An analysis of the process and results of manual geocode correction

Yolanda J. McDonald,[1] Michael Schwind,[2] Daniel W. Goldberg,[1] Amanda Lampley,[1] Cosette M. Wheeler[3]

[1]Department of Geography, College of Geosciences, Texas A&M University, College Station, TX; [2]College of Science & Engineering, Texas A&M University Corpus Christi, Corpus Christi, TX; [3]School of Medicine, University of New Mexico, Albuquerque, NM, USA

Abstract

Geocoding is the science and process of assigning geographical coordinates (*i.e.* latitude, longitude) to a postal address. The quality of the geocode can vary dramatically depending on several variables, including incorrect input address data, missing address components, and spelling mistakes. A dataset with a considerable number of geocoding inaccuracies can potentially result in an imprecise analysis and invalid conclusions. There has been little quantitative analysis of the amount of effort (*i.e.* time) to perform geocoding correction, and how such correction could improve geocode quality type. This study used a low-cost and easy to implement method to improve geocode quality type of an input database (*i.e.* addresses to be matched) through the processes of manual geocode intervention, and it assessed the amount of effort to manually correct inaccurate geocodes, reported the resulting match rate improvement between the original and the corrected geocodes, and documented the corresponding spatial shift by geocode quality type resulting from the corrections. Findings demonstrated that manual intervention of geocoding resulted in a 90% improvement of geocode quality type, took 42 hours to process, and the spatial shift ranged from 0.02 to 151,368 m. This study provides evidence to inform research teams considering the application of manual geocoding intervention that it is a low-cost and relatively easy process to execute.

Correspondence: Yolanda J. McDonald, Department of Geography, College of Geosciences, Texas A&M University, 3147 College Station, TX, USA.
Tel: +1.915.615.9088 - Fax: +1.979.863.4487.
E-mail: ymcdonald77@tamu.edu

Key words: Manual geocode correction; Geocode inaccuracies; Match rate improvement; Geocode.

Contributions: YJM research design, data processing, data analysis, tables, manuscript writing and reviewing, and references search; MS research design, data processing, tables, figures, manuscript writing, and references search; DWG research design and manuscript review; AL manuscript writing and references search; CMW data collection and manuscript review.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: this effort was supported by U54CA164336 (to CM Wheeler) from the US National Cancer Institute funded Population-Based Research Optimizing Screening through Personalized Regimens (PROSPR) consortium. The overall aim of PROSPR is to conduct multi-site, coordinated, transdisciplinary research to evaluate and improve cancer screening processes. Yolanda J. McDonald was funded by the UNM Cancer Center Support Grant P30CA118100 and Texas A&M College of Geosciences.

Ethical statement: ethical approval for the study was approved by the University of New Mexico Human Research Review Committee and by the Texas A&M University Human Subjects Protection Program Institutional Review Board (ID: IRB#2014-0078D).

Received for publication: 26 October 2016.
Revision received: 21 February 2017.
Accepted for publication: 1 March 2017.

Geospatial Health 2017; 12:526
doi:10.4081/gh.2017.526

Introduction

Geocoding is the process of matching postal addresses to their corresponding geographical coordinates (*i.e.* latitude, longitude) (Rushton *et al*., 2006). Sophisticated science, data sets, and algorithms underlie this complex process (Boscoe, 2008; Zandbergen, 2008). There are a large number of published studies (Goldberg, 2008; Ratcliffe, 2001) that describe the numerous algorithms that are used during the geocoding process to attempt to match an input address to an address stored in a reference database. The variability in algorithms, addresses, and databases can lead to a variety of errors in the geocoded results (Ratcliffe, 2001; Gilboa *et al*., 2006; Schootman *et al*., 2007; Zandbergen, 2008, 2011; Goldberg *et al*., 2013). There is no such thing as a *one size fits all* type of geocoding system that works perfectly in every situation and for every user. The accuracy of this complex process can range from the centroid of a rooftop to the centroid of a state (Jacquez and Rommel, 2009). This leads to the following questions: Should inaccuracies be incorporated into research or should they be omitted entirely? Should inaccuracies be corrected? Is there a threshold that inaccuracies should not exceed?

