Geospatial Health 5(2), 2011, pp. 265-273

Evaluation of the positional difference between two common geocoding methods

Dustin T. Duncan[1,2], Marcia C. Castro[3], Jeffrey C. Blossom[4], Gary G. Bennett[1,5], Steven L. Gortmaker[1,2]

[1]*Department of Society, Human Development and Health, Harvard School of Public Health, Boston, MA 02115, USA;* [2]*Harvard Prevention Research Center on Nutrition and Physical Activity, Harvard School of Public Health, Boston, MA 02215, USA;* [3]*Department of Global Health and Population, Harvard School of Public Health, Boston, MA 02115, USA;* [4]*Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA;* [5]*Department of Psychology and Neuroscience and Duke Global Health Institute, Duke University, Durham, NC 27708, USA*

Abstract. Geocoding, the process of matching addresses to geographic coordinates, is a necessary first step when using geographical information systems (GIS) technology. However, different geocoding methodologies can result in different geographic coordinates. The objective of this study was to compare the positional (i.e. longitude/latitude) difference between two common geocoding methods, i.e. ArcGIS (Environmental System Research Institute, Redlands, CA, USA) and Batchgeo (freely available online at http://www.batchgeo.com). Address data came from the YMCA-Harvard After School Food and Fitness Project, an obesity prevention intervention involving children aged 5-11 years and their families participating in YMCA-administered, after-school programmes located in four geographically diverse metropolitan areas in the USA. Our analyses include baseline addresses (n = 748) collected from the parents of the children in the after school sites. Addresses were first geocoded to the street level and assigned longitude and latitude coordinates with ArcGIS, version 9.3, then the same addresses were geocoded with Batchgeo. For this analysis, the ArcGIS minimum match score was 80. The resulting geocodes were projected into state plane coordinates, and the difference in longitude and latitude coordinates were calculated in meters between the two methods for all data points in each of the four metropolitan areas. We also quantified the descriptions of the geocoding accuracy provided by Batchgeo with the match scores from ArcGIS. We found a 94% match rate (n = 705), 2% (n = 18) were tied and 3% (n = 25) were unmatched using ArcGIS. Forty-eight addresses (6.4%) were not matched in ArcGIS with a match score ≥80 (therefore only 700 addresses were included in our positional difference analysis). Six hundred thirteen (87.6%) of these addresses had a match score of 100. Batchgeo yielded a 100% match rate for the addresses that ArcGIS geocoded. The median for longitude and latitude coordinates for all the data was just over 25 m. Overall, the range for longitude was 0.04-12,911.8 m, and the range for latitude was 0.02-37,766.6 m. Comparisons show minimal differences in the median and minimum values, while there were slightly larger differences in the maximum values. The majority (>75%) of the geographic differences were within 50 m of each other; mostly <25 m from each other (about 49%). Only about 4% overall were ≥400 m apart. We also found geographic differences in the proportion of addresses that fell within certain meter ranges. The match-score range associated with the Batchgeo accuracy level “approximate” (least accurate) was 84-100 (mean = 92), while the “rooftop” Batchgeo accuracy level (most accurate) delivered a mean of 98.9 but the range was the same. Although future research should compare the positional difference of Batchgeo to criterion measures of longitude/latitude (e.g. with global positioning system measurement), this study suggests that Batchgeo is a good, free-of-charge option to geocode addresses.

Keywords: geocoding, positional difference, ArcGIS, Batchgeo, addresses, USA.

Introduction

Geographical information systems (GIS) data are increasingly being used to investigate geospatial aspects of health (including health behaviour), as well as to conduct disease surveillance and to map disease clusters (Moore and Carpenter, 1999; Elliot et al., 2000; Cromley and McLafferty, 2002; Rushton, 2003). Geocoding, the process of matching street addresses to geographic coordinates (latitude and longitude), is a necessary first step when utilising point data available with addresses. However, geocoding methodology varies substantially across studies and can be fraught with problems, including providing inaccurate geographic coordinates (Drummond, 1995; Cromley and McLafferty, 2002; Rushton et al., 2006).

