Spatial and Spatio-temporal Epidemiology 3 (2012) 69–82

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets

P.A. Zandbergen [b,*], T.C. Hart [a], K.E. Lenzer [b], M.E. Camponovo [b]

[a] Department of Criminal Justice, University of Nevada, Las Vegas, NV, United States
[b] Department of Geography, University of New Mexico, NM, United States

ARTICLE INFO

Article history:
Available online 11 February 2012

Keywords:
Geocoding quality
Positional accuracy
Error propagation

ABSTRACT

The quality of geocoding has received substantial attention in recent years. A synthesis of published studies shows that the positional errors of street geocoding are somewhat unique relative to those of other types of spatial data: (1) the magnitude of error varies strongly across urban–rural gradients; (2) the direction of error is not uniform, but strongly associated with the properties of local street segments; (3) the distribution of errors does not follow a normal distribution, but is highly skewed and characterized by a substantial number of very large error values; and (4) the magnitude of error is spatially autocorrelated and is related to properties of the reference data. This makes it difficult to employ analytic approaches or Monte Carlo simulations for error propagation modeling because these rely on generalized statistical characteristics. The current paper describes an alternative empirical approach to error propagation modeling for geocoded data and illustrates its implementation using three different case-studies of geocoded individual-level datasets. The first case-study consists of determining the land cover categories associated with geocoded addresses using a point-in-raster overlay. The second case-study consists of a local hotspot characterization using kernel density analysis of geocoded addresses. The third case-study consists of a spatial data aggregation using enumeration areas of varying spatial resolution. For each case-study a high quality reference scenario based on address points forms the basis for the analysis, which is then compared to the result of various street geocoding techniques. Results show that the unique nature of the positional error of street geocoding introduces substantial noise in the result of spatial analysis, including a substantial amount of bias for some analysis scenarios. This confirms findings from earlier studies, but expands these to a wider range of analytical techniques.

1. Introduction

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 most common type of geocoding is address geocoding where the input consists of street addresses. The quality of geocoding has received substantial attention in recent years and it has been recognized that errors in geocoding need to be understood in order to determine the robustness of spatial analysis techniques which employ the results of geocoding. The general purpose of this paper is threefold: (1) to synthesize the existing knowledge on the nature of positional errors in geocoding; (2) to present a framework for modeling the effect of these errors on the spatial analysis of geocoded datasets; and (3) to present several case-studies that illustrate the implementation of this framework.

[*] Corresponding author. Address: Department of Geography, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131, United States. Tel.: +1 505 277 3105.
E-mail address: zandberg@unm.edu (P.A. Zandbergen).

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2012.02.007

2. Background

2.1. Geocoding foundations

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. 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 and positional error in the resulting spatial datasets (e.g., Rushton et al., 2006; Goldberg et al., 2007).

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. Several common address models exist, each with a particular set of supporting materials and characteristic errors. Commonly used address models include street networks, parcels, and address points which have been reviewed by Zandbergen (2008a, 2009). Street networks have historically been the most widely employed address data model, especially in the US. 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. Many different reference datasets are available for this type of geocoding.

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 jurisdictions in the US (Rushton et al., 2006). To overcome the limitations of parcels for geocoding, address points have emerged as an alternative address data model. Address points typically represent the locations of all addressable structures within a jurisdiction and are created from a combination of primary field data collection (GPS, field surveys) and secondary data interpretation (parcels, imagery, building footprints). In the US, address point geocoding is not yet in very widespread use. However, many local governments have started to create address point databases and several commercial geocoding firms provide address point geocoding for selected coverage areas.

2.2. Geocoding quality

A substantial body of literature has emerged on the quality of datasets obtained through address geocoding. 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 actual location of the structures of interest. 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.

The focus of the current paper is the positional accuracy of geocoded locations, defined as the Euclidean distance between the geocoded point location and the actual location of the structure associated with the address. Different components contribute to the error, including: (1) match to an incorrect street segment; (2) incorrect placement along the street segment; (3) incorrect offset from the street segment; and (4) positional error in the street segment. In most empirical studies, these components are not addressed separately and the measured error is therefore the aggregate effect of all four components. Several empirical studies in recent years have determined the positional accuracy of street geocoding, as reviewed by Zandbergen (2009) (Table 1). Despite differences in the design of the various studies, several general observations can be made as follows:

1. *The magnitude of positional errors varies strongly along urban–rural gradients.* Based on the review of published studies by Zandbergen (2009) using median values the "typical" positional error for residential addresses ranges from 2201 m. This is a very broad range and much of this can be attributed to differences across urban–rural gradients. 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. Several other studies have found similar differences, confirming a clear general trend that geocoding is much more accurate in urban areas compared to rural areas.
2. *The distribution of the magnitude of positional errors of street geocoding does not follow a normal distribution.* Formal testing by Zandbergen (2008b) has shown that the distribution approximates a log-normal distribution when the distribution of the direction of errors is uniform. In a similar study, Zimmerman et al. (2007) have shown that mixtures of bivariate *t* distributions with two or three components are required to characterize the distribution of the magnitude of positional error when the distribution of the direction is strongly influenced by the gridded nature of the street network.
3. *The direction of positional errors of street geocoding* (*i.e.*, the angle in degrees of the line connecting the actual structure with the geocoded location) *is not random and is closely related to the local properties of the street network.* Studies to date reveal mixed results with several finding no significant difference from a uniform distribution (Cayo and Talbot, 2003; Strickland et al., 2007) while others finding a significant difference (Schootman et al., 2007; Zimmerman et al., 2007). However, aggregate statistics for direction ignore local

Table 1
Error matrix for detailed land cover classification for Jackson County, OR.

		Geocoded locations (StreetMap USA)														
		Open water	Developed, open space	Developed, low intensity	Developed, medium intensity	Developed, high intensity	Barren land (rock/sand/ clay)	Deciduous forest	Evergreen forest	Mixed forest	Shrub/ scrub	Grassland/ herbaceous	Pasture/ hay	Cultivated crops	Woody wetlands	Total
Reference locations	Open water		1													1
	Developed, open space		26	33	4				3		2	2	4		1	75
	Developed, low intensity		31	325	80	1			1		2		3	2		445
	Developed, medium intensity		4	61	107	13						1	1			187
	Developed, high intensity			2	5	3										10
	Barren land (rock/sand/clay)															0
	Deciduous forest		2													2
	Evergreen forest		13	8	1				8		9	1				40
	Mixed forest		3						1		2					6
	Shrub/scrub		36	8	1				1	1	19	4	4			74
	Grassland/herbaceous		22	7			1				5	3			2	40
	Pasture/hay		37	38	6	1						1	18			101
	Cultivated crops		6	5	1								2			14
	Woody wetlands		1	2	1										1	5
	Total	0	182	489	206	18	1	0	14	1	39	12	32	2	4	1000

patterns. Since 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 Carrol County, IA, where the road network forms a near perfect grid of segments running East–West and North–South.

