Some of the limitations of Soundex that must be taken into account include (Patman & Shaefer, 2001): sensitivity to spelling variations, the algorithm's dependence on the initial letter, noise intolerance (mistyping, extra consonants, swapped consonants), differing transcription systems, names containing particles, perceptual differences, silent consonants, and the use of initials, among other potential errors. Testing of Soundex has shown it to produce a high number of incorrect matches (Stanier, 1990), and improvements have been suggested (Christian, 1998). Despite its limitations, Soundex is currently implemented in most geocoding software, but other types of probabilistic record linkage that do not rely on Soundex have been developed (Christen, Churches, & Zhu, 2002).

3. Address data modeling

One of the main challenges to accurate geocoding is the availability of good reference data. This includes a set of geographic features that are needed to match against as well as robust address characteristics (attribute data) that enable matching address records to feature locations in a GIS. This requires a sturdy address model to organize the reference data components in a logical, maintainable and site-specific way.

There are many challenges to building good reference data (Arctur & Zeiler, 2004). Addresses can be associated with many kinds of feature classes in a reference database; for example, road centerlines, parcel boundaries, address points, building structures, etc. The complexities of address component relationships might also dictate that some address elements be organized in separate, related tables since addresses and features in the GIS can share complex relationships (such as many-to-many). A feature might also have sub-addresses. For example, a parcel may house a duplex with two separate addresses. Sets of address components can also vary by locale and culture.

Several common address models exist. Each has a particular set of supporting materials and characteristic errors. The first one can be characterized as the "geographic unit" model. These geographic units can consist of postal codes (such as ZIP codes in the United States), 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. Location within the unit is not specified, but analyses can be carried out using data associated with the geographic unit. Postal codes are particularly attractive since this type of information is much easier to obtain than individual street address information and postal code data also tends to be very complete and accurate. For example, most people know their postal code and are less likely to provide misspellings or alternative descriptors than for street addresses. The utility of the results is obviously related to the size of the geographic units. For example, in the United States 5-digit ZIP codes tend to be quite large, typically larger than census tracts, making them less attractive when spatially detailed information is required. 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. The Postal Code Conversion File developed by Statistics Canada and Canada post contains the geographic coordinates of each postal code. In major urban areas a single 6-character postal code typically corresponds to a single block-face (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 using straight-line distance.

For many application that do not require individual-level locations, geocoding at the postal code level might be very appropriate, in particular since match rates at this level are typically very high. When geocoding at the level of postal codes is not sufficient, several alternatives exist, including street networks, parcel boundaires and address points. Each of these three address models will be described in more detail below.

3.1. Model 1: Street network data model

The most widely employed address data model is based on street network data. In this approach a street network is represented as street line segments that hold street names and the range of house numbers and block 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 placing a point along the segment based on a linear interpolation within the range of house numbers. An optional off-set can be employed to show on which side of the street line segment the address is located. This approach to geocoding an address is referred to as "street geocoding" and has become the most widely used form of geocoding. Nearly all commercial firms providing geocoding services and most GIS software with geocoding capabilities rely primarily on street geocoding.

The street network address model facilitates storing different names and address ranges for different sides of the street and enables validation of cases where there is no address range for one side of the street. It also supports cases where streets have multiple address ranges and names. Some additional attribute characteristics include the use of full block address ranges for major roads, while true address ranges are commonly used for residential roads. For better interpolation results, it is generally preferred to geocode with as much block-face accuracy as possible (that is, against true address ranges). While this results in a better spatial location for known valid addresses, this can also be problematic. When approximated addresses are geocoded against the centerline the records fail to match since the value does not fit into the existing range. An example of this situation would be an address like "300 [block] E Main St" when the known address range may run only from 315 to 345. Thus, some padding may be required even for true address ranges, and other means are needed for geocoding approximated data. Mixed parity issues also exist for some roads which throws off interpolation techniques. Street name alias fields may exist in the attribute table since naming standards can vary. This also allows for some flexibility in modifying street names to better fit geocoding rule base expectations.

3.2. Model 2: Parcel boundaries data model

Parcel boundaries are traditionally the most spatially accurate data with address information available. Geocoding against parcels allows for matching against individual plots of land (or the centroids of those polygons) rather than interpolating against a street centerline. This is particularly useful in areas where parcels are not regularly addressed (such as on roads with mixed parity) or those parcels that may be quite a distance from the centerline.

A principal difference between parcel and street geocoding is that a single parcel usually has a single house number, while a single street segment has an address range. This implies that a match is only obtained in parcel geocoding if there is a perfect match for the house number; for street geocoding a match is obtained if the house number being matched falls within the address range for the street segment. In effect street geocoding does not provide for a check if the house number actually exists and can therefore more easily result in false positives (i.e. produce a match for a non-existing address location). This is one of the reasons why parcel geocoding typically results in a lower match rate. Another perhaps more important reason why parcel geocoding produces lower match rates than street geocoding is that a single parcel can be associated with many addresses; for example, duplex units, condominiums, apartment complexes, commercial sites, etc. While the parcel may have an address, the addresses of individual structures or units on the parcel are not always captured in the parcel database.

While geocoding using parcels is more spatially accurate than geocoding using streets, parcel data may not necessarily constitute all valid addresses within an area. Additionally, not all parcels have a true address. Some may have an abstract number or a non-standard reference listed in the address fields.

Despite the often lower match rates, parcel geocoding is generally considered more spatially accurate and is now becoming widespread given the development of parcel level databases by many cities and Counties in the United States (Rushton et al., 2006).

3.3. Model 3: Address point address data model

To overcome the limitations of parcels for geocoding, address points have emerged as a third address data model. The address point data model is often derived from a master address file (MAF) of all known addresses, which is frequently available in the form of an E911 address list compiled for emergency response purposes. Address point data can also be constructed from several existing data layers such as parcel data. Address points are 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, condominium units, duplexes etc. which are not recorded as separate properties in the parcel data. Field data collection or verification of building locations using digital aerial imagery can be used to further supplement the address point file.

Both Australia and the United Kingdom have developed national address point databases. In Australia, this database is part of the Geocoded National Address File. In the United Kingdom, this database has been developed by the Ordnance Survey and is referred to as the ADDRESS-POINT dataset. These two efforts have set the stage for other jurisdictions to develop similarly detailed and comprehensive address point databases. At present, however, there has been limited published research on the quality of the geocoding based on either GNAF or ADDRESS-POINT.

In the United States address point geocoding at present is not in very widespread use. However, many local governments have started to create address point databases and several commercial goecoding firms have started to provide address point geocoding for selected urban areas.

4. Geocoding quality

For the results of geocoding to be meaningful, the geocoding process needs to meet certain quality expectations. Despite the widespread use of geocoding in a range of disciplines, the errors of geocoding have not received widespread attention in the literature. Much research that uses geocoding as one of its methods does not include any mention of the quality of the geocoding; if a reference is made to the quality, usually only the match rate or geocoding completeness is mentioned. Commercial geocoding firms also commonly emphasize high match rates to describe and promote their services, with little attention to other aspects of geocoding quality. Recent research is suggesting the emphasis on match rates is somewhat misplaced and potentially misleading (Whitsel et al., 2004).

The overall quality of any geocoding result can be characterized by the following components: completeness, positional accuracy 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. Repeatability indicates how sensitive the geocoding results are to variations in the street network 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 repeatable.

Several studies have been published that seek to investigate and evaluate the effectiveness of geocoding techniques and the quality of the final result. Most prevalent are those associated with health database mapping due to the dramatic increase in the number of public health applications using geocoding to assess geographical distributions of health-related issues such as zones of exposure and rates of disease. Three major categories of geocoding error can be identified: (1) data input errors, (2) reference data errors, and (3) errors related to the underlying geocoding process. These errors can be broken down into more precise categories for analysis in order to facilitate problem solving. Traditionally difficult addresses include apartment units, commercial suites, shopping center suites not addressed to the street centerline, and other troublesome address data anomalies.

4.1. Match rates

The simplest measure of geocoding quality is the match rate, or the percentage of records that produce a reliable match. 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 what minimum match rate is needed to obtain a reliable pattern of crime incidents. Results indicated that to generate a statistically reliable pattern a match rate of 85% was necessary. In general, however, match rates reported by studies that have employed geocoding vary greatly since they depend on many factors. There is no consensus on a universal standard for an acceptable geocoding match rate.

The match rates increases if efforts are made to increase the quality of the address file and the geographic reference file. Interpreting match rates, however, is very subjective since 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 minimum match score results in a decrease in accuracy and therefore geocoding quality.

