

Positional Accuracy of Geocoded Addresses in Epidemiologic Research

Matthew R. Bonner,* Daikwon Han,† Jing Nie,* Peter Rogerson,† John E. Vena,* and Jo L. Freudenheim*

Background: Geographic information systems (GIS) offer powerful techniques for epidemiologists. Geocoding is an important step in the use of GIS in epidemiologic research, and the validity of epidemiologic studies using this methodology depends, in part, on the positional accuracy of the geocoding process.

Methods: We conducted a study comparing the validity of positions geocoded with a commercially available program to positions determined by Global Positioning System (GPS) satellite receivers. Addresses (N = 200) were randomly selected from a recently completed case-control study in Western New York State. We geocoded addresses using ArcView 3.2 on the GDT Dynamap/2000 U.S. Street database. In addition, we measured the longitude and latitude of these addresses with a GPS receiver. The distance between the locations obtained by these two methods was calculated for all addresses.

Results: The distance between the geocoded point and the GPS point was within 100 m for the majority of subject addresses (79%), with only a small proportion (3%) having a distance greater than 800 m. The overall median distance between GPS points and geocoded points was 38 m (90% confidence interval [CI] = 34–46). Distances were not different for cases and controls. Urban addresses (median = 32 m; CI = 28–37) were slightly more accurate than nonurban addresses (median = 52 m; CI = 44–61).

Conclusions. This study indicates that the suitability of geocoding for epidemiologic research depends on the level of spatial resolution required to assess exposure. Although sources of error in positional

accuracy for geocoded addresses exist, geocoding of addresses is, for the most part, very accurate.

Key Words: geographic information systems, geocoding, address matching, epidemiology

(*Epidemiology* 2003;14: 408–412)

Geographic information systems (GIS) are increasingly used by epidemiologists to screen and test hypotheses about environmental exposures and disease.^{1–4} GIS techniques lend themselves to assessing residential or occupational proximity to exposures and to assessing spatial variation in epidemiologic measures. An important first step is often to geocode the study participant addresses.⁴ Geocoding, also referred to as address matching, is the process whereby the relative positions of addresses are linked to a reference theme, which is a database that contains both address information and locational information (ie, latitude and longitude). A reference theme for geocoding, therefore, can be considered an electronic version of a street map. Geocoding is an attractive method for epidemiologists because GIS software is relatively inexpensive, uses routinely collected address data, and is very efficient at locating addresses. Verification of each subject's address location by other methods would require considerable time and resources, especially for a large study. Furthermore, verifying address locations is not feasible when subjects' lifetime series of addresses are considered because the number of these addresses can become extremely large.

The validity of epidemiologic studies using GIS and geocoding methods depends on the proportion of addresses that can be geocoded as well as the positional accuracy of the geocoding process. Several studies have assessed the address matching rate of commercial geocoding companies, and found that matching rates are typically 60–80%.^{1,5} No previous published studies have assessed the positional accuracy of geocoding in epidemiologic research. Positional inaccuracy of geocoded addresses may be an important source of exposure misclassification in environmental epidemiology.

Editor's note: an invited commentary on this article appears on page 384. Submitted 17 September 2002; final version accepted 18 March 2003.

From the *Department of Social and Preventive Medicine, School of Medicine and Biomedical Sciences, and †Department of Geography, University at Buffalo, Buffalo, NY.

This study was supported in part by Medical Sociology, Epidemiology, and Control of Cancer CA-09051 NCI, Breast Cancer: Residential Environment and Genetics 5 R21CA8713802 NCI, and Environmental Exposure at Birth and at Menarche and Risk of Breast Cancer DAMD-170010417 U.S. Army Medical Research and Materiel Command.

Correspondence: Matthew R. Bonner, Department of Social and Preventive Medicine, School of Medicine and Biomedical Sciences, Room 270 Farber Hall, University at Buffalo, Buffalo, NY 14214. E-mail: Mrbonner@buffalo.edu.