Corresponding author:
Dustin T. Duncan
Department of Society, Human Development and Health
Harvard School of Public Health
677 Huntington Avenue, Kresge Building, 7th Floor
Boston, MA 02115, USA
Tel. +1 617 384 8732; Fax +1 617 384 8730
E-mail: dduncan@hsph.harvard.edu

Inaccurate location assignment leads to misclassification of environmental exposures, which can produce biased estimates and reduce the statistical power to detect true associations. This is known to be especially problematic for small area analysis (e.g. studies characterising neighbourhoods using finer spatial scales such as a 400 m buffer distance around participant residences) (Zimmerman and Sun, 2006; Whitsel et al., 2006; Zandbergen and Green, 2007; Mazumdar et al., 2008). The use of smaller spatial scales has surged in recent years in geospatial health studies, which may be due to increased interest in GIS and spatial analyses of health and health behaviour particularly at the individual level (Moore and Carpenter, 1999; Elliot et al., 2000; Cromley and McLafferty, 2002; Rushton, 2003); perhaps also due to criticisms associated with the use of administrative boundaries (e.g. zip codes and census tracts in the USA) as neighbourhood definitions (Coulton et al., 2001; Krieger et al., 2002; Osypuk and Galea, 2007; Lee et al., 2008). Finally, the increased use of smaller scales has benefited from the improved availability of data and better computing capability to handle large datasets (Miller, 2009).

Of note, researchers and practitioners are increasingly geocoding data "in-house", likely because of the increased user-friendliness and accessibility of geocoding software. "In-house" geocoding may offer other important advantages, including improved technical transparency and facilitated input, possible cost savings and faster turnaround times (McElroy et al., 2003; Ward et al., 2005). Such rewards might be especially important in light of some evidence demonstrating that the accuracy of geocoding varies widely between different commercial geocoding firms (Krieger et al., 2001; Whitsel et al., 2004, 2006; Zandbergen and Green, 2007). One study found that the use of a commercial geocoding firm did not improve geocoding accuracy as compared to geocoding data "in-house" with ArcGIS software (Environmental System Research Institute; Redlands, CA, USA) and in several instances the geocodes provided by the commercial firm had more positional error than that of those provided by ArcGIS (Ward et al., 2005). Although a variety of tools can be used to geocode data "in-house", ArcGIS is widely recognised by the industry as the most commonly used commercial geocoding software. Batchgeo, a free service publicly available online (http://www.batchgeo.com/), can also be used to geocode addresses and is easy to use for individuals who have little experience with geocoding. Although Batchgeo has made significant enhancements in 2010, there are no studies to our knowledge examining the accuracy of its geographic coordinates. The objective of this study was to compare the positional (i.e. longitude and latitude) difference between two geocoding methods: ArcGIS, the largely used commercial application, and Batchgeo, a freely available tool.

Materials and methods

Address data

This study used address data collected as part of the YMCA-Harvard After School Food and Fitness Project, a multi-site, quasi-experimental, after-school obesity prevention intervention targeting children aged 5-11 years and their families. The intervention was delivered to after-school programmes, administered by YMCA[1] and located in four geographically diverse metropolitan areas in the USA. For anonimity, we discuss the metropolitan areas by general geographic region only: the Pacific Northwest (n = 180), the Midwest (n = 170), the South (n = 238) and the East (n = 166). Baseline data were collected in the fall of 2006, and the follow-up was in the spring of 2007. Since our focus was on address geocoding, the analyses for the present study include full baseline addresses collected from the parents of children in the after school sites (n = 754), not just those who actually participated in various intervention activities. It is also important to note that because the data were not designed specifically for geocoding, or for future spatial analyses, postal addresses rather than physical street addresses were collected. Each address element was later entered into a separate field of a Microsoft Excel file.

Address cleaning

All postal addresses (i.e. street, city, state, zip code) were preprocessed before geocoding to improve standardisation and quality. First, we removed any address that had P.O. boxes (n = 6) (Hurley et al., 2003). We then reviewed the data for misspelled address information using Google Maps and remedied any incorrect home addresses (e.g. incorrect street names). In addition, we removed all extraneous characteristics and standardised the spelling to the United States

[1] The leading nonprofit organization in the USA for youth development, healthy living and social responsibility (http://www.ymca.net/)

Postal Service format (e.g. we changed "street" to "St", "avenue" to "Ave", and "circle" to "Cir") (United States Postal Service, 2000).