4.2. Positional accuracy

Several studies have determined quantitative estimates of the positional accuracy of geocoding. Estimates of 'typical' positional errors for residential addresses range from 25 to 168 m (Bonner et al. 2003; Cayo & Talbot 2003; Dearwent, Jacobs, & Halbert 2001; Karimi & Durcik 2004; Ratcliffe 2001; Schootman et al., 2007; Strickland, Siffel, Gardner, Berzen, & Correa, 2007; Ward et al. 2005; Whitsel et al., 2006; Zandbergen, 2007; Zhan, Brender, De Lima, Suarez, & Langlois, 2006; Zimmerman, Fang, Mazumdar, & Rushton, 2007) based on median values of the error distribution. Results in urban areas are generally more accurate than in rural areas (Bonner et al. 2003; Cayo & Talbot 2003; Ward et al. 2005). It should also be noted that the occurrence of major positional errors is relatively common. For example, in one of the more thorough studies by Cayo and Talbot (2003) 10% of a sample of urban addresses geocoded with errors larger than approximately 96 m and 5% geocoded with errors larger than 152 m. For rural addresses these distances were 1.5 and 2.9 km, respectively.

The positional error in geocoded addresses may adversely affect spatial analytic methods. Specific effects includes inflation of standard errors of parameters estimates and a reduction in power to detect such spatial features as clusters and trends (Jacquez & Waller, 2000; Waller, 1996; Zimmerman, 2007). Even relatively small positional errors can have an impact on local statistics for detecting clusters (Burra, Jerrett, Burnett, & Anderson, 2002). Research on this topic has been mostly confined to the health field. For example, typical street geocoding is not sufficiently accurate for the analysis of exposure to traffic-related air pollution of children at short distances of 250–500 m (Zandbergen, 2007; Zandbergen & Green, 2007). Similar errors in misclassification of exposure potential have been identified by Whitsel et al. (2006).

4.3. Repeatability

The repeatability of geocoding has not received as much attention as positional accuracy. In one recent study by Whitsel et al. (2006) using a large sample ($n = 3615$) of addresses in 49 United States, substantial differences were found between four commercial vendors. There were important differences among vendors in address match rate (30–90%) concordance between established and vendor-assigned census tracts (85–98%) and distance between established and vendor assigned coordinates (mean of 228–1809 m). This confirmed earlier findings by Whitsel et al. (2004) for a much smaller sample that the repeatability of commercial geocoding is not very good. The exact causes for the lack of repeatability are unknown, since the geocoding algorithms and data quality procedures of commercial vendors are not disclosed.

In a comparison of three geocoding algorithms (LocMatch, ArcView 3.2 and Tele Atlas North America) using the same TIGER reference data, Karimi and Durcik (2004) found that the differences between the results were not significant. This suggests that differences in reference data are at least in part responsible for the observed differences between commercial vendors.

4.4. Study objective

Several different address models for geocoding have emerged, but very limited research has been carried out to determine their relative strengths and weaknesses. The objective of the empirical component of this study, therefore, is to compare the reliability of the address point, parcel and street network data models for geocoding. This comparison is accomplished by geocoding the same address databases using the three different address data models for the same geographic areas. To strengthen the comparison several different types of address databases from three different jurisdictions are used. The comparison emphasizes geocoding completeness (i.e. match rates) since positional accuracy is inherently tied to the type of address data model used (i.e. address point and parcel geocoding produce more spatially accurate results than street geocoding).

5. Methods

5.1. Study area

Reliable and complete reference information for address point, parcel and street geocoding is not available for all areas. As a result, this study employed an extensive search strategy to identify Counties with this type of reliable reference data in GIS compatible format. The search was limited to the State of Florida; most Counties in Florida have undertaken major investments in GIS data over the last two decades and access to address data of various types is also generally good in part due to the requirements of the Sunshine Law (Florida Statutes Chapter 286) to make public records available.

For each of Florida's 67 Counties, GIS Departments and Property Appraiser's Offices were contacted with a request for digital copies of address point, parcel and street centerline data in GIS format. A few Counties remained unresponsive to repeated requests, and several Counties do not maintain a GIS database, but ultimately digital data was obtained from 62 of the 67 Counties.

Street centerline data was available for all 62 Counties, and in most cases contained the proper fields required for geocoding. Parcel data was also available for all 62 Counties, but did not always contain the proper fields. The first priority in maintaining parcel data is not for geocoding, and therefore the completeness of the data is not always sufficient for geocoding. Sometimes address information is completely lacking and only legal descriptions are provided. Sometimes the address information is stored in a single field, making the creation of an address locator complicated. Despite these limitations, data from 35 Counties was deemed sufficient for geocoding. Development of address point data has not received the same level of effort as street centerlines and parcels, and was available for only 11 Counties. Since geocoding is often one of the main objectives in developing an address point database, their quality for this purpose is generally good and all 11 databases were considered adequate.

Upon review of the three databases for each County, only seven Counties were identified as having a reliable database for all three types. Of these seven, the three Counties with the largest population were selected based on sample size considerations: Bay, Collier and Seminole County. The location of these three Counties is shown in Fig. 1. Based on this selection process, the databases for the three Counties are by not truely representative of what a typical GIS datatabase at the County level looks like. Instead, they represent examples of the very best data available in terms of completeness, currency and appropriateness for

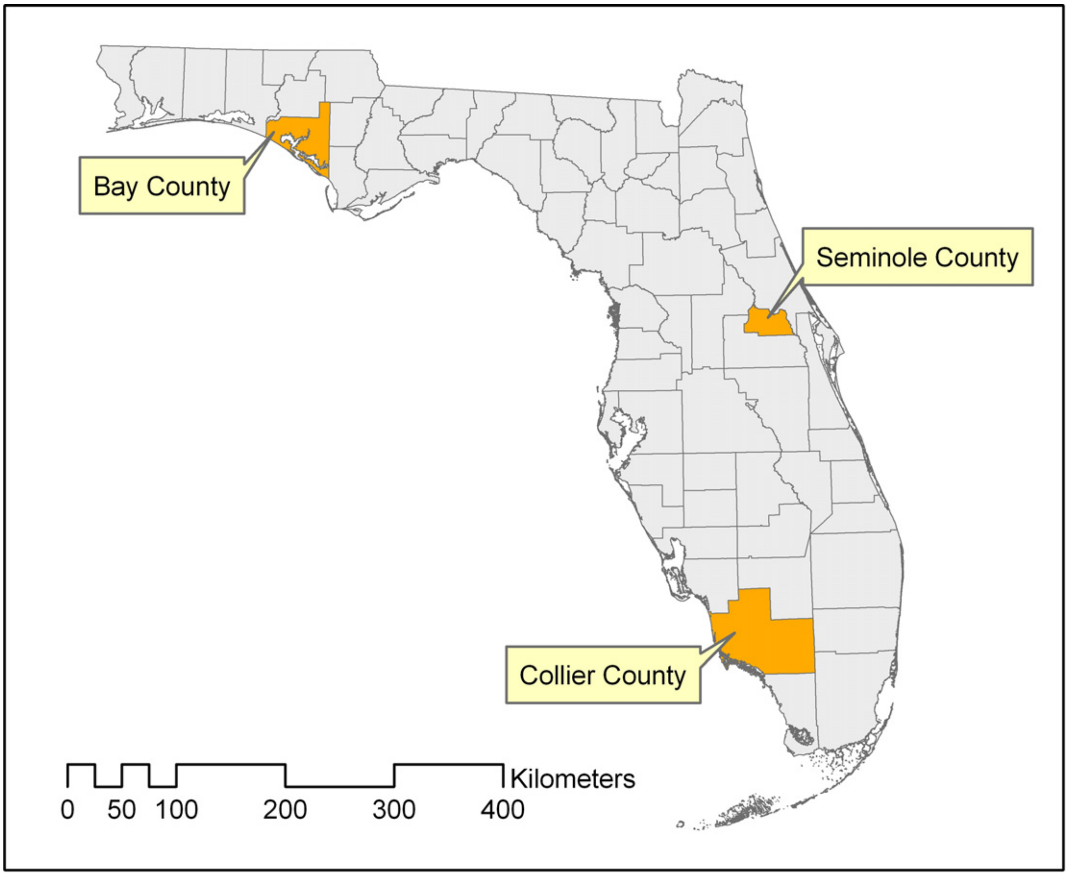

Fig. 1. Location of Counties used in this study.

geocoding. For the purpose of this study they represent a case-study of a best-case scenario which is not currently feasible at the State or national level in the United States, but represents an objective to strive towards. The case-study will illustrate the current performance of this best-case scenario.