4. *Positional errors in geocoding are spatially autocorrelated.* There is evidence that this spatial autocorrelation is related to the properties of the local street network used for reference data (Zimmerman and Li, 2010; Zimmerman et al., 2010). However, this has received limited attention in the research to date and at present there is no widely applicable spatial autocorrelation model that can be applied to different study areas.

2.3. 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, 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 the geocoding (Zandbergen, 2009). The errors in geocoded locations may adversely affect spatial analytic methods which has started to receive attention in the literature. 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 250–500 m (Zandbergen, 2007; Zandbergen and Green, 2007). 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 over-estimate the number of potentially exposed children at small distances up to 250 m. 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 over-estimation. Wu et al. (2005) found that the use of road networks of poor positional accuracy also leads 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 since the distances of concern are very short. 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 (TRI) 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 shockingly 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: (1) magnitude of the distances considered; and (2) type of spatial analysis technique employed.

For many applications, the purpose of geocoding addresses is to associate the individual-level information with the properties of area-level information, such as demographic and other socio-economic variables using census enumeration units. Several studies have determined the effects of positional error on the assignment of individual location to the correct polygon boundaries. 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. 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. The percentage of incorrectly placed locations was substantially higher in rural areas.

There have also been a number of studies in the crime literature on the effects of positional errors in geocoding 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 (Bilcher and Balchak, 2007). 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 meters). 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 depend on the nature of the analysis. This includes both the scale (or spatial resolution) of the analysis, as well as the specific nature of the analysis technique. As with any spatial-analytical technique, a good understanding of the input data quality and knowledge of the robustness of the technique itself are 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 active research. The current study seeks to contribute to this body of knowledge by developing a framework for error propagation modeling of the effect of positional errors in geocoding.

3. Framework for error propagation modeling

Error propagation modeling makes it possible to assess the error in analysis outputs resulting from the propagation of analysis input error through the processing and analysis steps (Heuvelink, 1998; Lanter and Veregin, 1992). Several techniques exist for spatial error propagation modeling. Analytical solutions can sometimes be developed, but these are only available for a limited number of relatively simple spatial processes. Most techniques employ stochastic variables in Monte Carlo simulation modeling. In this approach, multiple realizations of input variables are created using a stochastic description of these variables based on an understanding of the (spatial) nature of the error. The analysis is then carried out on the set of realizations (typically several hundred), and sample statistics are determined for the set of outcomes. This approach has become widespread in several domains of spatial science (e.g., Arbia et al., 1998; Karssenberg and De Jong, 2005; Stanislawski et al., 1996). For example, in the area of terrain modeling, the Monte Carlo simulation approach has become very widely adopted to determine the sensitivity of terrain derivatives to errors in the terrain data inputs (e.g., Eclschlaeger, 1998; Hengl et al., 2010; Lindsay, 2006; Lindsay and Evans, 2008;Oksanan and Sarjakowski, 2005; Zandbergen, 2010). This can include relatively sophisticated geostatistical methods to estimate the nature of spatial autocorrelation to be employed in the simulation model (e.g., Hengl et al., 2010).

Despite the widespread use of Monte Carlo simulation for spatial error propagation modeling, it is not without its limitations. For example, Heuvelink (2002) has pointed out that most errors in spatial datasets are spatially correlated and cannot be characterized by a single accuracy metric. Other issues which represent challenges include

Table 2
Error matrix for detailed land cover classification for Washington, DC.

		Geocoded locations (StreetMap USA)										
		Developed, open space	Developed, low intensity	Developed, medium intensity	Developed, high intensity	Barren land (rock/sand/clay)	Deciduous forest	Evergreen forest	Pasture/hay	Cultivated crops	Woody wetlands	Total
Reference locations	Developed, open space	14	21	5			1				1	42
	Developed, low intensity	24	133	99	1		2	1	1			261
	Developed, medium intensity	6	79	432	54					1		572
	Developed, high intensity		2	44	49							95
	Barren land (rock/sand/clay)	3					3				1	7
	Deciduous forest	1	5				3					9
	Evergreen forest	1					1					2
	Pasture/hay	2										2
	Cultivated crops					1	1		1			3
	Woody wetlands	2	1								4	7
	Total	53	241	580	104	1	11	1	2	1	6	1000

the use of multiple uncertain inputs, the difficulty arising from incorporating uncertainty in categorical inputs with a large number of classes, and the scale dependency of modeling results (Heuvelink, 2002).

Implementing Monte Carlo simulation for spatial error propagation requires a very good understanding of the error in the input data. The general nature and magnitude of positional errors in geocoding are relatively well understood as discussed in the previous section. However, several characteristics limit the use of Monte Carlo simulation. First, the error distribution is distinctly non-normal, while most stochastic approaches are built on assumptions of normality. Second, the nature of the spatial autocorrelation of the error is not well understood. As a result, for a given set of geocoded locations, it is not (yet) possible to create a realistic error model for use in Monte Carlo simulation.

One alternative approach has been presented by Jacquez and Rommel (2009). In this approach, the sensitivity of analysis results to perturbations (resulting from positional error) is calculated as a local statistic. The proposed local indicators of geocoding accuracy (LIGA) method is the only formalized error propagation model to date specifically developed for positional errors in geocoding. Despite its merit, the method is somewhat limited because: (1) it employs a Gaussian error model, which is not very realistic for positional errors in geocoding; (2) sensitivity to perturbations is determined using a spatial weights matrix, which limits the range of spatial-analytical techniques to be explored; and (3) spatial autocorrelation of the errors is not explicitly incorporated.

The current study proposes an alternative error propagation model for examining the effects of positional errors in geocoding. A brief description of the framework follows and is also illustrated in Fig. 1.

The proposed framework for error propagation modeling is built around the use of address point datasets. Address point datasets contain the coordinates for all addressable structures within a jurisdiction and have typically been created for the specific purpose of geocoding. Positional accuracy of address points is very good and geocoding match rates are often very similar to those obtained using street geocoding (Zandbergen, 2008a). A table with individual-level addresses is geocoded using address points to create a set of reference locations (RL). The same addresses are also geocoded using one or more alternative geocoding techniques to create sets of geocoded locations (G1, G2, etc.). To account for differences in matched records, the sets of locations are filtered to maintain only those records which produced a match for all geocoding techniques. This makes it possible to examine the effect of positional accuracy while controlling for differences in matched records. Once the matching sets of locations are created, all sets are run through the same spatial analysis procedure, for example point-in-raster overlay, point-in polygon overlay, proximity or network analysis, or local clustering. Fig. 1 shows only three examples for illustration. This produces analysis results for each set of locations. Error estimates are developing through pair-wise comparison between the results for the reference locations and the results for each set of geocoded locations. The nature of this