Address geocoding

Geocoding is based on a linear interpolation of an address within the address range for the street segment in a reference street file (Drummond, 1995; Rushton et al., 2006). Using two methods, we geocoded the 748 addresses that remained in our database after cleaning the data: the Pacific Northwest (n = 176), the Midwest (n = 170), the South (n = 236) and the East (n = 166). In the first step, addresses were geocoded to the street level and assigned longitude and latitude coordinates, using the Tele Atlas US street address locator via the ArcGIS Online World Geocoding service with ArcGIS, version 9.3. Addresses were matched using a minimum match score of 65, spelling sensitivity of 60, and side offset of 10 feet, i.e. the default settings of ArcGIS. We then conducted interactive rematching in ArcGIS, where addresses can be reviewed and corrected on a case-by-case basis as necessary, for addresses with a match score of ≥80 that had ties. The match score is the agreement between the address that is being geocoded and the address location in the geocoding engine, indicating the geocoding software's certainty about the geocode accuracy. For a perfect match, the match score (range: 0-100) is 100. For this analysis, the ArcGIS minimum match score required was 80; therefore we deemed addresses with a match score ≥80 as "high quality". This was based on a quality assurance assessment in which we randomly checked 5% of addresses, comparing addresses with a match rate <80 and ≥80 produced in ArcGIS to Google Maps. Results indicated that only addresses with a match rate ≥80 were positionally accurate. Subsequently, we geocoded the same addresses using Batchgeo following the point-and-click procedures outlined at the website. The latitude and longitude coordinates from Batchgeo were generated with Google's geocoding service, which is the default setting. All addresses were geocoded by the two methods in late June of 2010.

Data analysis

After having obtained the results from ArcGIS and Batchgeo, we recorded descriptive information on the gecoding methods (e.g. the match rate which is the percentage of addresses that were successfully geocoded) and then used the ArcGIS project command to transform the data into the appropriate North American Datum (NAD) 1983 state plane coordinate system for each of the four metropolitan areas. The geographic transformation World Geodetic System (WGS) 1984 to NAD 1983, option number 1, was used during the transformation. This enabled us to also calculate the positional difference in longitude (x, east-west) and latitude (y, north-south) directions in meters. Specifically, we calculated the longitudinal difference (ArcGIS longitude - Batchgeo longitude coordinates) and latitude difference (ArcGIS latitude - Batchgeo latitude coordinates) between the two methods based on the absolute value difference (e.g. -36 was transformed to 36). When the positional difference between the two methods was closer to zero, the geographic location differences were closer in geographic proximity. These analyses were conducted for the data overall and for each of the four metropolitan areas. For the purpose of visualisation, we created maps showing examples of the discrepancy distance of the geocoded addresses for the two methods. We used a random number generator to select four examples to show the positional discrepancy (one per metropolitan area). To protect confidentiality, no disclosure of the metropolitan areas or other specific geographic information is shown on the maps created. Finally, we quantified the descriptions of the geocoding accuracy provided by Batchgeo (i.e. "approximate" - least accurate; "geometric center" - center of the city or zip code; "range interpolated" - the traditional geocoding method; and "rooftop" - most accurate, based on the actual building at that address) with the match scores obtained from ArcGIS. Data analyses were conducted in SAS version 9.2 (SAS Institute Inc.; Cary, NC, USA).

Results

Of the 748 addresses in the analytic sample, we found a match rate of 94% (n = 705), 2% (n = 18) were tied and 3% (n = 25) were unmatched using ArcGIS. For all addresses with candidate ties that had match scores ≥80 (n = 9), we performed interactive rematching in ArcGIS, which resulted in one address change. Forty-eight addresses (6.4%) were not matched in ArcGIS with a match score ≥80 (and therefore only 700 addresses were included in our positional difference analysis). Specifically, these analyses included addresses in the Pacific Northwest (n = 155), the Midwest (n = 161), the South (n = 226) and the East (n = 158) of the USA. Overall, 613 (87.6%) of these addresses had a match score of 100 and

Batchgeo yielded a 100% match rate for the 700 addresses that ArcGIS geocoded.