5.2. Addresses for geocoding

Six different databases were obtained for use in the comparison of geocoding using three different address models. The selection of these databases was driven by a number of considerations. First, the database had to be publicly available to facilitate data access. Second, the database had to be recently updated (2005 or 2006) to prevent temporal bias. Third, the database had to be available for the entire State of Florida to allow for comparisons among the three Counties. Fourth, sufficient sample size for each County was needed. And fifth, a range of different types of addresses was needed, including residential, commercial and other types. The following six databases were decided upon: commercial banks, child care facilities, properties with elevators, establishments with food permits, saltwater recreational fishing license holders, and registered sex offenders. Each will be briefly described below.

Addresses for all branches of licensed commercial banks in Florida were obtained from the Federal Deposit Insurance Corporation (FDIC) in March 2006 ($n = 5{,}138$). Banks were selected as an example of commercial properties with a very good address description due to their strict licensing. It was anticipated that the address information from the FDIC would be very complete and standardized.

Addresses for all registered child care facilities in Florida were obtained from the Florida Department of Children and Families (FDCF) in March 2006 ($n = 13{,}564$). The child care facilities database was selected for this study since they include both commercial and residential properties, i.e. licensed home child cares with a maximum of 10 children are part of this database. It was expected that this database would not be as complete or standardized as some of the other databases, since the licensing of child care is handled mostly by local authorities, which may vary across the State.

Addresses for all properties with a licensed elevator were obtained from the Florida Department of Business and Professional Regulation (FDBPR), Bureau of Elevator Safety ($n = 45{,}998$). The elevator database was selected for this study since these properties contain both commercial and residential multi-family units which are known to be a challenge in geocoding. Since elevators are regulated and inspected by the State of Florida, a high degree of address standardization and completeness was expected.

Addresses for all saltwater recreational fishing license holders were obtained from the Florida Fish and Wildlife

Conservation Commission (FFWCC) in December 2005 ($n = 744,149$). The fishing licenses were selected for this study as an example of a large database of mostly residential addresses. It was also anticipated that the addresses in this database would not be very complete or standardized since the address information provided by the applicant is completely self-reported, with very little data validation or checking.

Addresses for all food establishment in Florida were obtained from the Florida Department of Agriculture and Consumer Services (FDACS) in March 2006 ($n = 40,780$). Food establishments contain all those facilities were food items are processed and sold; the majority consists of supermarkets, grocery stores and convenience stores. Food establishments were selected for this study as an example of commercial properties with a large sample size.

Addresses for all sex offenders registered in Florida were obtained from the Florida Department of Law Enforcement (FDLE) in December 2005. From the original database, only those offenders not in jail and with their latest known residence in the State of Florida were selected ($n = 18,551$). Sex offenders were selected for this study as an example of mostly residential addresses, although it is known that some offenders reside in transient housing, including hotels and motels.

Each of the six databases contained fields for address, city, County and 5-digit ZIP code. From each database the records associated with the three Counties of interest were selected.

Once the County-level databases were established (six types for three Counties for a total of 18 databases), the address fields were examined for any blanks, and these blanks were removed prior to geocoding. Blank addresses were particularly common for the residential child care facilities and fishing licenses. Removing these blanks may introduce some bias. For example, residential child care facilities in the database may have a blank address while commercial child care facilities do not. However, the objective in this study is to compare geocoding techniques, and not an assessment of the availability of child care. Therefore, the removal of blanks was considered appropriate. Table 1 reports the final sample size used for geocoding and the number of blanks removed prior to geocoding where applicable.

5.3. Reference data for geocoding

Address point data, parcel data and street centerlines data were obtained from Bay, Collier and Seminole County in April 2006. Currency of these data varied, but all had been updated in mid-2005 or later, and the data obtained presented the most up-to-date and complete datasets available directly from the Counties' GIS Department and/or Property Appraiser. Each of these reference datasets contained several attributes for the address; although specific fields varied, all had the following as a minimum: number, prefix direction, street name, street type and suffix. Several datasets had fields for City or 5-digit ZIP code, but this was not consistent – as a result, the use of a "zone" field (which commonly uses the ZIP or City field) was not feasible for all reference data.

Table 1
Sample size of address databases used for geocoding

Database	Bay		Collier		Seminole	
	n[a]	Blanks	n[a]	Blanks	n[a]	Blanks
Commercial banks	57	–	127	–	119	–
Child care – commercial	54	–	104	–	124	–
Child care – residential	23	20	48	58	82	55
Elevators – commercial	316	–	1181	–	787	–
Elevators – residential	251	–	1439	–	42	–
Fishing licenses	10,336	1108	11,116	1132	9815	1047
Grocery stores	452	–	691	–	891	–
Sex offenders	289	–	189	–	306	–

[a] Sample size after removal of records with blank addresses.

5.4. Geocoding process

Address locators were created in ArcGIS 9 for the three reference datasets for each of the three Counties for a total of nine address locators. Fields included in each locator included number, prefix direction, street name, street type and suffix. Additional fields were available in some cases (usually a field for prefix type) but were not used to maintain consistency between the address locators. No field was used for "zone" since this was not consistently available in the reference datasets. City or ZIP fields are normally used for "zone" and this is often required in geocoding since it speeds up database searches and prevents the occurrence of a large number of ties. For example, an address like 123 Main Street is expected to occur in almost every major city; the use of a City or ZIP field as a search criteron in addition to the address itself prevents these ties or potentially incorrect matches. Since a separate address locator was built for each County, the use of a zone was not necessary. In the geocoding results, any ties were investigated and none of these ties were a result of not using a "zone" field in the address locator.

For each address locator, settings for spelling sensitivity and match score were set to identical thresholds. After experimentation with a sample dataset, the minimum match score was set to a value of 60 (out of 100). If the house number was not a one-to-one perfect match (for address points and parcels) or did not fall within the house number range for a street segment (for streets), the maximum score obtained by the ArcGIS 9 rule-based geocoding algorithm was 52. As a result, using the minimum match score of 60 in effect ensured that a match was only obtained if the house number was an unambiguous perfect match. In addition, ties were permitted, but identified separately in the results.

5.5. Analysis

For each database geocoded, the number of perfect matches and ties (score = 100), the number of additional matches and ties (score < 100), and the number of unmatched cases were determined. The overall match rate for each database was determined by calculating the sum of all matches and ties as precentages of all address records. The percentage of ties as a percentage of all matches was also determined. For the child care facilities and properties with elevators the analysis was carried out for the entire database as well as separately for residential and commercial addresses.

6. Results and discussion

6.1. Description of reference data

A summary of the number of features in each of the reference datasets is provided in Table 2. The number of residents per parcel for the three Counties ranges from 1.46 to 2.59. This number is strongly influenced by the presence of multi-family units with a large number of residents residing on a single parcel. The highest number of 2.59 for Collier County is therefore not surprising, since multi-family units are much more common here than in the other two Counties. The number of address points for Bay and Seminole County is quite a bit smaller than the number or parcels, while for Collier County the number is slightly higher. This reflects in part the same difference in multi-family housing: a single parcel with multi-family units may contain many address points.

When comparing the number of address points to the number of residents in each County, the ratios are much more similar with values ranging from 2.43 to 2.59. These values are similar to the average household size reported in the 2000 Census, but this comparison is confounded by the fact that many address points are not residential, and that the relationship between address points and households is not consistent. For example, many multi-storey apartment complexes with multiple units may get assigned a single address points for every single building structure which may contain many units.

Table 2
Descriptive summary of reference data used in geocoding

	Bay	Collier	Seminole
Population (2005 Census)	161,558	307,242	401,619
Parcels	110,651	173,787	154,919
Residents per parcel	1.46	1.77	2.59
Address points	75,928	125,329	155,208
Residents per address point	2.43	2.45	2.59
Average household size (2000 Census)	2.48	2.39	2.59
Parcels w/o address point	37,352	66,427	18,229
Parcels with 1 address point	71,874	104,927	132,088
Parcels with 2 address points	1119	1047	3412
Parcels with >2 address points	306	1386	1190
Number of street line segments	15,892	14,125	21,580
Parcels per street segment	6.96	12.30	7.18
Address points per street segment	4.78	8.87	7.19

Table 2 also reveals that there are many parcels without an address point. Most undeveloped parcels do not get assigned an address point, or even an address for that matter. The majority of parcels have only a single address point – this would be typical of single family residential housing, but also applies to many other types of commercial, industrial and institutional properties. A smaller number of parcels have two address points and this would be typical of residential duplex units. An even smaller number of parcels has more than two address points, and these would be typical of larger multi-family complexes and commercial sites with many individual businesses located on the same parcel.