Previous studies have indicated that researchers should attempt to correct inaccurate data so that real world variances can be incorporated into analysis (Krieger, 2003; Zandbergen, 2007; Goldberg *et al*., 2008; Goldberg and Cockburn, 2012; Murray *et al*., 2011; Zandbergen, 2012). The practical application of reducing geocode inaccuracies is to improve the source data (*i.e.*

geocoded data) used for spatial analysis (Strickland *et al*., 2007). However, despite calls to pay heed to geocode quality by type and to employ manual geocode correction methods, there are few documented case studies that evaluate the cost effectiveness of this practice, or the improvements that can be expected by undertaking such an effort (Goldberg *et al*., 2008). The purpose of this study was to quantify the effort (*i.e*. time) required to manually correct the geocodes in a health related dataset, as well as the match rate improvement between the original geocoded and the corrected geocode, and the corresponding spatial shift by geocode quality type resulting from the corrections. The results of this study can be used to help guide researchers as they decide whether or not to undertake manual geocoding correction to improve the geocode quality type of a dataset.

Materials and Methods

Web based geocoding and interactive geocoding correction procedures were performed using the Texas A&M University (TAMU) Geoservices Online Geocoding service, version 4.01, which was developed by the study authors (Goldberg *et al*., 2008). The corrections were performed by the study authors, a Ph.D. student and an honors undergraduate student. This web-based system allows for rapid manual intervention of previously geocoded data by drawing from online satellite imagery, street maps, and additional geocoding engines to determine an improved geocode for each record (Goldberg *et al*., 2008).

This system allows a user to upload a dataset and analyse each record one at a time. It compares the current location of each geocode to that of another location provided by an alternate geocoder (*i.e*. Google Maps) within the TAMU online geocoding platform, and allows the user the flexibility to execute a manual intervention process to determine a more accurate geocode. The user can select which geocoder produced a more accurate location and the dataset can be updated with the corrected coordinates. In the event that neither geocoder provides an accurate location, the user can utilise online sources to refine an address (*e.g*. misspelling of an address) as well as aerial imagery and street views to attempt to find the location intuitively, and visually verify a location using Google Maps. The TAMU Geoservices Online Geocoding service utilises publicly accessible data so person-hours are the only cost associated with the geocode correction processes. It is free to all researchers (https://geoservices.tamu.edu/), and the source code can be made available upon request to researchers and/or organisations that wish to use it.

To analyse the impact of the geocode correction process, a health related dataset was used. This dataset contained 784 addresses of health service facilities located within the state of New Mexico that offered cervical screening (Pap and/or Human Papillomavirus testing), diagnostic testing (colposcopy), and excisional pre-cancer treatment (loop electrosurgical excision procedure or cone biopsy). Although this data is publically available, it is not practical to obtain information on specific tests offered by individual clinics or providers. This unique health service facilities dataset was provided by the New Mexico HPV Pap Registry