Copyright © 2003 by Lippincott Williams & Wilkins
1044-3983/03/1404-0408

DOI: 10.1097/01.EDE.0000073121.63254.c5

We describe here a study comparing the location of addresses measured by global positioning system (GPS) receivers (devices that use satellite signals to estimate the latitude and longitude of any given location) to positions geocoded with a commercially available reference theme. We assessed three areas that may be important to determine the appropriateness of geocoding in epidemiologic research. First, we compared the positional accuracy of historical addresses. Many exposures that are relevant to disease outcomes are historical in nature and positional inaccuracy of geocoding historical addresses may be a source of error in estimating these historical exposures. Second, we investigated whether positional inaccuracy of geocoded addresses would result in differential exposure misclassification between cases and controls. Third, we compared differences in positional accuracy between urban and nonurban areas. This urban–rural differential is particularly important because the reference themes commonly available were designed for non-epidemiologic purposes and are generally thought to be more accurate and complete in urban areas than in nonurban areas.

METHODS

We obtained a random sample of 200 addresses from a recently completed case–control study in Erie and Niagara Counties in Western New York State. Lifetime residential histories were collected from 3,286 subjects for a total of 20,240 addresses. These included study participants' current addresses and all previous addresses dating back to 1918. For the remainder of this article, addresses before the current address of each participant are termed "historical addresses." Most addresses ($N = 15,903$) were for residences in Erie and Niagara Counties. We geocoded Erie and Niagara County addresses using ArcView 3.2 (ESRI, Inc., Redlands, CA) and the Dynamap/2000 US Street Database (Geographic Data Technologies, Inc., Lebanon, NH) for Erie and Niagara Counties as the reference theme. Essentially, the Dynamap/2000 is an enhancement of the Topologically Integrated Geographic Encoding and Referencing file (TIGER/line file) that was developed by the US Bureau of the Census. These are data files that contain street address ranges and census tract/block boundaries.⁶ We matched 10,356 (65%) of the original 15,903 addresses using the initial geocoding process.

To ensure an adequate number of cases and controls for comparison of urban and nonurban positional accuracy, we randomly selected 200 addresses in a random block selection scheme to obtain 50 cases and 50 controls from urban areas and 50 cases and 50 controls from nonurban areas. We defined urban addresses as addresses within the city limits of Buffalo, Niagara Falls, and Kenmore, NY. All other addresses were considered nonurban. If an address could not be located for the GPS measurements, then that address was discarded and a new address was randomly selected from the same block.

We determined the geocoded latitude and longitude for each address with ArcView 3.2 by first geocoding the addresses and then changing the map projection to Universal Transverse Mercator-1983. This projection compensates for the Earth's curvature and generates more precise estimates of latitude and longitude than other projections. Latitude and longitude were then converted into x and y coordinates (arbitrary values representing a point on a plane) for each address. These x and y coordinates were measured in meters for this study.

We measured the actual geographic position of the 200 addresses with an Etrex GPS receiver manufactured by Garmin (Garmin International, Inc., Olathe, KS). This GPS receiver reported latitude and longitude in decimal degrees to five places using the World Geodetic System 1984 map datum. Before making site visits, the GPS receiver was turned on and automatically searched for least three satellite signals. Once the satellite signals were detected, the GPS receiver provides real-time current latitude, longitude, speed, and direction. Investigators then visited each address and used the GPS receiver to record latitude and longitude from the street directly in front of each address.

The observed GPS latitudes and longitudes were then converted to x and y coordinates in Universal Transverse Mercator-1983 projection for comparison between the GPS and the geocoded positions. We calculated the distance between the two points for each address by determining the Euclidean length of the hypotenuse of the right triangle formed by the two points: $[(x_1 - x_2)^2 + (y_1 - y_2)^2]^{1/2}$, where x_1 is the GPS latitude, x_2 is the geocoded latitude, y_1 is the GPS longitude, and y_2 is the geocoded longitude. This formula does not correct for the curvature of the Earth. However, this uncorrected formula does not introduce sizable error in the distance calculation between the two points because the distances between the points were relatively small.