The number of street segments within each County is much lower than the number of parcels or address points. The number of parcels and address points per street segment varies from 4.78 to 12.30. While there is no significance to the particular values for this ratio, it provides a general measure of the difference in resolution of the three types of reference data.

A closer examination of the relationship between address points, parcels and street networks also reveals some interesting examples. For example, Tyndall Air Force Base in Bay County is classified as a single parcel owned by the United States Air Force. However, a detailed street network within the base is part of the street centerlines database, and the base contains no fewer than 504 address points for a residential complex located on the base. For this particular example geocoding using only the parcels would not generate any matches, but both the address points and street network will likely produce matches. While an Air Force Base presents a very special case, large parcels with many individual addresses are fairly common. Collier County contains no less than 42 parcels with more than 100 address points. Most of these are mobile home parks, RV parks, apartment complexes or other types of rental housing where many separate structures are located on the same parcel.

6.2. Placement of address points

The placement of address points further illustrates some of the differences between the address data models. One common approach to the placement of the address point is at the centroid of the main structure on the parcel. Fig. 2 shows a typical example for a single family residential neighborhood in Colliler County. There is only one structure per parcel and a single address point is placed within each parcel at (approximately) the building centroid. This results in one address point per residential unit.

The situation for multi-family residential areas is different as illustrated in Fig. 3 for Collier County. For duplexes

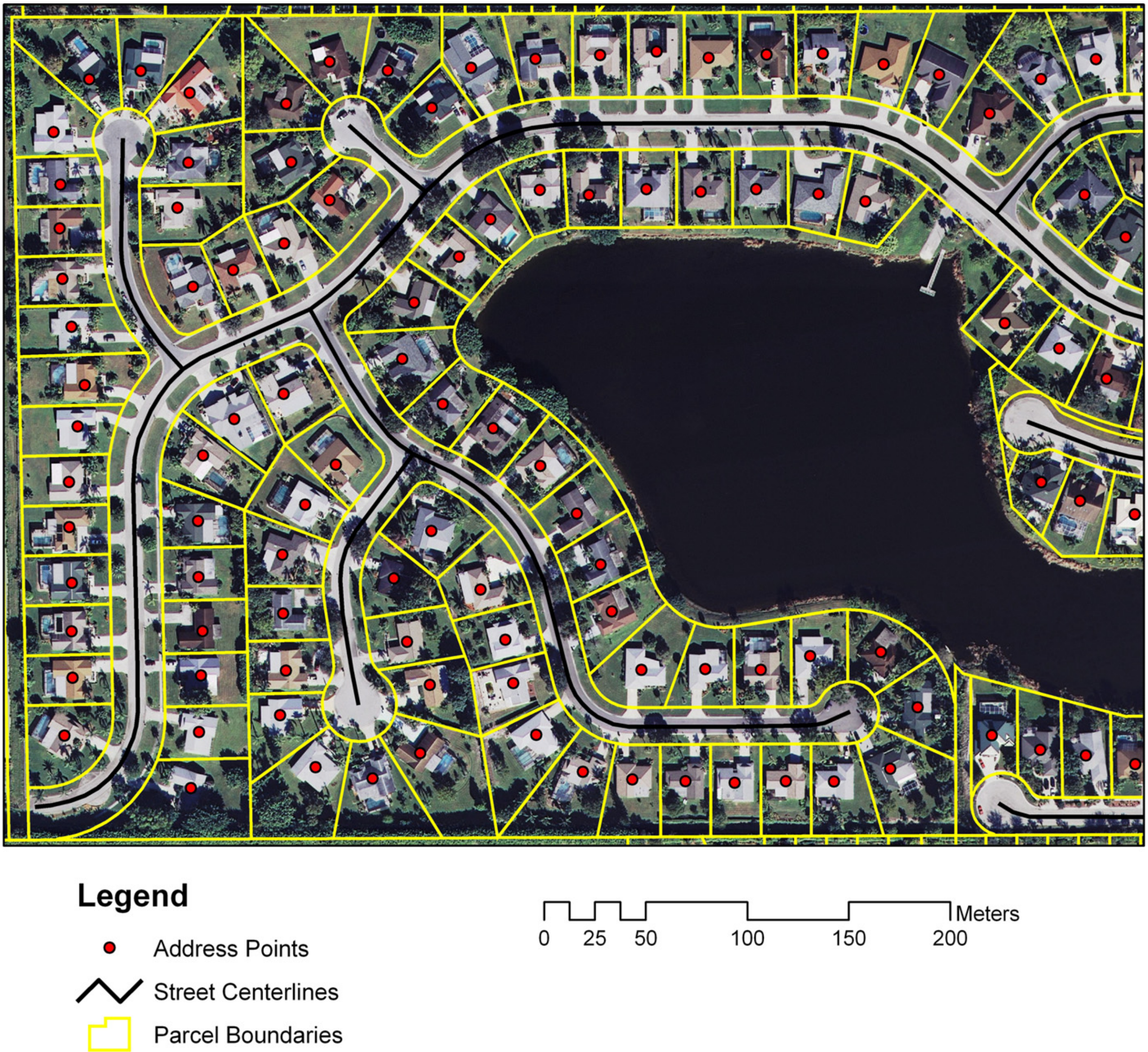

Fig. 2. Example of address points and parcel boundaries for single-family residential area in Collier County, FL.

and townhouses (bottom right of Fig. 3), one address point is placed for each residential unit, resulting in two or more address points per structure. These residential units have unique street numbers. For multi-unit apartment complexes (top center of Fig. 3), only a single address point is placed for each structure which may contain many residential units. These residential units share a single street number, and units are uniquely identified by their unit number (e.g. #101, 102, etc.). The examples in Figs. 2 and 3 represent the most widely used approaches to the use of address points for residential units within the datasets examined.

For commercial, industrial and institutional properties, the situations is different again. Fig. 4 shows a typical commercial area in Collier County. A number of parcels contain only a single structure with a single unit, and the address point is (mostly) placed at (approximately) the building centroid. However, several of the structures in the shopping plaza contain multiple businesses, each with their own street number, and an address point is placed for each of these businesses. The placement of these address points is somewhat arbitrary, but appears to correspond to the (approximate) centroid of the portion of the structure occupied by the business.

The examples in Figs. 2–4 illustrate the logic most widely followed: a unique address point is placed for every unique street number, which may represent many units if they share the same street number. The location of the address points varies somewhat and can be near the centroid of the main structure or near the front of the