Table 1 summarises the distribution of the longitudinal and latitudinal difference of coordinates comparing ArcGIS versus Batchgeo for the data overall and for each geographic region. The median for longitude and latitude coordinates for all data was just over 25 m. Overall, the range for longitude was 0.04-12,911.8 m and the range for latitude was 0.02-37,766.6 m. Comparisons of each geographic region show that there were minimal differences in the median and minimum values, but there were differences in the maximum values. For example, the longitude and latitude maximum for the East was 12,911.8 and 37,766.6 m, respectively, while the longitude and latitude maximum for the Pacific Northwest was 6,807.0 and 2,619.8 m, respectively. The majority (>75%) of the longitude and latitude differences, comparing ArcGIS to Batchgeo, were within 50 m of each other; mostly <25 m from each other (about 49%). Only about 4% overall were ≥400 m apart (Table 2). We found geographic differences in the proportion of addresses that fell within certain meter ranges. The South and East longitude/latitude differences were mostly less than 25 m from each other (>50%). However, relative to these

Table 1. Positional difference of coordinates (in m) between ArcGIS 9.3 and Batchgeo geocoding methods, including all data and for each geographic region.

Distribution of the positional difference (m)	Longitude (X) difference	Latitude (Y) difference
Overall (n = 700)		
Median	25.4	25.2
Minimum	0.04	0.02
25th percentile	9.5	9.7
75th percentile	48.2	49.2
Maximum	12,991.8	37,766.6
Pacific Northwest (n = 155)		
Median	26.4	29.2
Minimum	0.22	0.21
25th percentile	14.5	17.8
75th percentile	45.8	60.3
Maximum	6,807.0	2,619.8
Midwest (n = 161)		
Median	26.9	26.6
Minimum	0.04	0.22
25th percentile	19.2	20.1
75th percentile	47.9	38.2
Maximum	8,322.5	7,145.4
South (n = 226)		
Median	24.1	17.4
Minimum	0.14	0.15
25th percentile	5.0	4.4
75th percentile	79.9	49.9
Maximum	5,744.9	9,632.8
East (n = 158)		
Median	17.3	18.3
Minimum	0.42	0.02
25th percentile	7.8	10.5
75th percentile	38.9	46.2
Maximum	12,911.8	37,766.6

Table 2. Percent and number within different distance ranges (in m) of the positional difference of coordinates between ArcGIS 9.3 and Batchgeo geocoding methods, including all data and for each geographic region.

Distance ranges of the positional difference (m)	Longitude (X) difference % (n)		Latitude (Y) difference % (n)	
Overall (n = 700)				
<25	48.9	(342)	49.4	(346)
25-49	26.9	(188)	26.6	(186)
50-74	6.6	(46)	10.0	(70)
75-99	3.9	(27)	3.6	(25)
100-199	6.3	(44)	4.1	(29)
200-299	2.4	(17)	2.0	(14)
300-399	1.1	(8)	0.1	(1)
>400	4.0	(28)	4.1	(29)
Pacific Northwest (n = 155)				
<25	43.2	(67)	40.0	(62)
25-49	34.2	(53)	27.1	(42)
50-74	7.1	(11)	16.8	(26)
75-99	3.9	(6)	5.2	(8)
100-199	7.1	(11)	6.5	(10)
200-299	0.7	(1)	1.9	(3)
300-399	--	(0)	--	(0)
>400	3.9	(6)	2.6	(4)
Midwest (n = 161)				
<25	41.0	(66)	37.3	(60)
25-49	36.7	(59)	45.3	(73)
50-74	9.9	(16)	8.1	(13)
75-99	0.6	(1)	0.6	(1)
100-199	6.2	(10)	--	(0)
200-299	1.2	(2)	1.9	(3)
300-399	1.2	(2)	--	(0)
>400	3.1	(5)	6.8	(11)
South (n = 226)				
<25	51.3	(116)	56.2	(127)
25-49	15.0	(34)	19.0	(43)
50-74	5.8	(13)	8.0	(18)
75-99	7.1	(16)	2.7	(6)
100-199	7.5	(17)	5.8	(13)
200-299	4.9	(11)	3.1	(7)
300-399	1.8	(4)	0.4	(1)
>400	6.6	(15)	4.9	(11)
East (n = 158)				
<25	58.9	(93)	61.4	(97)
25-49	26.6	(42)	17.7	(28)
50-74	3.8	(6)	8.2	(13)
75-99	2.5	(4)	6.3	(10)
100-199	3.8	(6)	3.8	(6)
200-299	1.9	(3)	0.6	(1)
300-399	1.3	(2)	--	(0)
>400	1.3	(2)	1.9	(3)