The mean distance, its standard deviation, and the median distance between the geocoded points and the GPS points were calculated with SPSS version 10.1 for the total sample, for case and control addresses, for urban and nonurban addresses, and for cases and controls stratified by urban/nonurban status. We grouped distance into nine categories: <25 m, 25–50 m, 51–75 m, 76–99 m, 100–199 m, 200–399 m, 400–599 m, 600–799 m, and >800 m; the proportion of addresses in each of these categories was then computed. Bootstrapped 90% confidence intervals for the medians of distance were computed with Resampling Stats (Resampling Stats, Inc., Arlington, VA).

RESULTS

The majority of the 200 randomly selected addresses ($N = 133$) were historical in nature; subjects did not currently occupy these addresses (Table 1). The median distances

TABLE 1. Distance (m) Between Geocoded Position and GPS Position for 200 Addresses, by the Year Moved Out of Address

Year Moved Out of Address*	% of Addresses (N = 200)	Median Distance (90% CI)	Minimum Distance	Maximum Distance
1930–1939	1.5	33 (16–50)	16	50
1940–1949	8.0	40 (32–53)	17	1225
1950–1959	11.5	38 (26–53)	9	502
1960–1969	11.5	29 (26–38)	9	760
1970–1979	14.5	36 (29–71)	7	2552
1980–1989	12.0	38 (28–59)	5	313
1990–2000	7.5	34 (28–70)	12	209
Currently occupy	32.5	49 (38–59)	6	763
Unknown	1.0	96 (19–172)	19	172
Total	100	38 (34–46)	5	2552

GPS = global positioning system.
*Indicates the decade when a study participant moved out of the address.

between the GPS and the geocoded position did not vary greatly across decades. For the current addresses, there was a slightly larger median distance (49 m; 90% confidence interval [CI] = 38–59) between the measured address location and the geocoded location, whereas the distances for the historical addresses tended to be between 30 and 40 m. However, the three addresses with distances greater than 1,000 m were all historical addresses.

Tables 2 and 3 present comparisons of the geocoded

and GPS positions. The distribution of distances between geocoded and GPS positions is skewed to the right, as evidenced by the median distance for all addresses (38 m; 90% CI = 34–46) being considerably smaller than the mean distance (113 m; Table 2). Consequently, the median is more accurate than the mean as a measure of central tendency. The distance between the geocoded point and the GPS point was within 100 m for the majority of the all subject addresses (79%), with only a small proportion (3%) having a distance greater than 800 m. Distances were not different for cases and controls.

Positional accuracy was somewhat better for the urban addresses (32 m; 90% CI = 28–37) than for the nonurban addresses (52 m; 90% CI = 44–61) (Table 3). In addition to having a smaller median distance, urban addresses had a higher proportion of addresses within 100 m (89%) than the nonurban addresses (69%). Within the urban strata, cases and controls were more similar than the cases and controls in the nonurban strata. In the nonurban strata, there was a 9-m difference in the medians between cases (45 m; 90% CI = 40–70) and controls (54 m; 90% CI = 38–66). In addition, urban cases (86%) and controls (92%) had a higher proportion of addresses within 100 m than did nonurban cases (70%) and controls (68%).

DISCUSSION

Overall, the positional accuracy of addresses geocoded with the Dynamap/2000 was good. The majority of addresses were located within 100 m of the real address as determined by on-site GPS latitude and longitude measurements; the historical addresses tended to have smaller median distances than the current addresses. There was, however, some difficulty in assessing all the selected historical addresses. In eight

TABLE 2. Distance Between Geocoded Position and GPS Position for Case and Control Addresses

	Cases (N = 100)	Controls (N = 100)	All Addresses (N = 200)
Distance (m)			
Median	41	38	38
90% CI	31–47	33–49	34–46
Mean (SD)	119 (247)	107 (286)	113 (266)
Minimum distance	5	5	5
Maximum distance	1151	2552	2552
Distance (%)			
<25 m	28	26	27
25–50 m	33	34	33.5
51–75 m	12	16	14
76–99 m	5	4	4.5
100–199 m	10	9	9.5
200–399 m	5	8	6.5
400–599 m	1	1	1
600–799 m	2	0	1
≥800 m	4	2	3

GPS = global positioning system; SD = standard deviation.