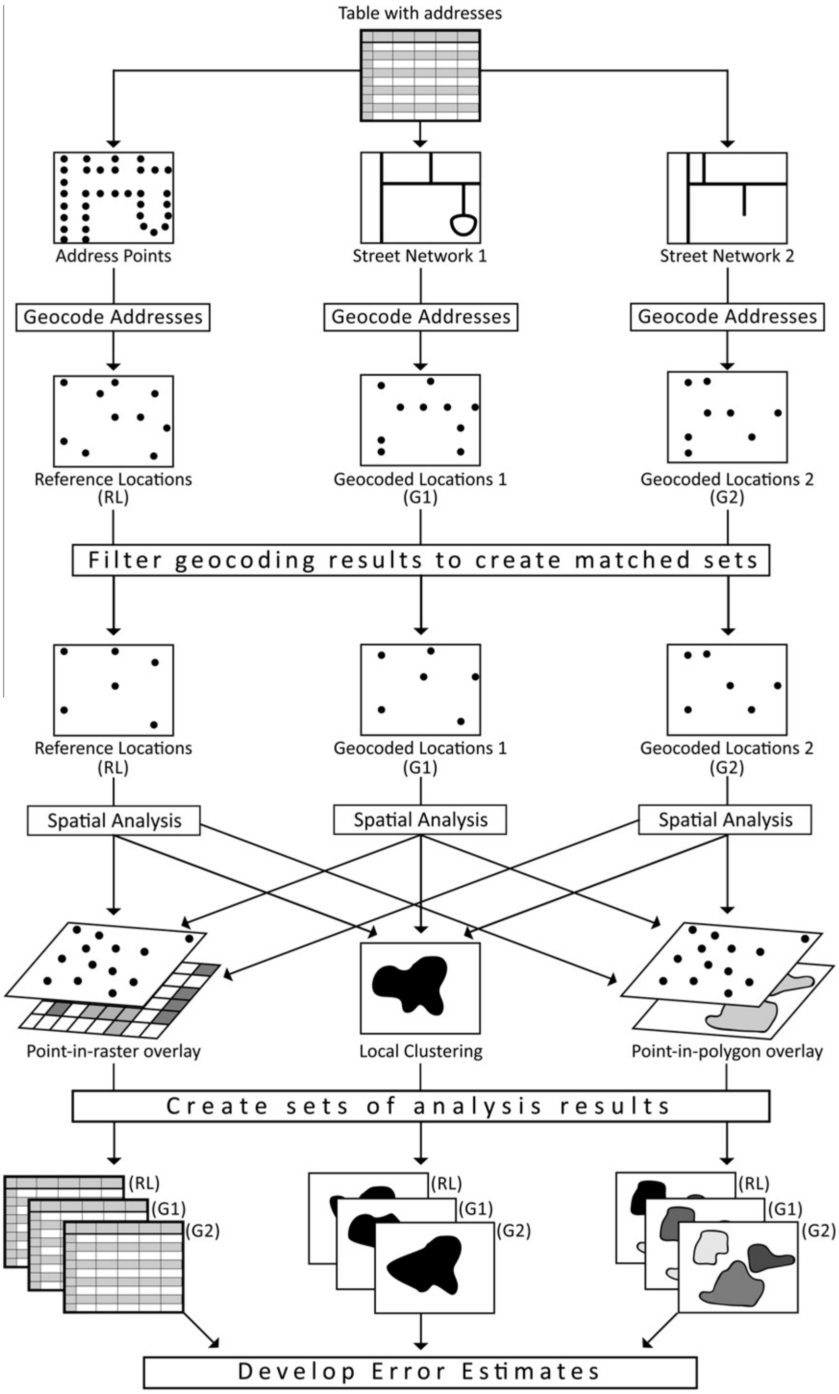

Fig. 1. Conceptual illustration of the error propagation modeling approach to examine the effects of geocoding quality on spatial analysis.

error analysis will vary with the type of spatial analysis being explored, but will typically exist of tabular and/or cartographic measures of agreement of the results.

The proposed model has a number of strengths:

- It does not make any assumptions about the nature of the positional error in geocoding and is strictly based on an empirical error model for the study area.
- It is relatively simple and does not require any complex geostatistical characterizations to determine parameters for the error model.
- It can be applied to any spatial-analytical technique that employs a geocoded set of points as one of the inputs.
- It does not require field work or other resource intensive methods to validate the accuracy of geocoded locations and can therefore be applied to very large datasets.

The proposed model also has a number of assumptions/limitations:

- The approach requires high quality address point data for at least a portion of the study area.
- Only those locations present in address points are used, meaning that records with poor address quality will be excluded. This may result in a somewhat conservative error model.
- It does not address issues of geocoding completeness, although it could be modified to incorporate this.
- It does not incorporate any errors in the other inputs of the spatial analysis, although it could be extended to incorporate this additional complexity using Monte Carlo simulation for those inputs if a robust error model is available.

Since the model relies on high quality address point datasets for the study area, this beckons the question why another less accurate form of geocoding would be used to create geocoded locations for the study area? First, address point coverage is limited in the US. For many datasets at the state or national scale, address points for a small representative portion of the study area may be available for error propagation modeling, but the entire dataset still needs to be geocoded using regular street geocoding methods. Second, some types of addresses geocode poorly using address point as reference data (e.g., block-side addresses, intersections), making it unlikely that address point geocoding would be the only type of geocoding employed for a particular dataset.

Following are the methods and results for three case-studies to illustrate the implementation of the proposed error propagation model.

4. Methods

4.1. Case-study 1: point-in-raster overlay

The first case-study consists of a point-in-raster overlay. Increasingly, individual-level health datasets are geocoded and then overlaid with raster datasets derived using remote sensing, such as imagery, climate variables, and land cover. For example, Estes et al. (2009) used land cover and land surface temperature data to determine the relationship between living environments and blood pressure. Overlaying geocoded locations with raster datasets presents a promising approach to determine the relationship between environmental variables and health outcomes. The case-study determines how sensitive the land cover associations for residential addresses are to positional errors in the geocoded locations.

Six US counties were selected as study areas, reflecting a range of different population densities and land cover types: Clay County, MN; Jackson County, OR; New Hanover County, NC; San Francisco County, CA; Travis County, TX; and Washington, DC. Land cover datasets were obtained in the form of the 2006 National Land Cover Dataset (NLCD). These data are derived from Landsat imagery and consist of raster datasets at 30-meter resolution. The NLCD 2006 data is provided in a continental Albers projection. To avoid issues resulting from raster resampling, all other datasets were projected into the same projection prior to analysis.

For each study area, address point datasets were obtained from local jurisdictions. Descriptive attributes for the address points were used to select only those points representing residential structures. These residential address points were then stripped of their coordinates to result in a table with only the residential address information. This table of addresses was geocoded using StreetMap USA 2009, a widely used address locator that is distributed with the ArcGIS software. The underlying reference data is based on TIGER 2000 but has been "enhanced" by commercial data providers. Minimum match score was set to 80 to ensure only very good matches were included in the final results. For each study area a random selection of 1000 geocoded addresses was created, resulting in two sets of 1000 points: one set representing the original address points (reference locations) and one set representing the geocoded locations using the StreetMap USA geocoder.

Both sets of 1000-point locations were assigned a land cover category using a point-in-raster overlay with the NLCD 2006 data. Two versions of the land cover were used: (1) the original NLCD land cover classes, which includes four different urban land cover types and many non-urban types; and (2) a more generalized version of the land cover data, reclassified into urban (high and medium density developed), suburban (low density and open space developed) and rural (all non-urban categories). The results for the two sets of point locations were compared using an error matrix and associated measures of agreement.