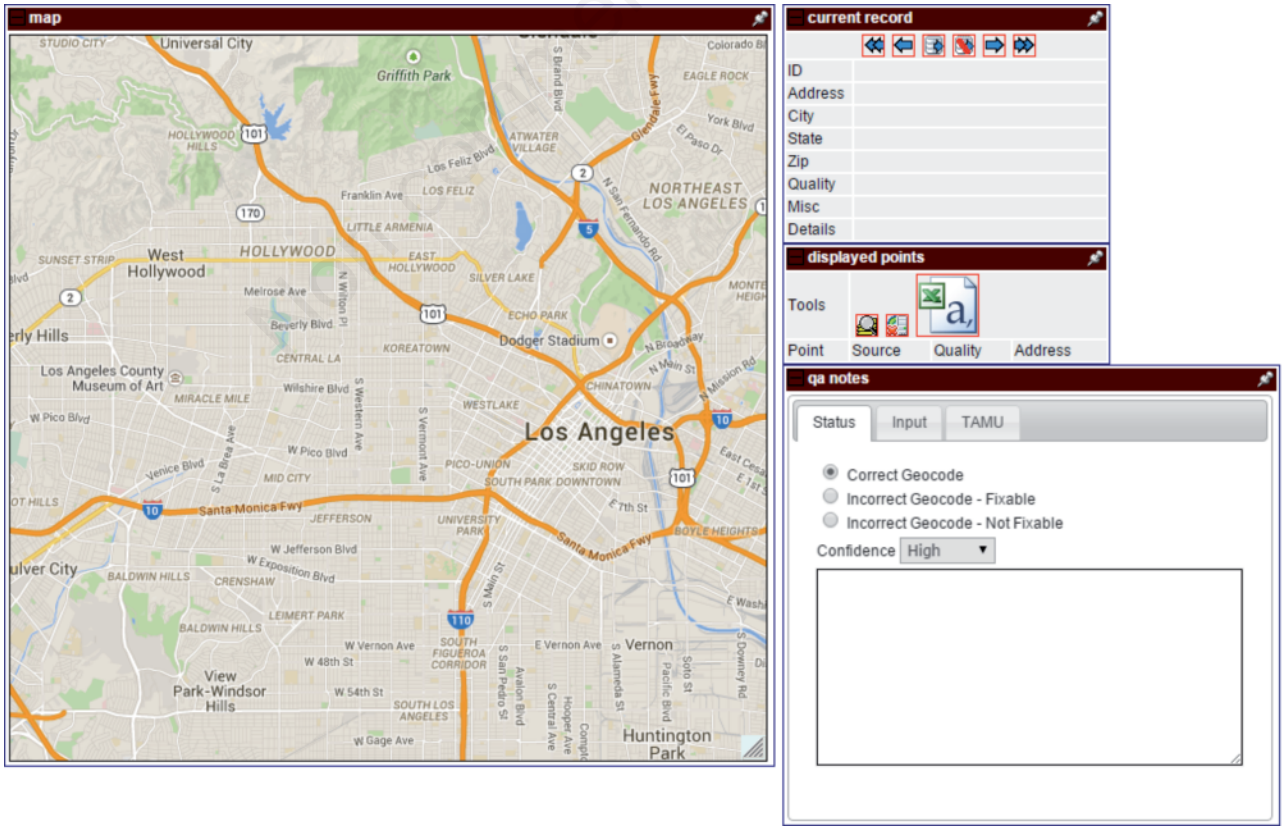

Figure 1. Manual geocode correction tool interface.

(NMHPVPR). The NMHPVPR is the first population-based statewide cervical screening registry in the United States; it includes address-level data on healthcare facilities providing aforementioned services in rural and urban areas. Due to the uniqueness of this data set, the authors invested the effort to have the most accurate geocoding possible.

The first step of processing was to geocode the entire set of addresses using the TAMU Geoservices Online Geocoding service. The version of the geocoding service used for this research included the 2015 Navteq Address Points database, the 2010 USPS ZIP+4 reference files, the 2010 Boundary Solutions National Parcel Data Layer, and the 2010 US Census TIGER/Lines the reference, and the US Census Bureau 2010 Cartographic Boundary files for Minor Civil Divisions, Zip Code Tabulation Areas, Counties, and States. Once the results were obtained, the geocoded file was uploaded to the TAMU Geoservices Online Geocoding Correction Service; Figure 1 displays the geocode correction tool interface. This service provides a user interface that displays a map that shows the point obtained from the TAMU geocoding system and the point obtained from the alternate geocoder, *i.e.* Google Maps. If the alternate geocoder is able to find a match that is more accurate than the original match, a button can be pressed that updates the original geocode with the more accurate geocode. As previously noted, in the case that both geocodes appear to be inaccurate, the next step would be to attempt manual interactive geocoding. Online resources can be used to refine the address contained within the input file and often photo(s) of the building to be geocoded are available online. In addition, the user can study aerial imagery and street views of the location and attempt to manually locate the site; Figure 2 displays the correction prompt. If the site is located, the user marks that spot on the map and the geocode will be updated. These processes were used to update and correct the health service facility dataset analysed for this study. The final file contained information about the original geocodes and the corrected geocodes, which were used for comparative analysis.