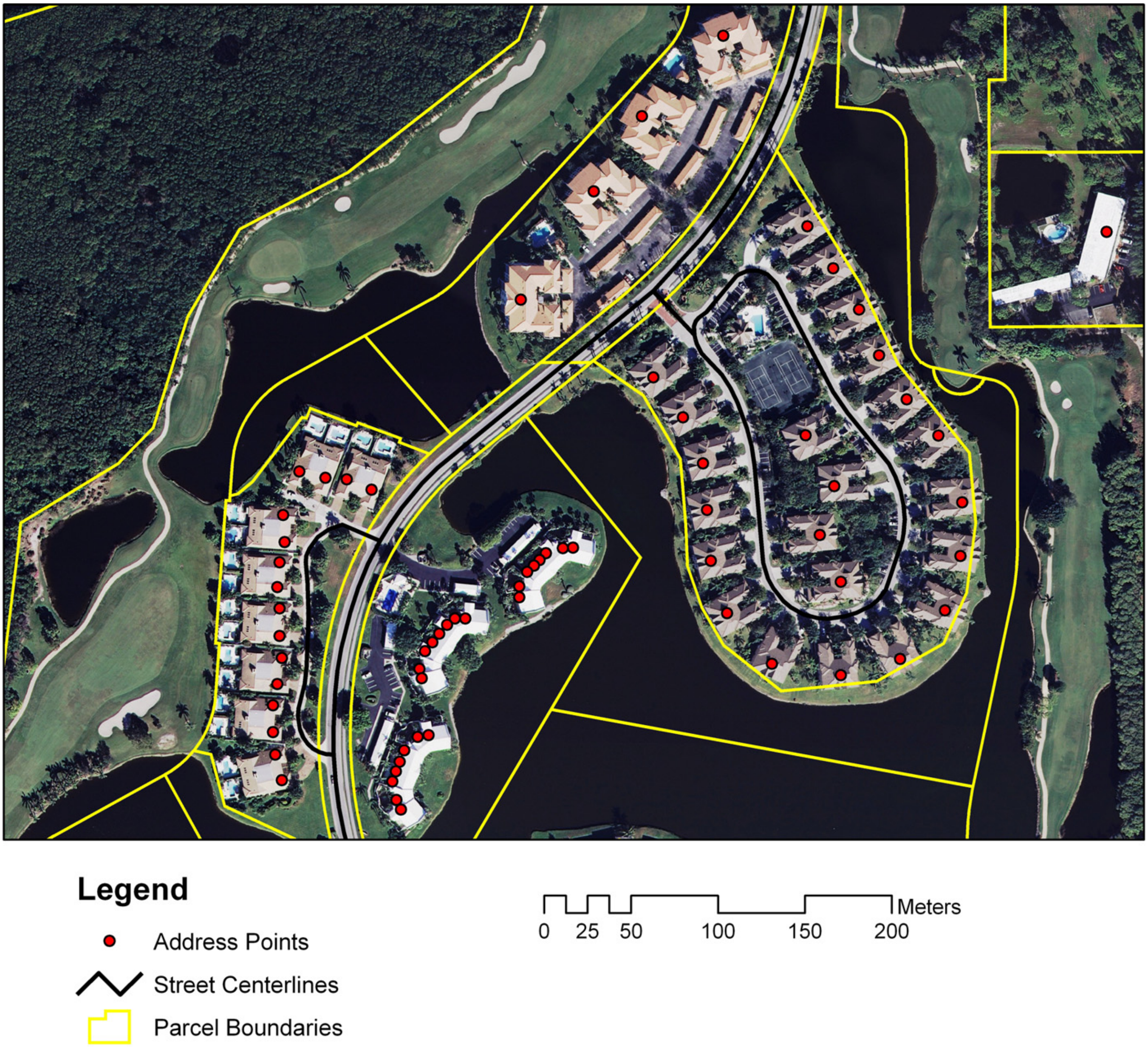

Fig. 3. Example of address points and parcel boundaries for multi-family residential area in Collier County, FL.

structure. For example, Fig. 5 shows a mixed residential/institutional area in Seminole County, and the placement of the address points is clearly not near the building centroid, but in front of the building towards the street that the address is located on.

6.3. Geocoding match rates

Table 3 reports the results for the match rates for the six address databases for each of the three Counties. The number of perfect matches and ties (score = 100), additional matches (sore < 100) and unmatched cases are reported separately. To facilitate the interpretation of the results, the overall match rates are also plotted in Fig. 6.

A number of relevant trends can be derived from Table 3 and Fig. 6. The match rates in general are highest for street geocoding, followed by address points and parcels. Match rates for street and address point geocoding are generally relatively close, with match rates for parcels being a distant 3rd and rarely exceeding 70%. This general trend confirms the hypothesis that parcel geocoding results in lower match rates; parcel databases only associate one address with a parcel, while in reality a single parcel may contain many addresses. The slightly higher match rates for street geocoding compared to address point geocoding can in part be attributed to the way in which street numbers are stored in the two address data models. A single address point contains a single house number, while a

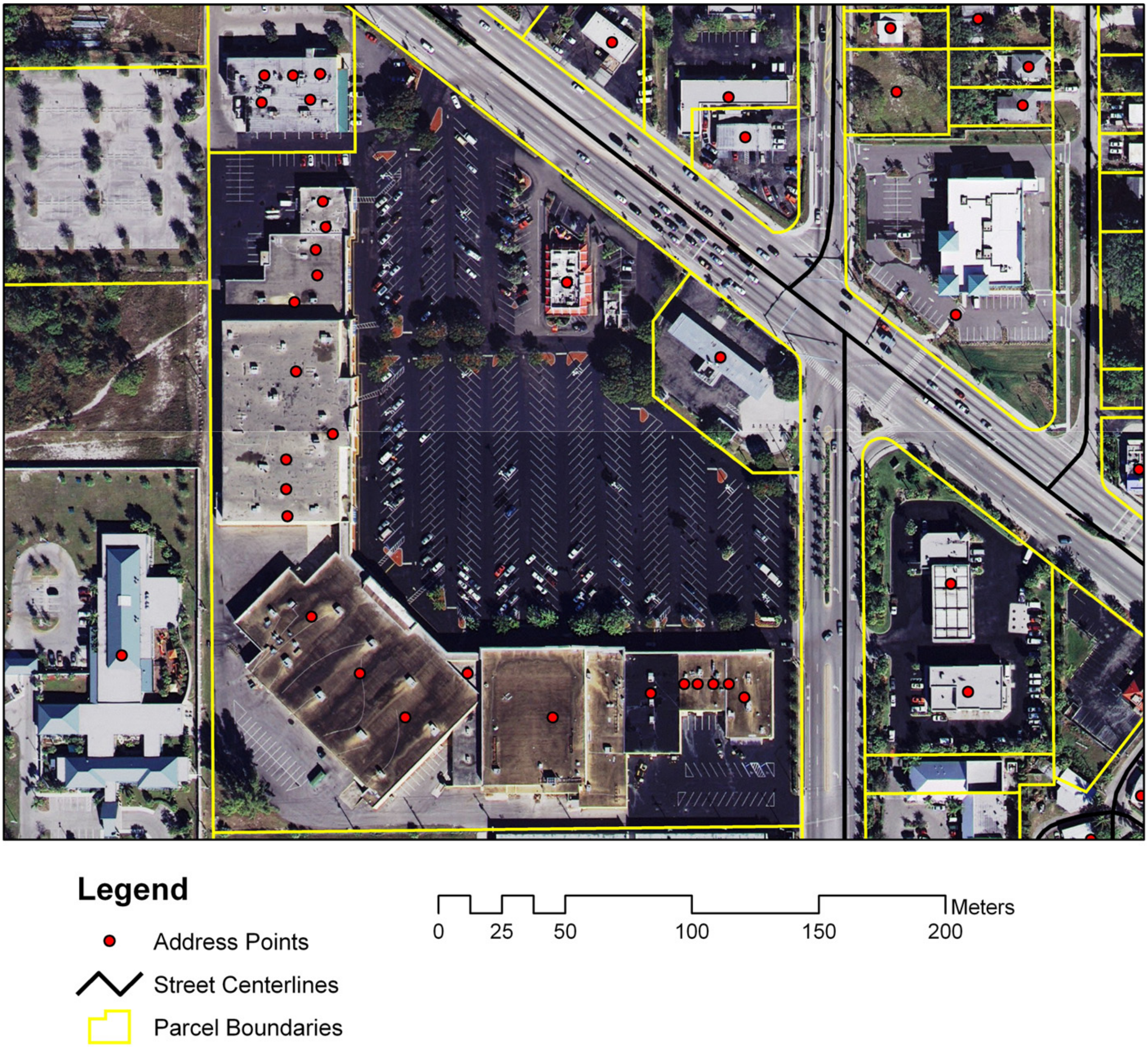

Fig. 4. Example of address points and parcel boundaries for commercial area in Collier County, FL.

street segment contains a range of house numbers. Address point geocoding is therefore much more sensitive to data entry errors since a match is only obtained when a perfect one-to-one match can be made for the house numbers. The addresses that were matched using street geocoding but did not produce a match using address point geocoding could therefore in fact represent false positives, since street geocoding does not provide a mechanism to determine if a particular street number exists. Without extensive field validation however, it is not possible to say with certainty that the higher match rate obtained using street geocoding produced a better result.

Geocoding match rates vary strongly between the six address databases considered. Match rates for the addresses of sex offenders are consistently highest for all three Counties considered, but the pattern for the other databases is not very consistent. The lack of agreement in the match rates between the three Counties points to differences in the reference data. For example, for Collier County street geocoding consistently produces match rates of greater than 80%, while there is much more variability in the street geocoding match rates for the other two Counties. For parcel geocoding on the other hand, results for Collier County include some dramatically low match rates, in particular for commercial properties (banks, elevators and grocery stores) which are all below 50%. Another difference can be observed between address point and parcel geocding. For both Bay and Collier County, match rates for address point geocoding are much higher than for parcel geocoding, while for Seminole County the difference is