4.2. Case-study 2: kernel density hotspots

The second case-study consists of a local clustering analysis of individual events. Kernel density was selected as the local clustering technique, which employs a search radius to determine the local density (in events per square km) of a point pattern. Kernel density has been widely employed in criminology, epidemiology, and spatial ecology.

The event data for this illustration consists of crime events. Specifically, crime incidents in 2008 that occurred in Charlotte, NC were obtained from the Charlotte-Mecklenburg Police Department. Offense codes were used to limit the event data to assaults, resulting in a total of 10,793[1] incident records. The crime event locations were geocoded using five different reference datasets: (1) local address points from Mecklenburg County; (2) local street centerlines from Mecklenburg County; (3) commercial street network form TeleAtlas; and (4) street networks from StreetMap USA 2009. Address locators were created in ArcGIS in similar fashion for each of the four reference datasets and crime

[1] A total of 1988 records were removed from the assault data file because the incident locations for these records were recorded as an intersection. Since address point datasets do not typically contain intersections, the incidents with only an intersection recorded in the address field were removed prior to geocoding. This reduced the total number of assault records that could have been successfully geocoded to 8805.

incidents were geocoded using each of these four address locators. A subset of incidents was created which successfully geocoded using all five address locators ($n = 4915$). This subset was used to examine the effect of positional accuracy on kernel density hotspots, using the results of address point geocoding as the reference scenario.

Kernel density hotspots were created for each of the four sets of geocoded crime incidents. A final hotspot map was created from the kernel density grid by selecting those areas where the density is at least three times the mean density, after removing the area with a density of zero. This threshold reflects a relatively common approach in crime hotspot analysis (Eck et al., 2005). The analysis was repeated using five different search radii (100, 200, 300, 400 and 500 meters) to examine the effect of this parameter on the hotspot delineation. The threshold parameter was kept constant across the search radii.

The final hotspot maps for a given search radius were compared to determine the effect of positional error on hotspot delineation. The hotspot maps derived from address point geocoding were used as a reference and compared to the results from the other geocoding techniques. Quantitative pair-wise comparisons were made by determining areas of agreement, false positives, and false negatives.

4.3. Case-study 3: point-in-polygon overlay

The third and final case-study presents a point-in-polygon analysis, which is widely employed to create associations between individual-level and area-level datasets. The case-study employs the example of the residential address of public school children and the school zone boundaries used to assign attendance.

A dataset of children attending public school in Orange County, FL for the 2004/2005 school year was obtained from the Orange County Public Schools (OCPS). The dataset includes residential address, racial/ethnic classification based on self-identification by the parents (Asian, Black, Hispanic or White), and grade level. School zone boundaries for elementary, middle and high schools for the same school year were also provided by OCPS.

The residential addresses were geocoded using three different reference datasets: (1) local address points from Orange County; (2) local street centerlines from Orange County; and (3) street networks from TIGER 2000. Address locators were created in ArcGIS in similar fashion for each of the three reference datasets and residential addresses were geocoded using each of these three address locators. A subset of locations was created which successfully geocoded using all three address locators ($n = 120,122$). This subset was used to examine the effect of positional accuracy on point-in-polygon overlay, using the results of address point geocoding as the reference scenario.

Each geocoded dataset was split based on school type (elementary, middle, high) and aggregated to the corresponding school boundaries using a point-in-polygon overlay method. For each individual student, a determination was made whether or not the street centerline and TIGER geocoded location resulted in an allocation to the correct school, as indicated by the results from address point geocoding. Results were summarized by school type and race/ethnicity. The percentage error (again using the results from address points as the reference) was also determined for each individual school.

5. Results and discussion

5.1. Case-study 1: point-in-raster overlay

The first case-study examines the effect of geocoding quality on the land cover associations of residential addresses for six different study areas. An error matrix was created for each study area to determine the agreement between the results for the reference locations and the geocoded locations using StreetMap USA. Table 1 and 2 show the error matrices for Jackson County, OR and Washington, DC, respectively, using the original NLCD land cover classes. Cells on the diagonal axis show counts of addresses for which the land cover associations are correct, while off-diagonal cells are incorrect. For Jackson County, OR, 510 addresses were associated correctly (51.0%), while for Washington, DC this number is 635 (63.5%). Two general categories of errors occur in the error matrices in Table 1 and 2. In the first category, reference locations in urban land cover types are associated with different urban land cover classes, for example developed/low intensity vs. developed/high intensity. This category can be referred to as "minor" errors. This category is common in both study areas. In the second category, reference locations located in non-urban land cover types (forest, shrub, grassland, pasture) are associated with urban land cover types, most notably developed/open space. This category can be referred to as "major" errors. This category is common in Jackson County, OR, but nearly absent in Washington, DC.

Tables 3 and 4 show the error matrices for the same study areas for the generalized urban–rural land cover classification. In Jackson County, OR rural locations associated with suburban land cover are most common, while in Washington, DC suburban locations associated with urban land cover are most common. In general, reducing the number of land cover classes has a modest effect on the accuracy.

The error matrices for the other study areas are not shown, but in general follow some of the same general patterns. Results for Clay County, MN, New Hannover County, NC and Travis County, TX resemble those of Jackson County, OR, while the results for San Francisco County, CA resemble those of Washington, DC. This reflects the differences in population density and the land cover composition of the study areas.

Table 5 represents summary agreement measures for the six study areas, including accuracy (%), kappa index of agreement and weighted kappa index of agreement. Accuracy for the detailed land cover varies from a low of 46.4% for New Hannover County, NC to a high of 72.9% for San Francisco County, CA. For all study areas, the accuracy increases when considering the more generalized land cover, confirming the effect of granularity of the land cover. Values for kappa follow the same general pattern as for accuracy. Values for the weighted kappa are slightly higher, as

Table 3
Error matrix for generalized urban–rural land cover classification for Jackson County, OR.

		Geocoded locations (StreetMap USA)			
		Urban	Suburban	Rural	Total
Reference locations	Urban	128	67	2	197
	Suburban	85	415	20	520
	Rural	11	189	83	283
	Total	224	671	105	1000

Table 4
Error matrix for generalized urban–rural land cover classification for Washington, DC.

		Geocoded locations (StreetMap USA)			
		Urban	Suburban	Rural	Total
Reference locations	Urban	579	87	1	667
	Suburban	105	192	6	303
	Rural	0	15	15	30
	Total	684	294	22	1000

expected, since they reduce the influence of "minor" errors relative to "major" errors. The exception to this is Clay County, MN where the values for weighted kappa are lower as a result of a large number of these major errors. Land cover in Clay County, MN is dominated by agriculture, but many of these rural addresses are associated with urban land cover types using the geocoded locations.