Results

This section provides a description of the results that were obtained from manually correcting the 784 geocodes. The same method used in prior research (Goldberg *et al*., 2008) was used to classify an improved record as one of two criteria (Rushton *et al*., 2006). A record that was originally non-geocodable and a geocode was obtained after processing was categorised as criteria one. A record that was previously geocodable and the accuracy of the geocode was improved after processing was categorised as criteria two (Boscoe, 2008). It should be noted that we considered a record that has a lower North American Association of Central Cancer Registries (NAACCR) GIS Coordinate Quality Code (Goldberg, 2008) after it has been processed, to be an improvement in accuracy according to criteria 2. We acknowledge that without direct field observation, it is not possible to assess with 100% accuracy that the original geocode was improved. All of the records in the dataset were geocodeable in the original file, therefore no records met criteria one. For measuring improvement, we followed the geocode output type hierarchy of the NAACCR GIS Coordinate Quality Code.

Of the 784 records, 709 met criteria two. Ninety percent of the original addresses were corrected to a higher accuracy after the manual correction processes and 10% did not change. Of the 75 records that did not change, 21 were of the Exact Parcel Centroid quality, 50 were of Address Range Interpolation, and four records were of the USPS Zip Centroid quality. Table 1 shows that of the 71 addresses that matched to either Exact Parcel Centroid or Address Range Interpolation these records were already either the second or the third highest ranked geocode quality types (Goldberg, 2008).

Table 1 contains the original and corrected geocode quality type for the dataset. The original dataset contained zero records that were geocoded to the Building Centroid quality type. The corrected dataset contains 638 (81.38%) geocodes of this quality. It is notable that the original geocoded dataset contained 204 (26%) geocodes that matched to the USPS Zip Centroid quality type and after manual geocoding correction there were only four (<1%) records.

Discussion

Processing time

The correction process of the entire dataset consisting of 784 records was completed in 42.21 hours. The average processing time was 194 seconds per record. In the following sections, we will discuss the quality improvement of the dataset. The purpose of analysing both the time taken and the geocode quality improvement is to illustrate the effort that is involved versus the improvement in geocode accuracy gained.

Spatial shift

Of the 784 geocodes, 709 were assigned a new set of coordinates during the correction process. In this section we will review the spatial shift that the majority of the geocodes underwent. This

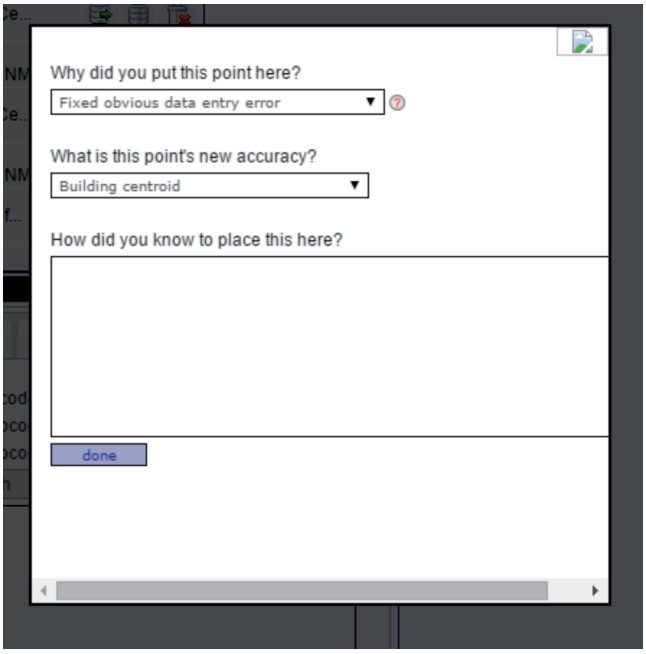

Figure 2. Prompt for new accuracy description.

distance was measured in meters (m) using the XY to Line tool within ArcGIS 10.1. Of the addresses that met criteria 2, the spatial shift improvements ranged from the smallest (0.018851 m) to the largest (151,368 m), the mean was 1963 m, and the median was 114 m (Table 2). For the smallest spatial shift improvement category, *i.e.* Exact Parcel Centroid to Building Centroid, we found that these geocode quality types were closely aligned and required minimal processing time (in seconds), mean 100 seconds and the median 52. In the event that the original geocode location of an Exact Parcel Centroid quality type was already accurate but needed to be updated to Building Centroid, the building was selected to reflect its true level of accuracy. The newly selected point was located proximate to the original point, resulting in the small difference between the original and corrected geocodes. For the largest spatial shift the geocode quality improved from USPS Zip Centroid to Street Centroid and the processing time was 1276 sec (21.2 min). Figure 3 illustrates an example of the spatial shift between the original and corrected geocoded points. In the bottom left of the diagram, it can be seen that many corrected geocoded points were derived from the same original point. In this case, many addresses were originally geocoded to a zip code centroid and then corrected to more accurate single location-based geocode.