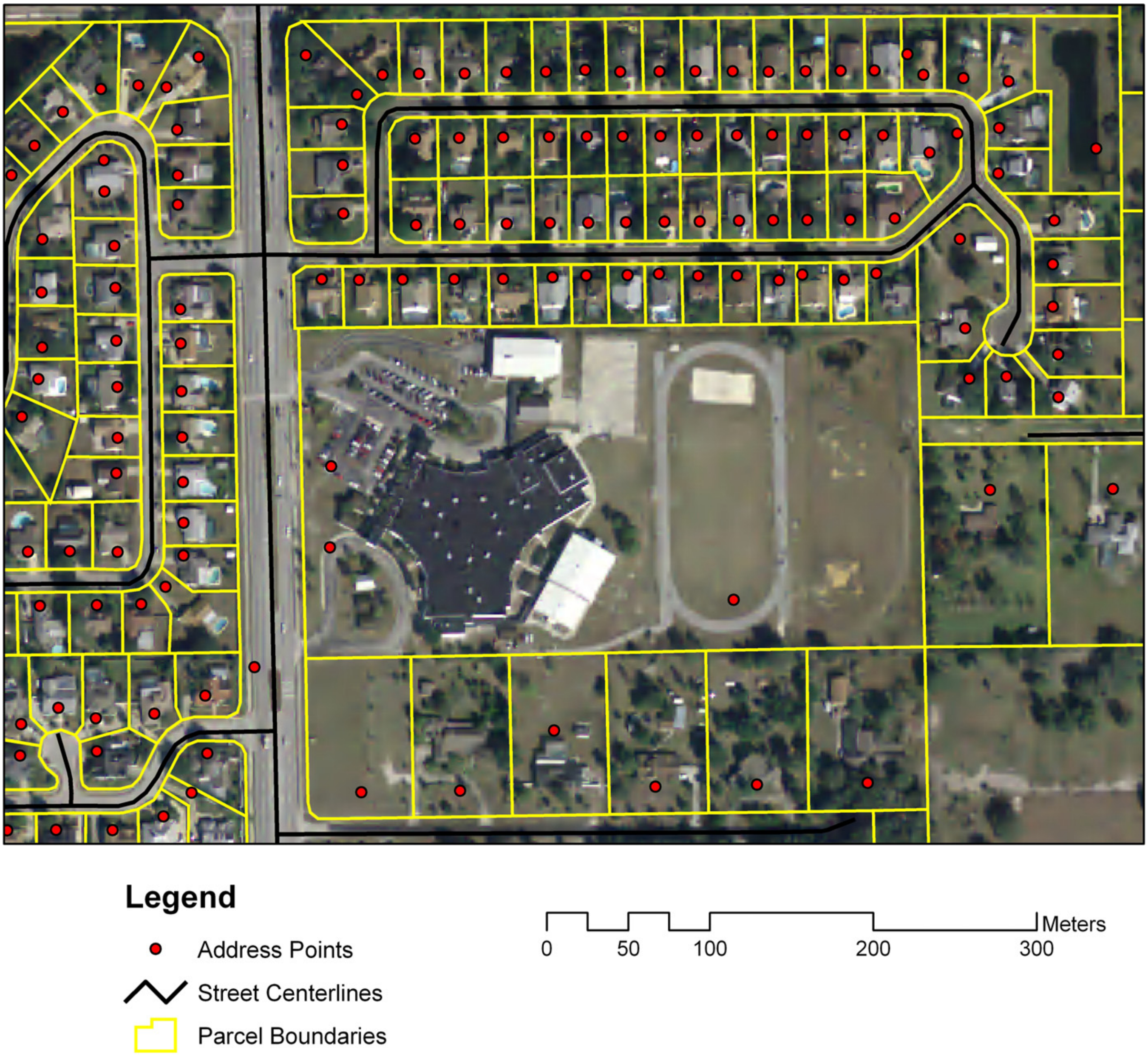

Fig. 5. Example of address points, parcel boundaries and street centlines mixed residential and institutional area in Seminole County, FL.

very small; in fact, for banks the match rates are the same and for child care facilities the match rate for parcel geocoding is higher.

The variability in match rates also points to the influence of the data input quality. Differences in match rates between the six different databases are similar in magnitude to the difference in match rates between the three methods of geocoding for the same database. This strongly suggests that the quality of the input data and the quality of the geocoding process are both important contributors to the quality of the final output, in this case the geocoding match rate.

The match rates reported in this study for street geocoding are similar to those reported by other studies using the same geocoding method for large sample sizes. Whitsel et al. (2006) report match rates of 30%, 77%, 78% and 79% using four different commercial vendors. Zhan et al. (2006) report match rates of 79% and 89% using two different ArcGIS-based methods. Match rates are known to be lower in rural areas, which explains some of the variation between different studies. Cayo and Talbot (2003) report match rates of 62% for rural addresses, 87% for sub-urban areas and 94% for urban areas using street geocoding on a large sample. Fewer studies have employed parcel geocoding, but these few confirm the typically much lower match rate compared to street geocoding. Dearwent et al. (2001) report a match rate of 70% for parcel geocoding versus 89% for street geocoding of the same large sample. No

Table 3
Summary of geocoding match results for six address databases for three Florida Counties using address point, parcel and street geocoding

Category	Commercial banks			Child care facilities			Properties with elevators			Fishing licenses			Grocery stores			Sex offenders		
	Address	Parcel	Street	Address	Parcel	Street	Address	Parcel	Street	Address	Parcel	Street	Address	Parcel	Street	Address	Parcel	Street
	$n = 57$			$n = 77$			$n = 567$			$n = 10{,}336$			$n = 452$			$n = 289$		
Bay County																		
Matched (score = 100)	31	17	34	57	37	58	307	117	332	5995	5195	6200	270	164	319	229	180	253
Tied (score = 100)	0	4	0	0	5	0	0	175	1	2	317	74	0	26	5	0	17	2
Matched (score < 100)	15	14	15	7	7	7	45	11	66	1547	1424	1628	66	22	77	20	9	21
Tied (score < 100)	0	2	1	0	1	1	0	22	3	23	246	86	0	6	3	1	6	1
Unmatched (score < 60)	11	20	7	10	24	8	215	242	165	2769	3154	2348	116	234	48	39	77	12
Match rate (%)	80.70	64.91	87.72	86.49	67.57	89.19	62.08	57.32	70.90	73.21	69.49	77.28	74.34	48.23	89.38	86.51	73.36	95.85
	$n = 127$			$n = 152$			$n = 2620$			$n = 11{,}116$			$n = 691$			$n = 189$		
Collier County																		
Matched (score = 100)	70	47	81	91	77	101	1404	372	1435	6593	5216	6628	411	219	486	171	132	166
Tied (score = 100)	3	5	6	3	8	3	60	69	144	89	143	212	12	24	27	3	6	10
Matched (score < 100)	8	7	13	17	12	25	449	153	442	1728	1182	1702	28	23	74	7	7	3
Tied (score < 100)	2	1	3	3	0	5	16	37	72	88	139	501	2	5	10	0	0	5
Unmatched (score < 60)	44	67	24	38	55	32	691	1989	527	2618	4436	2073	238	420	94	8	44	5
Match rate (%)	65.35	47.24	81.10	75.00	63.82	80.72	73.63	24.08	79.89	76.45	60.09	81.35	65.56	39.22	86.40	95.77	76.72	97.35
	$n = 119$			$n = 206$			$n = 822$			$n = 9815$			$n = 891$			$n = 306$		
Seminole County																		
Matched (score = 100)	44	44	45	130	128	131	356	283	382	5299	5045	5606	356	264	431	238	208	241
Tied (score = 100)	1	2	2	2	4	7	28	44	22	141	47	159	22	29	25	6	0	11
Matched (score < 100)	18	17	24	25	28	32	137	127	209	1624	1591	1717	46	102	147	18	27	25
Tied (score < 100)	0	0	15	2	1	7	19	11	24	92	154	402	4	8	67	2	0	2
Unmatched (score < 60)	56	56	33	47	45	29	289	364	192	2659	2978	1931	463	488	221	42	71	27
Match rate (%)	52.94	52.94	72.27	77.18	78.16	85.92	65.14	56.09	76.84	72.91	69.66	80.33	48.04	45.23	75.20	86.27	76.80	91.18