Several of the results follow expected patterns. First, errors vary across study areas based on population density and land cover composition, with high density urban areas dominated by relatively homogenous urban land cover types (e.g., San Francisco County, CA) resulting in fewer errors compared to lower density areas with a heterogeneous mix of urban, rural and forested land cover types (e.g., Jackson County, OR). Second, errors are reduced when reducing the granularity of the land cover data, i.e., reclassifying the original land cover types into categories of urban, suburban and rural.

However, one result in particular requires further inspection since it is somewhat unexpected. For several study areas, in particular Clay County, MN and Jackson County, OR, there are many rural addresses which are associated with urban land cover types using the geocoded locations ("major" errors). Fig. 2 illustrates two sample study areas within Jackson County, OR to explore the underlying reason for this result. Fig. 2a shows a rural area where the reference locations are associated with land cover types such as shrub, grassland and pasture. The corresponding street geocoded locations are at some distance from the reference locations and are associated with developed/open space and developed/low intensity. Geocoded locations are placed in relatively close proximity to the road network and roads are typically classified as urban land cover types in the NLCD data. As a result, many rural addresses are associated with urban land cover types. In contrast, Fig. 2b shows an urban area within Jackson County, OR. Positional errors are typically somewhat smaller, but more importantly the study area is dominated by various types of urban land cover. Most errors are therefore limited to the "minor" errors discussed above. These results are significant since they reveal a strong bias in the land cover associations. For study areas with a mix of urban and rural land cover types, rural address are more likely to be associated with urban land cover types than vice versa; this is evident in Tables 1 and 3 for Jackson County, OR. No such bias is observed for study areas dominated by urban land cover, as is evident in Tables 2 and 4 for Washington, DC. Therefore, increased land cover heterogeneity not only introduces more error, it also introduces substantial bias.

A confounding factor in the use of the NLCD data is that the land cover data itself contains errors. While no formal accuracy assessment of NLCD products has been completed, initial accuracy estimates based on cross-validation suggest an overall average accuracy across all mapping zones of 83.9% (Homer et al., 2007). Accuracy is likely to be lower in areas with very heterogeneous land cover or containing linear features which result in mixed pixels. NLCD data are therefore expected to be less accurate in the vicinity of roads which are also the areas where geocoded locations tend to be. This contributes substantially to the overall error and the specific bias observed in land cover association for geocoded locations.

5.2. Case-study 2: kernel density clusters

The second case-study examines the effect of geocoding quality on the delineation of local clustering based on

Table 5
Summary of errors of land cover classifications in residential addresses.

	Clay, MN	Jackson, OR	New Hannover, NC	San Francisco, CA	Travis, TX	Washington, DC
Population density (#/sq. mile)	48.7	64.7	782.3	16,470.2	794	8379.4
Detailed land cover						
Accuracy (%)	57.9	51.0	46.4	72.9	54.3	63.5
Unweighted kappa	0.288	0.322	0.213	0.514	0.360	0.385
Weighted kappa[a]	0.214	0.428	0.304	0.541	0.387	0.524
Urban/suburban/rural						
Accuracy (%)	69.9	62.6	71.2	95.0	73.0	78.6
Unweighted kappa	0.237	0.352	0.359	0.520	0.393	0.529
Weighted kappa[a]	0.173	0.415	0.386	0.523	0.393	0.561

[a] Weighted kappa considers a ranking of land cover categories. Land cover categories were ranked from developed/high intensity to open water and linear weighting was employed in the calculation of kappa. This effectively means that minor errors are given less weight in the determination of agreement relative to major errors.

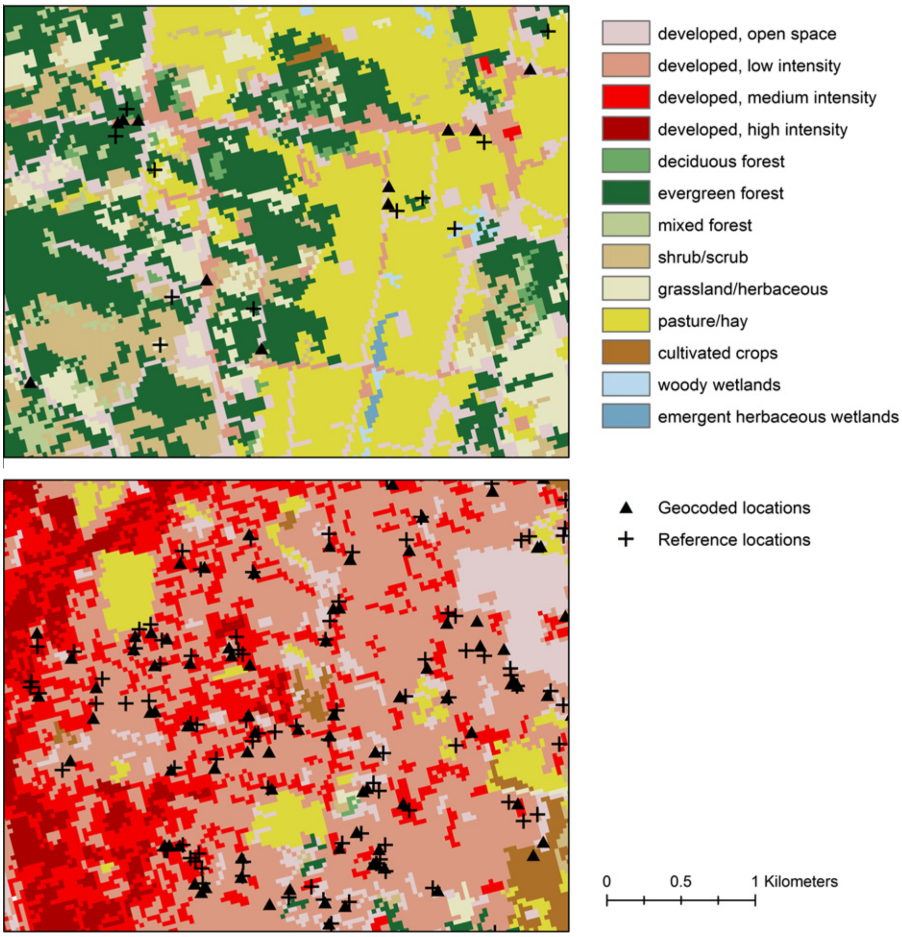

Fig. 2. Reference and geocoded locations overlaid on NLCD 2006 land cover raster datasets. Fig. 2a (top) shows a rural area while Fig. 2b (bottom) shows an urban area.

kernel density hotspot mapping. Fig. 3 shows a sample of results for a small portion of the entire study area. A total of nine pair-wise comparisons are shown, in each case consisting of the clusters boundaries from the reference locations ("reference clusters") and those from the geocoded locations ("geocoded clusters"). Results are shown for three different geocoding techniques and three different values for the kernel density bandwidth.

Errors in the boundary delineation are most clearly visible for the 100 m bandwidth. For all three geocoding techniques, several smaller geocoded clusters appear shifted relative to the reference clusters. This is a direct effect of the misplacement of street geocoded locations along the street segments. Clusters become smoother with larger values for the bandwith and visually the errors in the geocoded clusters are reduced. Fig. 3 also shows that for a given bandwidth the overall area of the clusters is very similar, despite differences in the number and shape of clusters. This is to be expected for kernel density clusters derived from point datasets with the same sample size.