regions, the Pacific Northwest and Midwest tended to have a greater proportion of address between 25 and 50 m of each other. Additionally, there were geographic region variations in the proportion of addresses that were within 50 m of each other. For example, based on longitude differences, the proportion of addresses within 50 m of each other in the South was just over 66%, while it was just over 85% in the East. The proportion of addresses that were within 50 m of each other, based on latitude differences, was slightly over 67% for the Pacific Northwest and almost 83% for the Midwest. In addition, although some geographic variability in the amount of positional differences between the two geocoding methods were noted visually, most observations appear to be minimal (Fig. 1).

Relative to ArcGIS, the matching scores range associated with the Batchgeo accuracy level "approximate" (least accurate) was 84 to 100 (mean = 92; n = 17); for the "geometric center" Batchgeo accuracy level scores ranged from 96 to 100 (mean = 99.3; n = 29); for the "range interpolated" Batchgeo accuracy level scores ranged from 84 to 100 (mean = 98.9; n = 245); and finally for the "rooftop" Batchgeo accuracy level (most accurate) the scores range from 84 to 100 (mean 98.9; n = 409).

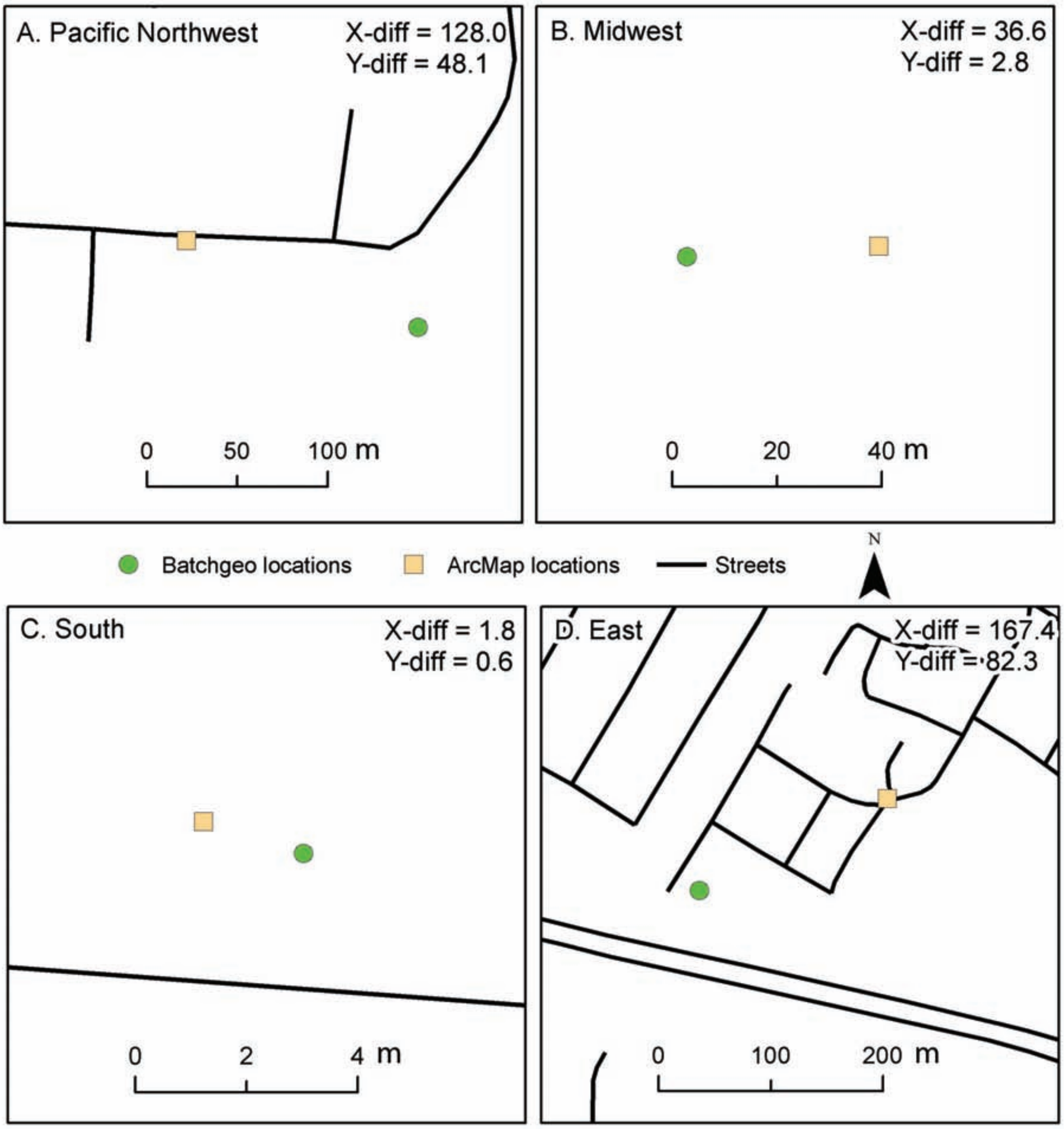

Fig. 1. Examples of the positional discrepancy of geocoded addresses between ArcGIS versus Batchgeo geocoding methods.

Discussion

Prior to conducting GIS and spatial analyses, health researchers and practitioners often need to geocode their address-based data and many do so "in-house". However, with more user-friendly, geocoding services available, the important decisions regarding geocoding sensitivity may be hidden from the user (which might not be identifiable to a non-specialist) and the geocodes obtained may be inaccurate (which can lead to substantial exposure misclassification). In this study, we compared postal addresses that were geocoded in a commonly used commercial geocoding software (ArcGIS) and also geocoded in Batchgeo, a readily and freely available tool.

The findings suggest that there is generally no disadvantage in using Batchgeo versus ArcGIS, but results may depend on the geographic region. Our findings also indicate that Batchgeo is capable to geocode addresses with higher match scores to its most accurate level, suggesting that Batchgeo's accuracy levels are generally likely to be correct. While several studies have evaluated the positional difference between different geocoding