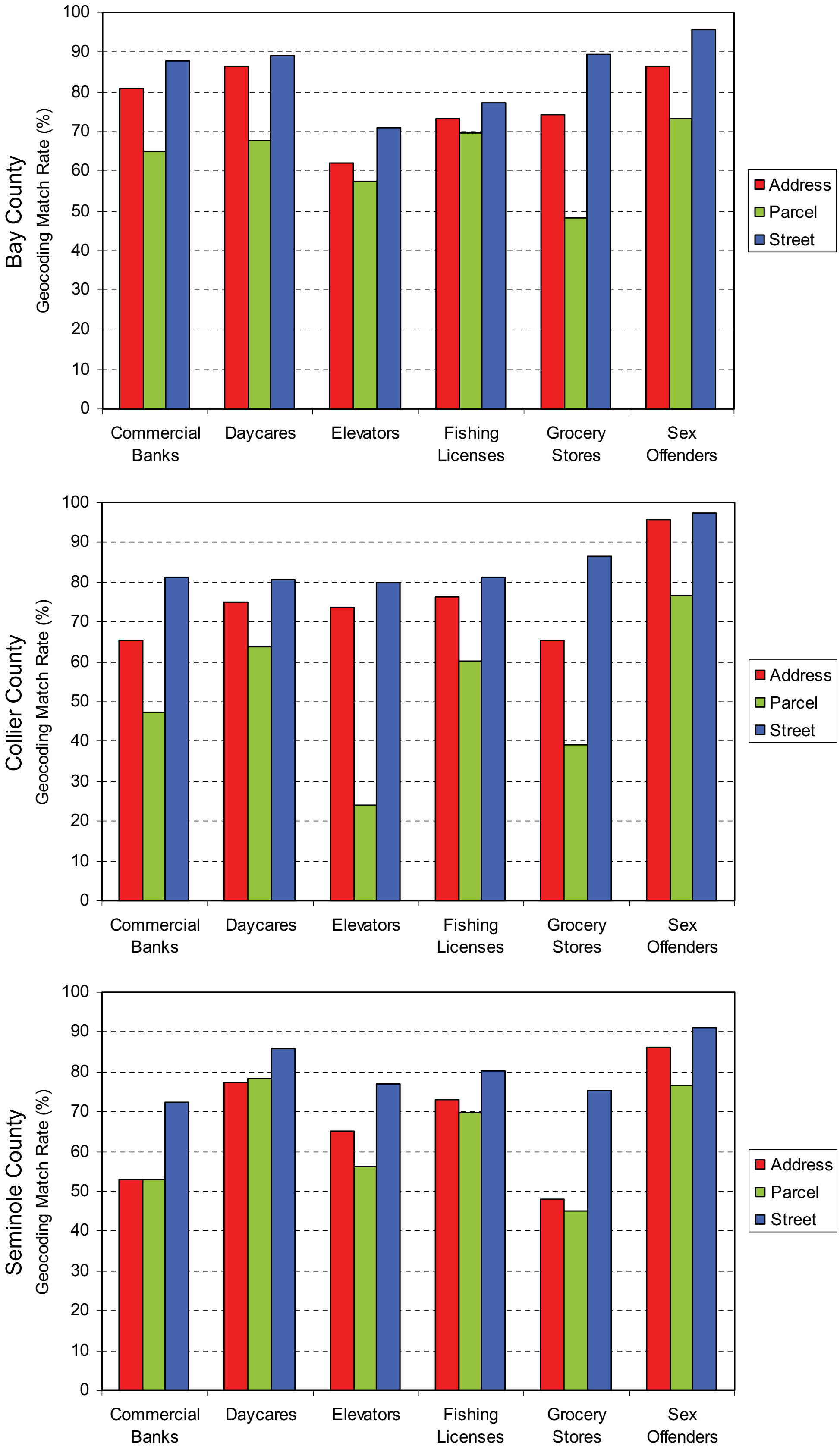

Fig. 6. Geocoding match rates by County.

published studies were identified that have employed address point geocoding in the United States.

6.4. Commercial versus residential

Some of the most difficult addresses to correctly geocode are commercial and multi-unit residential addresses. Table 4 shows a comparison of commercial properties with elevators and multi-unit residential properties with elevators. Results indicate that match rates for commercial properties are consistently lower than for residential properties for both address point and street geocoding. For residential properties the match rates for address points are only slightly lower than for street geocoding, suggesting that the address point reference data contains a fairly complete and reliable representation of multi-unit residential addresses. For commercial properties, the results for address point geocoding are not as good. Parcel geocoding results in much lower match scores for both types of properties, with dramatically low match scores for residential properties in Collier (13%) and Seminole County (12%). This confirms the poor performance of parcel geocoding for multi-unit residential addresses.

A second comparison between commercial and residential addresses is provided by the results for child care facilities in Table 5. In this case the residential addresses are mostly single family homes. Similar to the results for properties with elevators, the match rates for residential address are higher. In this case, however, the difference is much larger, with relative high match scores (>80%) for all three types of geocoding for all three Counties. The difference in match rates between the three types of geocoding is in fact quite small, suggesting that for single family residential address the choice of address data model does not influence match rates very strongly, in sharp contrast to multi-unit residential addresses.

6.5. Ties

Ties represent a concern in geocoding since they normally require manual inspection to determine which of the ties represents the correct match. Even with manual inspection, no determination may be possible due to ambiguities in either the reference data, the address input data, or both. A low number of ties, therefore, is an indication of a more reliable result.

Table 6 reports the number of ties as a percentage of all matches (score > 60) for all the address databases and each of the three geocoding techniques. Address point geocoding consistently produces the lowest number of ties, while the results for the other two techniques is more variable. The percentage ties for street geocoding is generally low for Bay County (<1%) but much higher for Collier (6–10%) and Seminole County (5–20%). Variability in the

Table 4
Comparison of geocoding match rates for commercial and residential properties with elevators

Category	Commercial elevators			Residential elevators		
	Address	Parcel	Street	Address	Parcel	Street
	$n = 316$			$n = 251$		
Bay County						
Matched (score = 100)	144	70	155	163	47	177
Tied (score = 100)	0	53	1	0	122	0
Matched (score < 100)	35	11	55	10	0	11
Tied (score < 100)	0	18	1	0	4	2
Unmatched (score < 60)	137	164	104	78	78	61
Match rate (%)	56.65	48.10	67.09	68.92	68.92	75.70
	$n = 1181$			$n = 1439$		
Collier County						
Matched (score = 100)	562	266	674	842	106	761
Tied (score = 100)	39	42	53	21	27	91
Matched (score < 100)	119	109	128	330	44	314
Tied (score < 100)	4	27	27	12	10	45
Unmatched (score < 60)	457	737	299	234	1252	228
Match rate (%)	61.30	37.60	74.68	83.74	13.00	84.16
	$n = 787$			$n = 42$		
Seminole County						
Matched (score = 100)	335	281	356	21	2	26
Tied (score = 100)	18	42	18	10	2	4
Matched (score < 100)	136	127	204	1	0	5
Tied (score < 100)	16	10	23	3	1	1
Unmatched (score < 60)	282	327	186	7	37	6
Match rate (%)	64.17	58.45	76.37	83.33	11.90	85.71

Table 5
Comparison of geocoding match rates for commercial and residential child care facilities

Category	Commercial child care			Residential child care		
	Address	Parcel	Street	Address	Parcel	Street
	$n = 54$			$n = 23$		
Bay County						
Matched (score = 100)	38	22	39	19	15	19
Tied (score = 100)	0	4	0	0	1	0
Matched (score < 100)	4	3	4	3	4	3
Tied (score < 100)	0	0	1	0	1	0
Unmatched (score < 60)	12	25	10	1	2	1
Match rate (%)	77.78	53.70	81.48	95.65	91.30	95.65
	$n = 104$			$n = 48$		
Collier County						
Matched (score = 100)	56	44	64	35	33	37
Tied (score = 100)	3	7	3	0	1	0
Matched (score < 100)	13	8	8	4	4	17
Tied (score < 100)	3	0	5	0	0	0
Unmatched (score < 60)	29	45	24	9	10	8
Match rate (%)	72.12	56.73	76.92	81.25	79.17	87.10
	$n = 124$			$n = 82$		
Seminole County						
Matched (score = 100)	66	61	64	64	67	67
Tied (score = 100)	2	4	4	0	0	3
Matched (score < 100)	18	21	24	7	7	8
Tied (score < 100)	2	1	7	0	0	0
Unmatched (score < 60)	36	37	25	11	8	4
Match rate (%)	70.97	70.16	79.84	86.59	90.24	95.12

Table 6
Ties as a percentage of all matches by geocoding method

County	Geocoding method	Commercial banks	Child care	Elevators	Fishing licenses	Grocery stores	Sex offenders
Bay County	Address	0.00	0.00	0.00	0.33	0.00	0.40
	Parcel	16.22	12.00	60.62	7.84	14.68	10.85
	Street	2.00	1.52	1.00	2.00	1.98	1.08
Collier	Address	6.02	5.26	3.94	2.08	3.09	1.66
	Parcel	10.00	8.25	16.80	4.22	10.70	4.14
	Street	8.74	5.97	10.32	7.88	6.20	8.15
Seminole	Address	1.59	2.52	8.70	3.26	6.07	3.03
	Parcel	3.17	3.11	11.83	2.94	9.18	0.00
	Street	19.77	7.91	7.22	7.12	13.73	4.66

percentage of ties for parcel geocoding is even greater, and generally higher than for street geocoding for both Bay and Collier County. Several very high values, including a value of 61% for elevators in Bay County, highlight the lack of reliability obtained using parcel geocoding. Ties are also higher in general for the commercial addresses, for all three geocoding techniques.