The errors illustrated in Fig. 3 are described more quantitatively in Table 6. The area of agreement (A) represents the overlap between the reference clusters and the geocoded clusters. The area of false negatives (FN) represents the portion of reference clusters not overlaid by the geocoded clusters. Conversely, the area of false positives (FP) represents the portion of the geocoded clusters not overlaid by the reference clusters. An overall error metric is calculated as (FN + FP)/A. Higher values for this metric indicate larger errors and a value of zero would indicate the absence of error. Results in Table 6 indicate that the accuracy of the geocoded clusters for a bandwidth value of 100 m is very poor. The combined area of false negatives and false positives is several times larger than the area of agreement, regardless of reference data type. For larger bandwidth values, the accuracy gradually improves and the error metric drops to around 0.25–0.32 for the 500 m bandwidth scenario. This corresponds to the visual interpretation in Fig. 3, which suggests the error in the boundaries is relatively minor for a bandwidth of 500 m.

The results for the different geocoding techniques are very similar for a given bandwidth, suggesting that this plays a minor role in the accuracy of the geocoded clusters. For this particular dataset, it has been established that geocoding to local street centerline reference data is the most accurate, closely followed by TeleAtlas, while results for

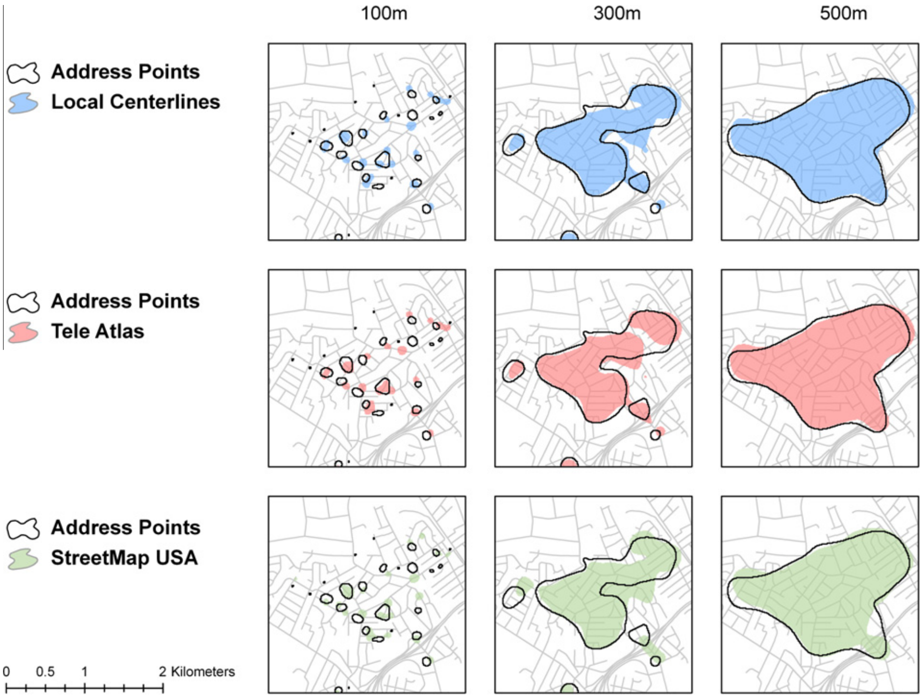

Fig. 3. Examples of kernel density hotspots resulting from geocoded locations compared to reference hotspots for varying kernel bandwidth values.

Table 6
Error analysis of kernel density hotspots for several geocoding techniques using varying values of kernel bandwidth.

Kernel bandwidth	Geocoding method	Area of agreement (A) (km^2)	Area of false negative (FN) (km^2)	Area of false positive (FP) (km^2)	(FN + FP)/A
100 m	Local street centerlines	1.66	2.82	2.67	3.315
	TeleAtlas	1.66	2.82	2.80	3.390
	StreetMap USA	1.51	2.96	2.91	3.882
200 m	Local street centerlines	9.10	4.90	4.31	1.012
	TeleAtlas	9.07	4.92	4.59	1.049
	StreetMap USA	8.62	5.38	4.86	1.188
300 m	Local street centerlines	19.05	5.49	4.87	0.544
	TeleAtlas	18.92	5.62	4.87	0.555
	StreetMap USA	18.13	6.41	5.64	0.665
400 m	Local street centerlines	28.84	5.64	4.98	0.368
	TeleAtlas	29.12	5.36	5.02	0.356
	StreetMap USA	27.91	6.56	5.78	0.442
500 m	Local street centerlines	37.65	5.47	4.97	0.277
	TeleAtlas	38.39	4.74	4.88	0.251
	StreetMap USA	37.01	6.12	5.76	0.321

StreetMap USA are much less accurate (Hart and Zandbergen, 2011). The error metric in Table 6 closely follows this pattern with very similar values for street centerlines and TeleAtlas, and slightly larger values for StreetMap USA. However, the differences between street geocoding techniques for a given bandwidth are much smaller than those between bandwidth values for a given geocoding technique. The effect of the positional error of geocoding is therefore very dependent on the scale of the analysis (in this case the kernel density bandwidth), while differences in geocoding techniques play a minor role. Combined with the results in Fig. 3, this also suggests that the errors in geocoding clusters from different geocoding techniques are very consistent, despite differences in the positional accuracy of these techniques.

5.3. Case-study 3: point-in-polygon overlay

The third case-study examines the effect of geocoding quality on the allocation of public school children to school zone boundaries using a point-in-polygon overlay operation. Table 7 provides a summary of the results, by school type, racial/ethnic category and geocoding type. The percentages indicate how many of the students are not

Table 7
Percentage of incorrectly allocated public school students by school type and racial/ethnic category.

Elementary schools	Asian (*n* = 2021)	Black (*n* = 18,330)	Hispanic (*n* = 18,218)	White (*n* = 21,350)	Total (*n* = 59,919)
Street centerlines	0.64	2.92	1.56	0.74	1.65
TIGER 2000	6.93	5.64	5.80	5.03	5.51
Middle schools	(*n* = 1023)	(*n* = 8464)	(*n* = 8104)	(*n* = 10,938)	(*n* = 28,529)
Street centerlines	0.20	0.87	0.51	0.30	0.53
TIGER 2000	4.20	2.28	3.06	2.23	2.55
High schools	(*n* = 1395)	(*n* = 8501)	(*n* = 7997)	(*n* = 13,781)	(*n* = 31,674)
Street centerlines	0.07	0.65	0.44	0.12	0.34
TIGER 2000	1.15	1.31	1.80	0.77	1.19