methods (Cayo and Talbot, 2003; Lovasi et al., 2007; Schootman et al., 2007; Mazumdar et al., 2008), few have done so with address data from multiple locations in the same study (Ward et al., 2005; Zhan et al., 2006). Several of these studies also have a smaller number of addresses (e.g. Ward et al., 2005; Zhan et al., 2006; Schootman et al., 2007; Mazumdar et al., 2008). Further, calculating the positional difference of latitude and longitudinal coordinates is infrequently performed. Doing so, however, can reveal directional biases not otherwise revealed by using straight distance positional offset comparisons. When geocoding data "in-house", the use of ArcGIS and Batchgeo can be combined for validity checks of certain geocodes. Of note, attention could be focused on those geocodes with a large degree of positional difference. Importantly, to our knowledge, no previous research has examined the geocodes produced by Batchgeo. This freely accessible service is easy to use and understand for individuals with limited geocoding experience. Moreover, the coordinates produced by Batchgeo can also be used for visualisation (i.e. mapping of addresses) on the Batchgeo website and embedded into other websites that the user specifies. This has implications for users such as, for example, many public health departments with minimal financial and human resources for geocoding could use Batchgeo in their geospatial work. While ArcGIS is probably the most popular commercial software for geocoding data "in-house", researchers and practitioners are beginning to use Batchgeo in their work, indicating that Batchgeo may become another widely used tool for geocoding addresses. However, it is important to note that the chosen geocoding method may depend on the nature of the project, cost restrictions and the skills of the analyst. Therefore, the increased flexibility of ArcGIS, in addition to more information known about the software, might be considered an advantage, suggesting that ArcGIS, if possible, should be the first choice to geocode addresses. In addition, ArcGIS, when used for geocoding, returns a "match score", which gives an indication of how close the address matches its actual location in the street database used. The minimum match score can be adjusted to only accept geocodes of high accuracy. A further advantage is that location and typographic adjustments can be made to individual addresses through the standard ArcGIS geocoding interface, allowing an interactive refinement of one address at a time.