Geocoding a list of addresses is often just the first step to a more extensive project (Rushton *et al*., 2006; Goldberg *et al*., 2007). This first step, however, is very important because it can ultimately dictate the accuracy and direction of the final result (Oliver *et al*., 2005; Zandbergen, 2009; Wey *et al*., 2009). Prior research has demonstrated that geocoded datasets should be evaluated not only for match rate but also by geocode quality type (Goldberg *et al*., 2008; Rushton *et al*., 2006). Based on the level of accuracy of geocodes and the research purpose, it is our recommendation that researchers pause and evaluate if it is necessary to invest time to improve the accuracy of the geocodes (Krieger *et al*., 2001; Bonner *et al*., 2003; Nuckols *et al*., 2004; Oliver *et al*., 2005; Grubesic and Matisziw, 2006; Schootman *et al*., 2007; Zandbergen, 2007, 2009). This study illustrates that a dataset of lower geocode quality types can be improved to a higher level of quality with very little investment of time, effort, or finances. The original dataset contained zero geocodes that matched to a building centroid. After 42 hours (~one week of work), 638 (81%) of the geocodes matched to a building centroid. Our spatial shift findings support previous studies demonstrating that inaccurate geocoding produces positional errors (Cayo and Talbot, 2003; Ward *et al*., 2005). These errors have the potential to impact health analysis ranging from inaccurate local disease rates to imprecise accessibility measures; these health analysis studies are frequently used to inform health policy decisions (Jacquez, 2012). The manual intervention geocoded dataset that was produced as part of this study is now more suitable to be used for analysis because it will yield more reliable results.

Table 1. Geocode quality types and descriptions ranked from most to least accurate and geocode quality types of the original and corrected dataset.

Quality type	Description	Original quality type Total (N=784)	%	Corrected quality type Total (N=784)	%
Building centroid	Matched to the centroid of the building	0	0.00	638	81.38
Exact parcel centroid point	Matched to the centroid of the parcel	194	24.75	44	5.61
Address range interpolation	Uses information about the address number ranges to estimate the position of a numbered address	386	49.23	79	10.08
Street centroid	Matched to the centroid of the street	0	0.00	18	2.29
USPS zip centroid	Matched to the zip code area centroid	204	26.02	4	0.51
City centroid	Matched to the centroid of the city	0	0.00	1	0.13
State centroid	Matched to the centroid of the state	0	0.00	0	0.00

USPS, United States Postal Service.

Table 2. Geocode quality types of the original and corrected dataset and spatial shift improvement by each geocode quality type correction.

Old geocode quality type	New geocode quality type*	Total (N=703) N	%	Spatial shift (m) Mean	Median	IQR (Q1, Q3)°	Minimum	Maximum
Address range interpolation	Building centroid	323	45.95	355.22	105.88	(54.21, 221.96)	3.49	33936.56
Address range interpolation	Exact parcel centroid	10	1.42	253.77	72.32	(42.75, 130.22)	7.04	1904.97
Exact parcel centroid	Building centroid	171	24.32	116.62	11.66	(2.29, 27.25)	0.02	8260.35
USPS zip centroid	Building centroid	143	20.34	5070.82	3094.47	(1446.09, 5455.60)	191.04	54717.53
USPS zip centroid	Exact parcel centroid	14	1.99	9903.80	5669.26	(3036.69, 11614.65)	871.14	41691.95
USPS zip centroid	Address range interpolation	29	4.13	6581.60	3405.08	(858.99, 12227.95)	114.31	23920.18
USPS zip centroid	Street centroid	13	1.85	22956.72	11708.03	(3959.76, 20884.24)	1734.06	151367.94
All corrections		703		1963.18	113.81	(24.64, 940.39)	0.02	151367.94

USPS, United States Postal Service. *Geocode quality type change of N≥5; °IQR, interquartile range.