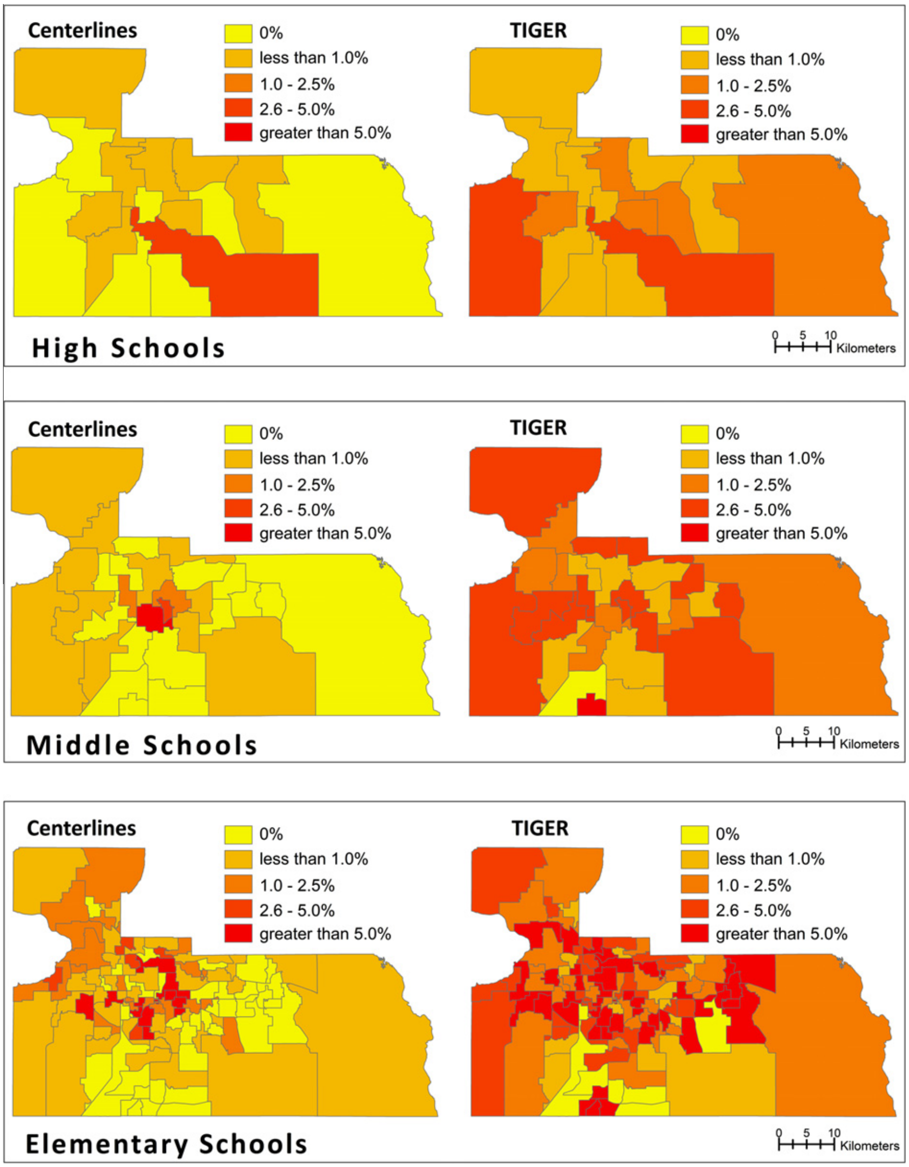

Fig. 4. Percent incorrectly allocated students by school zone for two street geocoding techniques.

allocated to the correct school zone boundary using the reference locations as comparison.

Several conclusions can be drawn from the results. First, there is a substantial difference between the two geocoding methods. Errors for the street centerlines are consistently lower compared to TIGER 2000 by a factor of at least two in most comparisons and often greater. The positional accuracy for local street centerlines is known to be much better compared to TIGER 2000 data (e.g., Zandbergen, 2011) which directly translates into lower errors in the aggregation. Second, the errors are inversely proportional to the size of the polygons. The average sizes for the school zones are: 4.4 km^2 for elementary schools ($n = 111$), 16.3 km^2 for middle schools ($n = 30$), and 30.5 km^2 for high schools ($n = 16$). The percentage errors for all students combined using the results for TIGER 2000 for illustration are: 5.51% for elementary schools, 2.55% for middle schools and 1.19% for high schools. Third, the errors vary by racial/ethnicity category. On average for a given school type and geocoding technique, minority students are more likely than white students to be incorrectly allocated. For example, for middle schools using street centerlines as an illustration, the error is 0.74% for white students, 1.56% for Hispanic students, and 2.92% for black students. These differences are consistent across all comparisons for Hispanic and black students, but not for Asians. The differences between racial/ethnic categories can be explained by examining the size of polygons relative to patterns of where minority students live. Fig. 4 shows the percentage for each school zone boundary for the three school types and the two geocoding techniques. Larger errors are more common for smaller polygons, especially for middle schools, which is consistent with the comparison of the three different sets of school zone boundaries themselves. Smaller polygons are located primarily in the higher density urban core of Orlando, which are predominantly non-white. The larger errors for non-white students, therefore, are a direct result of the racial/ethnic differences across gradients of urban densities.

6. Conclusions

Street geocoding is currently the most widely employed technique in the US to assign XY coordinates to individual addresses. Errors introduced by street geocoding include incompleteness, positional error and incorrect assignment to geographic units. The focus of the current paper is on the effects of positional error. A review of empirical studies suggests that positional errors of street geocoding are neither small nor random in nature, and that substantial bias may be introduced in spatial analysis that employs the results of geocoding. The positional errors of street geocoding are somewhat unique relative to those of other types of spatial data: (1) the magnitude of error varies strongly across urban–rural gradients; (2) the direction of error is not uniform, but strongly associated with the properties of local street segments; (3) the distribution of errors does not follow a normal distribution, but is highly skewed and characterized by a substantial number of very large error values; and (4) the magnitude of error is spatially autocorrelated and is related to properties of the reference data. This makes it difficult to employ analytic approaches or Monte Carlo simulations for error propagation modeling because these rely on generalized statistical characteristics.

The current paper has described an alternative empirical approach to error propagation modeling for geocoded data. The error propagation model relies on the utilization of a high quality address point dataset to create reference locations against which the results of alternative street geocoding techniques can be compared. The model has a number of strengths: it makes no assumptions about the nature of the positional error in geocoding; it does not require complex geostatistical characterizations to determine model parameters; it can be applied to any spatial-analytical technique that employs a geocoded set of points; and it does not require resource intensive methods to validate the accuracy of geocoded locations. The model also has a number of limitations: it requires high quality address point data for at least a portion of the study area; it only uses locations present in address points, resulting in a somewhat conservative error model; it does not address issues of geocoding completeness; and it does not incorporate any errors in the other inputs of spatial analysis. The implementation of the approach is illustrated using three different case-studies of geocoded individual-level datasets.

The first case-study determined the error in land cover associations of geocoded addresses using a point-in-raster overlay. Results indicate relatively poor accuracy of these land cover associations, in particular for areas with greater land cover heterogeneity. Reducing the number of land cover categories improved accuracy. Addresses in rural areas were consistently over-estimated as being associated with urban land cover types due to the particular nature of how road networks are classified in the specific land cover dataset. This observed bias presents a serious challenge to the use of geocoded locations in combination with categorical remote sensing datasets.