In addition to the freely available Batchgeo website, other free and low-cost GIS software packages such as Quantum GIS can be used for geocoding. Quantum GIS uses the Google API to geocode addresses, but only allows for entry of one address at a time unless a custom programme is written to allow for more. The University of Southern California maintains an extensive list of available geocoders at https://webgis.usc.edu/Services/Geocode/About/GeocoderList.aspx. Although many of these are free and allow many addresses to be geocoded, they all have one or more of the following limitations: (i) allows only geocoding one address at a time; (ii) requires the creation of a user account; or (iii) includes multi-page navigation before arriving at the geocoding interface. Batchgeo is extremely user friendly and does not have these restrictions. Importantly, while a number of studies have evaluated the geocodes produced by ArcGIS, much less research has evaluated the geocodes produced by several of these alternative software packages. As recommended by Krieger et al. (2001) and others (McElroy et al., 2003; Goldberg et al., 2008; Hay et al., 2009), we encourage the reporting of address geocoding methods regardless of which method is chosen. Most geospatial health studies still do not provide such information or provide only limited information on the geocoding method used, e.g. such as only providing the match rate.

Several caveats of this study merit consideration. Although we do not know the actual location of each address (as previously noted our study examined the positional difference, not positional error), we are confident that the geocodes produced by ArcGIS are generally positionally accurate. Since the accuracy of geocodes in part depends on the quality of the street reference maps used to generate the coordinates (Drummond, 1995; Cromley and McLafferty, 2002; Rushton et al., 2006), we used the most up-to-date maps (i.e. from the ArcGIS Online World Geocoding service which uses the most recent commercial street data from Tele Atlas). This study only included geocoded addresses with the highest positional accuracy (as defined by a match score of ≥80). The "true" geographic location of each address can be determined through aerial imagery or with global positioning systems (GPS) receiver data. Though these are gold standards, this was not practical nor a central focus of the parent study. In addition, Google and Yahoo (the two companies that can be used to produce Batchgeo's geocodes) maintain extensive geographic databases, which are frequently updated, ensuring that Batchgeo has strong address-matching capabilities and a sufficiently high positional accuracy. The street base map data used by the different geocoding services play a large part in determining accurate address matches. The mapping companies Tele Atlas and NAVTEQ map and sell these base map data to companies like ESRI, Google and Yahoo, which then include them in their geocoding services. Therefore, the base map data used by the different geocoding services at any given point may vary in quality and completeness. The quality and completeness may also vary by geographic region. Thus, it is important to also document (if possible) what base map data the geocoding service used. Currently, ESRI uses Tele Atlas data, Yahoo uses NAVTEQ data, and Google, as of October 2009, uses its own street database. However, even if two geocoding services use the exact same base map data, different address-matching sensitivity settings built into the geocoder may produce different positional placements. Further, while error might be introduced due to incorrect geocodes (with correctly recorded addresses), error can also arise due to the quality of the collected addresses (i.e. it could also be due to incorrect addresses such as incorrectly spelled street names) (Cromley and McLafferty, 2002; Rushton et al., 2006; Goldberg et al., 2008). For this reason, we manually cleaned the addresses for this study prior to geocoding. Although we geocoded the same addresses that had been cleaned, it is likely that the editing of the addresses impacted the geocoding findings (e.g. improved the match rate and probably also increased the positional accuracy) (Cromley and McLafferty, 2002; McElroy et al., 2003; Rushton et al., 2006; Goldberg et al., 2008). Additionally, we used interactive geocoding to investigate ties in order to yield the highest possible match rate and increase the positional accuracy. Our use of interactive rematching is likely to have affected the geocodes included in this study. It is also important to note that, in addition to the settings used, different programmes, or even different versions of the same geocoding software, might produce different results (Drummond, 1995). Since each of the elements discussed can influence the results, we suggest that future projects take these aspects into consideration when geocoding and examining differences between geocoding methods.

In conclusion, although this study indicates that positional differences between the two geocoding methods examined exist, the medians of the differences found with ArcGIS and Batchgeo were minimal and most addresses were placed only a short distance apart. Although future research should compare the positional difference of Batchgeo to criterion measures of longitude/latitude (e.g. with GPS measurement), we feel that Batchgeo is a free and powerful alternative when geocoding addresses, a much relevant task for health researchers and practitioners with limited experience in this field.