Figure 3. Spatial shift from original geocode to corrected geocode.

Conclusions

The current study provides additional motivation and evidence-based findings for the purpose of demonstrating that manual geocoding correction is both a feasible and economical method for improving the quality of geocoded data. And, we demonstrated that the manual intervention geocoded processes resulted in increased match rates, higher confidence in geocode quality, and improved geocode match types. Finally, this study supports prior research that has been conducted in the geocoding accuracy and analysis field, and supports that prior findings are transferable from one geographic region to another as well as across domains of health services (Goldberg *et al*., 2008). As demonstrated by this study, the TAMU Geoservices geocoder and the geocode correction tool, which is integrated in the online web service, is a low to no cost, easy to use option to improve geocode accuracy.

References

Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL, 2003. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 14:408-11.

Boscoe FP, 2008. The science and art of geocoding. In: Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL, eds. Geocoding health data: the use of geographic codes in cancer prevention and control, research, and practice. CRC Press, Boca Raton, FL, USA, pp. 95-109.

Cayo M, Talbot T, 2003. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2:10.

Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Langlois PH, Savitz DA, Herring AH, 2006. Comparison of residential geocoding methods in population-based study of air quality and birth defects. Environ Res 101:256-62.

Goldberg D, 2008. A geocoding best practices guide. North American Association of Central Cancer Registries, Springfield, IL, USA.

Goldberg D, Ballard M, Boyd J, Mullan N, Garfield C, Rosman D, Ferrante AM, Semmens JB, 2013. An evaluation framework for comparing geocoding systems. Int J Health Geogr 12:50.

Goldberg D, Cockburn M, 2012. The effect of administrative boundaries and geocoding error on cancer rates in California. Spatial Spatio-Temporal Epidemiol 3:39-54.

Goldberg D, Wilson J, Knoblock C, 2007. From text to geographic coordinates: the current state of geocoding. Urisa J 19:33-47.

Goldberg D, Wilson J, Knoblock C, Ritz B, Cockburn M, 2008. An effective and efficient approach for manually improving geocoded data. Int J Health Geogr 7:60.

Grubesic T, Matisziw T, 2006. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiologic data. Int J Health Geogr 5:1.

Jacquez GM, 2012. A research agenda: Does geocoding positional error matter in health GIS studies? Spatial Spatio-temporal Epidemiolo 3:7-16.

Jacquez GM, Rommel R, 2009. Local indicators of geocoding accuracy (LIGA): theory and application. Int J Health Geogr 8:60.

Krieger N, 2003. Place, space, and health: GIS and epidemiology. Epidemiology 14:384-5.

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

Murray AT, Grubesic TH, Wei R, Mack EA, 2011. A hybrid geocoding methodology for spatio-temporal data. Trans GIS 15:795-809.

Nuckols JR, Ward MH, Jarup L, 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Persp 112:1007-15.

Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW,

2005. Geographic bias related to geocoding in epidemiologic studies. Int J Health Geogr 4:29.

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

Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL, 2006. Geocoding in cancer research: a review. Am J Prev Med 30:16-24.

Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G, 2007. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Ann Epidemiol 17:379-87.

Strickland MJ, Siffel C, Gardner BR, Berzen, AK, Correra A, 2007. Quantifying geocode location error using GIS methods. Environ Health 6:10.

Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P, 2005. Positional accuracy of two methods of geocoding. Epidemiology 16:542-7.

Wey CL, Griesse J, Knightlinger L, Wimberly MC, 2009. Geographic variability in geocoding success for West Nile virus cases in South Dakota. Health Place 15:1108-14.

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

Zandbergen PA, 2008. A comparison of address point, parcel and street geocoding techniques. Comp Environ Urban Syst 32:214-32.

Zandbergen PA, 2009. Geocoding quality and implications for spatial analysis. Geogr Compass 3:647-80.

Zandbergen PA, 2011. Influence of street reference data on geocoding quality. Geocarto Int 26:35-47.

Zandbergen PA, 2012. Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. Spatial Spatio-Temporal Epidemiol 3:69-82.