The second case-study determined the effects of positional error on local clustering using kernel density analysis. Errors in the cluster delineation were found to be very sensitive to the kernel density bandwidth. Using a bandwidth of 100 meters resulted in very large errors but these gradually decrease for larger bandwidths. The effect of using different street geocoding techniques of varying quality was very minor by comparison.

The third case determined the effects of positional error on spatial data aggregation using the residential addresses of public school children and school zone boundaries. Errors in boundary allocation were found to be dependent of polygon size, with smaller polygons resulting in larger errors. Errors were also found to be sensitive to the quality of the street geocoding, with results for local street centerlines resulting in substantially lower errors compared to TIGER data. Minority students (including blacks and Hispanics) were more likely to be incorrectly allocated relative to white students. This was attributed to the smaller size of the school zones in the urban core where black and Hispanic students make up the majority of the student population.

In general, results from the case-studies demonstrated that the unique nature of the positional error of street geocoding introduces substantial noise in the result of spatial analysis, including a substantial amount of bias across urban–rural and socioeconomic gradients for some analysis scenarios. This confirms findings from earlier studies, but expands these to a wider range of analytical techniques. Findings also suggest that the effect of geocoding quality on spatial analysis depends strongly on the specific nature of the analysis technique. The "typical" positional error of street geocoding can have a strong effect on spatial analysis at very short distances of several hundred meters. As with any spatial-analytical technique, a good understanding of both the input data quality and knowledge of the robustness of the technique itself are required to determine the reliability of the analysis result. The proposed framework presents a formalized methodology to further the understanding of the effects of geocoding quality.

References

Arbia G, Griffith D, Haining R. Error propagation modeling in raster GIS: overlay operations. Int J Geogr Inf Sci 1998;12:145–67.

Bilcher G, Balchak S. Address matching bias: ignorance is not bliss. Policing 2007;30:32–60.

Burra T, Jerrett M, Burnett RT, Anderson M. Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. Can Geogr 2002;42:160–71.

Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2003;2.

Eck JE, Chainey S, Cameron JG, Leitner M, Wilson RE. Mapping crime: understanding hotspots. National Institute of Justice Special Report; 2005.

Eclschlaeger CR. The stochastic simulation approach: tools for representing spatial application uncertainty. Santa Barbara: University of California; 1998 [Ph.D. dissertation].

Estes M, Crosson W, Al-Hamdan M, Estes S, Quattrochi D, Kent S, et al. Use of remotely-sensed data to evaluate the relationship between living environment and blood pressure. Environ Health Perspect 2009;117:1832–8.

Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: the current state of geocoding. URISA J 2007;19:33–46.

Griffith DA, Millones M, Vincent M, Johnson DL, Hunt A. Impacts of positional error on spatial regression analysis: a case-study of address locations in Syracuse, New York. Trans GIS 2007;11:655–79.

Harada Y, Shimada T. Examining the impact of the precision of address geocoding on estimated density of crime locations. Comput Geosci 2006;32:1096–107.

Hart TC, Zandbergen PA. Effects of geocoding quality on predictive hotspot mapping. American Society of Criminology Annual Meeting, November 16-19, Washington, DC; 2011.

Hengl T, Heuvelink GBM, van Loon EE. On the uncertainty of stream networks derived from elevation data: the error propagation approach. Hydrol Earth Syst Sci Discuss 2010;7:767–99.

Heuvelink GBM. Error propagation in environmental modeling with GIS. London UK: Taylor and Francis; 1998.

Heuvelink GBM. Analyzing uncertainty propagation in GIS: why is it not so simple? In: Moody GM, Atkinson PM, editors. Uncertainty in remote sensing and GIS. West Sussex, UK: Wiley & Sons Inc.; 2002. p. 156–65.

Homer C, Dewitz J, Fry J, Coan M, Hossain N, Larson C, et al. Completion of the 2001 National Land Cover Database for the Conterminous United States. Photogramm Eng Remote Sensing 2007;73(4):337–41.

Jacquez GM, Rommel R. Local indicators of geocoding accuracy (LIGA): theory and application. Int J Health Geogr 2009;8.

Karssenberg D, De Jong K. Dynamic environmental modelling in GIS: 2. Modelling error propagation. Int J Geogr Inf Sci 2005;19:623–37.

Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. Health Place 2007;13:293–8.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91:1114–6.

Lanter DP, Veregin H. A research paradigm for propagating error in layer-based GIS. Photogramm Eng Rem S 1992;58:825–33.

Lindsay JB. Sensitivity of channel mapping techniques to uncertainty in digital elevation data. Int J Geogr Inf Sci 2006;20:669–92.

Lindsay JB, Evans MG. The influence of elevation error on the morphometrics of channel networks extracted from DEMs and the implications for hydrological modeling. Hydrol Process 2008;22:1588–603.

Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr 2008;7.

Oksanan J, Sarjakowski T. Error propagation analysis of DEM-based drainage basin delineation. Int J Remote Sens 2005;26:3085–102.

Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. Int J Geogr Inf Sci 2001;15:473–85.

Rushton G, Armstrong MP, Gittler J, Greene B, Pavlik CE, West MW, et al. Geocoding in cancer research: a review. Am J Prev Med 2006;30:S16–24.

Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, et al. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol 2007;17:464–70.

Stanislawski LV, Dewitt BA, Shrestha RL. Estimating positional accuracy of data layers within a GIS through error propagation. Photogramm Eng Rem S 1996;62:429–33.

Strickland MJ, Siffel C, Gardner BR, Beerzen AK, Correa A. Quantifying geocode location error using GIS methods. Environ Health 2007;6.

Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005;16:542–7.

Whitsel EA, Quilbrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, et al. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innov 2006;3.

Wu J, Funk T, Lurman FW, Winer AM. Improving spatial accuracy of roadway networks and geocoded addresses. Trans GIS 2005;9:585–601.

Zandbergen PA. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health 2007;7:37.

Zandbergen PA, Green JW. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environ Health Persp 2007;115:1363–70.

Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban 2008a;32:214–32.

Zandbergen PA. Positional accuracy of spatial data: non-normal distributions and a critique of the National Standard for Spatial Data Accuracy. Trans GIS 2008b;12:103–30.

Zandbergen PA. Geocoding quality and implications for spatial analysis. Geography Compass 2009;3:647–80.

Zandbergen PA. Accuracy considerations in the analysis of depressions in medium resolution lidar DEMs. Gisci Remote Sens 2010;47:187–207.

Zandbergen PA. Influence of street reference data on geocoding quality. Geocarto Int 2011;26(1):35–47.

Zandbergen PA, Hart TC. Geocoding accuracy considerations in determining residency restrictions for sex offenders. Crim Justice Policy Rev 2009;20:6–19.

Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol 2006;16:842–9.

Zimmerman DL, Fang X, Maxumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. Int J Health Geogr 2007;6.

Zimmerman DL, Li J, Fang X. Spatial autocorrelation among automatedgeocoding errors and its effects on testing for disease clustering. Stat Med 2010;29:1025–36.

Zimmerman DL, Li J. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. Int J Health Geogr 2010;9.