Acknowledgements

D.T. Duncan was supported by an Association of Schools of Public Health/Centers for Disease Control and Prevention/ Prevention Research Center Minority Health Fellowship, a Julius B. Richmond Fellowship from the Center on the Developing Child at Harvard University and by a grant from Robert Wood Johnson Foundation's Active Living Research Programme (grant # 67129). M.C. Castro was supported in part by a grant from the National Institutes of Health (grant # 5R21AI084015-02). S.L. Gortmaker was supported in part by grants from the Donald and Sue Pritzker Nutrition and Fitness Initiative and from the Centers for Disease Control and Prevention (Prevention Research Centers Grant # U48DP001946). We thank Nancy Krieger, Jarvis Chen, Pamela Waterman, Gina Lovasi, Steven Melly, Angie Cradock and Giovanni Zambotti for commenting on our initial ArcGIS geocoding methods and for providing useful comments on geocoding best practices, which inspired this study. We thank Rebecca Lundin for assisting with cleaning the addresses, as well as Rebekka Lee and Rebecca Mozaffarian for entering the addresses in Excel.

References

Cayo MR, Talbot TO, 2003. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2, 10.

Coulton CJ, Korbin J, Chan T, Su M, 2001. Mapping residents' perceptions of neighbourhood boundaries: a methodological note. Am J Community Psychol 29, 371-383.

Cromley EK, McLafferty SL, 2002. GIS and Public Health. The Guildford Press, New York, NY, USA, 340 pp.

Drummond WJ, 1995. Address matching: GIS technology for mapping human activity patterns. J Am Planning Assoc 61, 240-251.

Elliott P, Wakefield J, Best N, Briggs D, 2000. Spatial Epidemiology: Methods and Applications, Oxford University Press, Oxford, UK, 475 pp.

Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG, 2008. An effective and efficient approach for manually improving geocoded data. Int J Health Geogr 7, 60.

Hay G, Kypri K, Whigham P, Langley J, 2009. Potential biases due to geocoding error in spatial analyses of official data. Health Place 15, 562-567.

Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P, 2003. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 14, 386-391.

Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R, 2002. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas - the Public Health Disparities Geocoding Project. Am J Public Health 92, 1100-1102.

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, 1114-1116.

Lee BA, Reardon SF, Firebaugh G, Farrell CR, Matthews SA, O'Sullivan D, 2008. Beyond the census tract: patterns and determinants of racial segregation at multiple geographic scales. Am Soc Rev 73, 766-791.

Lovasi GS, Weiss JC, Hoskins R, Whitsel EA, Rice K, Erickson CF, Psaty BM, 2007. Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree? Int J Health Geogr 6, 12.

Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ, 2008. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr 7, 13.

McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA, 2003. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 14, 399-407.

Miller HJ, 2009. Geocomputation. In: The SAGE Handbook of Spatial Analysis. Fotheringham AS, Rogerson PA (eds). SAGE Publications, Thousand Oaks, CA, USA, 397-418.

Moore DA, Carpenter TE, 1999. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol Rev 21, 143-161.

Osypuk T, Galea S, 2007. What Level Macro? Choosing Appropriate Levels to Assess How Place Influences Population Health. In: Macrosocial Determinants of Population Health. S. Galea (ed). Springer Media, New York, NY, USA, pp.399-436.

Rushton G, 2003. Public health, GIS, and spatial analytic tools. Annu Rev Public Health 24, 43-56.

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, S16-24.

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, 464-470.

United States Postal Service, 2000. Postal Addressing Standards-Publication 28. Available at: http://pe.usps.com/cpim/ftp/pubs/Pub28/pub28.pdf

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, 542-547.

Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G, 2006. Accuracy of commercial geocoding: assessment and implications. Epidemiol Perspect Innov 3, 8.

Whitsel EA, Rose KM, Wood JL, Henley AC, Liao D, Heiss G, 2004. Accuracy and repeatability of commercial geocoding. Am J Epidemiol 160, 1023-1029.

Zandbergen PA, Green JW, 2007. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environ Health Perspect 115, 1363-1370.

Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH, 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol 16, 842-849.

Zimmerman DL, Sun P, 2006. Estimating spatial intensity and variation in risk from locations subject to geocoding errors. Available at: http://www.stat.uiowa.edu/techrep /tr363.pdf