7. Conclusions

This study has provided an empirical comparison of address point, parcel and street geocoding. In general, match rates for address point geocoding are only slightly lower than for street geocoding. The higher rate for street geocoding could in part be due to false positives, but confirming this requires extensive field validation. Match rates using parcel geocoding are much lower, but this varies by database type and geographic area.

Substantial differences were observed between commercial and residential addresses and between different types of residential addresses. In general, higher match rates are obtained for residential addresses relative to commercial addresses. For single family residential addresses, match rates are relatively high for all three geocoding techniques considered. For multi-unit residential addresses, however, parcel geocoding is very unreliable, while results for both address points and street geocoding are much better.

Geocoding match rates were found to vary substantially by type of address database and by geographic area, suggesting that determining an "acceptable" or "good" match rate requires very context specific considerations. Variability in match rates between address models is only one of several considerations. The lack of consistency in match rates between geographic areas using the same type of address database and the same address model also suggests that geocoding quality is very much a function of the quality and consistency of local reference data. Substantial differences in match rates between the six different databases also suggest that the quality of the input data is a very critical contributor to the final geocoding match rate.

One of the limitations of this study is that the comparison of geocoding methods is limited to three Counties in the United States. Specific results in other jurisdictions may be different, but the general nature of the differences between the three address models is likely to be similar. It should also be noted that the chosen study areas reflect current best practice in terms of digital spatial data, and the availability of digital parcel boundary and address point data for use in a GIS environment is not yet widespread in the United States.

Of the three geocoding methods considered, street geocoding is the most widely employed. Online geocoding services (Google Maps, Yahoo Maps, MapQuest) rely almost exclusively on street geocoding, as do most commercial geocoding firms. Digital street reference data is available for nearly all areas within the United States and for many other jurisdictions. Street geocoding has also become very affordable, in many cases even free. Many commercial GIS packages have built-in tools and reference data for street geocoding. Parcel geocoding is becoming more widely used, but typically in studies that are limited in geographic scope. Digital parcel data is not available at the national or even State level within the United States, and has to be obtained directly from local government agencies. The most recent estimates suggest that only about 60% of all approximately 140 million parcels in the United States is available in a format that can be utilized in a GIS environment (Stage & Von Meyer, 2003). Even where available, utilizing parcel data requires considerable more skill and effort than street geocoding, in part because parcel data is not specifically designed with geocoding in mind. Address point data is not widely used in the United States, mostly because data availability is limited. Commercial firms report that approximately 40 million address points are available for the United States, covering selected metropolitan regions (ESRI, 2007; TeleAtlas, 2006). Where available, address points are relatively easy to use for geocoding in a GIS environment because geocoding is one of the principal objectives in the collection of address points by local governments. Address point data is available as a national dataset in both the United Kingdom and Australia, but to date no comparative analysis or quality assessment has been performed between address points from multiple jurisdictions.

Address points appear very promising as an address data model for geocoding. They represent excellent positional accuracy, produce match rates only slightly lower than those for street geocoding, and result in a low number of ties. In addition, they provide an extra validation of the address input data, since it is less likely a false positive will be introduced through a non-existing street number as may be the case for street geocoding. While it may only be a matter of time before address point data is available for most of the United States, standardization efforts would provide a logical framework for the development of a national address point database and an opportunity to learn from the efforts in other jurisdictions.

Future research efforts in this area should focus on refinements of the address point data model, such as the occurrence of multiple units with the same street number (currently represented as one address point), vertical representation of units, and consistency in the placement of address points. Possible refinements of the parcel data model consist of capturing multiple addresses within a single parcel, as well as residences with street addresses that are different from the legal street address of the parcel itself. Finally, improved quality control during the original capture of input data is paramount to improving geocoding match rates. Continued improvements in the address data models and reference data will be in vain unless address standardization and validation procedures during input are also improved.

References

Arctur, D., & Zeiler, M. (2004). *Designing geodatabases: Case studies in GIS data modeling*. Redlands, CA: ESRI Press.

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

Bow, C. J. D., Waters, N. W., Faris, P. D., Seidel, J. E., Galbraith, P. D., Knudtson, M. L., et al. (2004). Accuracy of city postal code coordinates as a proxy for location of residence. *International Journal of Health Geographics, 3*(5).

Burra, T., Jerrett, M., Burnett, R. T., & Anderson, M. (2002). Conceptual and practical issues in the detection of local disease clusters: A study of mortality in Hamilton, Ontario. *The Canadian Geographer, 46*, 160–171.

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

Christen, P., Churches, T., & Zhu, J. X. (2002). Probabilistic name and address cleaning and standardization. *The Australasian Data Mining Conference*, Canberra, Australia, December 3, 2002.

Christen, P., Churches, T., & Willmore, A. (2004). A probabilistic geocoding system based on a national address file. *The Australasian Data Mining Conference*, Cairns, Australia, December 6, 2004.

Christian, P. (1998). Soundex – Can it be improved. *Computer in Geneaology, 6*(5).

Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Medical Informatics and Decision Making, 2*(9).

Dearwent, S. M., Jacobs, R. J., & Halbert, J. B. (2001). Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis and Environmental Epidemiology, 11*, 329–334.

ESRI (2007). ArcGIS Business Analyst 9.2 now shipping. *ArcNews*, Spring 2007.

Goldberg, D. W., Wilson, J. P., & Knoclock, C. A. (2007). From text to geographic coordinates: The current state of geocoding. *URISA Journal, 19*(1), 33–46.

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. Technical report 03/83, CSIRO Mathematical and Information Sciences, Australia.

Jacquez, G. M., & Waller, L. (2000). The effect of uncertain locations on disease cluster statistics. In H. T. Mowrer & R. G. Congalton (Eds.), *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing* (pp. 53–64). Chelsea, Michigan: Arbor Press.

Jaro, M. (1984). *Record linkage research and the calibration of record linkage algorithms*. Bureau of the Census, Statistical Research Division Report Series, SRD Report No. Census/SRD/RR-84/27.

Karimi, H. A., & Durcik, M. (2004). Evaluation of uncertainties associated with geocoding techniques. *Computer-Aided Civil and Infrastructure Engineering, 19*, 170–185.

Patman, F., & Shaefer, L. (2001). *Is Soundex good enough for you? On the hidden risks of Soundex-based name searching*. Language Analysis Systems Inc.

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

Ratcliffe, J. H. (2001). On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science, 15*(5), 473–485.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B., Pavlik, C. E., West, M. W., et al. (2006). Geocoding in cancer research: A review. *American Journal of Preventative Medicine, 30*(2S), S16–S24.

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology, 17*(6), 464–470.

Stage, S., & Von Meyer, N. (2003). An assessment of parcel data in the United States. Federal Geographic Data Committee's Subcommittee on Cadastral Data.

Stanier, A. (1990). How accurate is Soundex matching? *Computers in Geneaology, 3*(7), 286–288.

Statistics Canada. (2002). *Statistics Canada postal code conversion file reference guide*. Statistics Canada, Ministry of Industry, Ottawa, ON 92F0153GIE.

Strickland, M. J., Siffel, C., Gardner, B. R., Berzen, A. K., & Correa, A. (2007). Quantifying geocode location error using GIS methods. *Environmental Health, 6*(10).

TeleAtlas (2006). *TeleAtlas Address Point product description*. TeleAtlas, Lebanon, OH.

Waller, L. A. (1996). Statistical power and design of focused clustering studies. *Statistics in Medicine, 15*, 765–782.

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., et al. (2005). Positional accuracy of two methods of geocoding. *Epidemiology, 16*(4), 542–547.

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

Whitsel, E. A., Quibrera, P. M., Smith, R. L., Catellier, D. J., Liao, D., Henley, A. C., & Heiss, G. (2006). Accuracy of commercial geocoding: Assessment and implications. *Epidemiological Perspectives and Innovations, 3*(8).

Winkler, W. E. (1999). *The state of record linkage and current research problems.* RR99/03, United States Bureau of the Census.

Zandbergen, P. A., & Green, J. W. (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives, 115*(9), 1363–1370.

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

Zhan, F. B., Brender, J. D., De Lima, I., Suarez, L., & Langlois, P. H. (2006). Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology, 16*(11), 842–849.

Zimmerman, D. L., Fang, X., Mazumdar, S., & Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics, 6*(1).

Zimmerman, D.L. (2007). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*, in press.