TABLE 3. Distance Between Geocoded Position and GPS Position for Case and Control Addresses, by Urban/Nonurban Residential Status

	Urban			Non-Urban		
	Cases (N = 50)	Controls (N = 50)	Total (N = 100)	Cases (N = 50)	Controls (N = 50)	Total (N = 100)
Distance (m)						
Median	31	32	32	45	54	52
90% CI	27–41	28–37	28–37	40–70	38–66	44–61
Mean (SD)	122 (285)	70 (177)	96 (237)	116 (204)	125 (362)	129 (293)
Minimum distance	6	5	5	5	12	5
Maximum distance	1551	1225	1551	1223	2552	2552
Distance (%)						
<25 m	36	30	33	20	22	21
26–50 m	34	46	40	32	22	27
51–75 m	14	12	13	10	20	15
76–99 m	2	4	3	8	4	6
100–199 m	4	2	3	16	16	16
200–399 m	0	4	2	10	12	11
400–599 m	2	0	1	0	20	10
600–799 m	2	0	1	2	00	10
≥800 m	6	2	4	20	20	20

GPS = global positioning system.

instances, subjects' homes appeared to have been demolished, leaving a vacant lot. The uncertainty about whether these lots were the correct street address prevented GPS measurements of these addresses. This indicates that in areas where there has been major redevelopment or neglect the positional accuracy of historical addresses may be more difficult to determine.

Positional accuracy was not different between cases and controls, suggesting that errors resulting from the geocoding of addresses may not result in differential misclassification of exposure. In addition, the difference in positional accuracy between urban addresses and nonurban addresses was small. However, even with good overall positional accuracy, there were several sources of error. First, the Dynamap/2000 is largely derived from the US Bureau of the Census TIGER/line files, and inaccurate methods were used to create these TIGER/line files. The Geography Division of the US Bureau of the Census has reported that the median distances between GPS measured positions and the TIGER/line file positions in eight US counties ranged between 30 and 121 m.⁷ Additionally, this report also indicated that TIGER/line file updates since 1990 are less accurate than the pre-1990 versions. These inaccuracies may have important implications for epidemiologic use because, as the updates to the TIGER/line files provide more complete coverage and increase the address-matching rate, the decrease in positional accuracy may lead to increased error and misclassification of

exposure when estimating exposures based on geographic positioning.

The second source of error with geocoding arises from the geocoding process itself. We found that positional accuracy was slightly decreased in nonurban addresses compared with urban addresses, likely a result of the geocoding process. Geocoding uses interpolation to estimate the relative position of an address on a line segment in the reference theme.^{4,8} The likelihood of inaccurate interpolation by the geocoding process is higher for an address location in areas with long street segments than in areas where there are many short street segments, regardless of urban/nonurban location; however, nonurban areas generally have longer street segments than urban areas.

In addition to error in positional accuracy from the Dynamap/2000 and geocoding, GPS receivers are also prone to error. These errors are generally small but remain a limitation of this study in that the GPS receiver was used as the standard. GPS errors in positional accuracy arise from three general sources: satellite related errors, signal propagation errors, and receiver errors. Garmin reports that the Etrex has positional accuracy between 1 and 5m. Field tests, however, indicate that civilian GPS receivers are only accurate within 15–40 m.⁹ The Dynamap/2000 may actually be more accurate than we report here because most of the distances between the GPS point and the geocoded point were within the range of error for the GPS unit.

These errors in the positional accuracy of geocoding study participant addresses and sources of exposure have several implications for epidemiologic research. First, numerous epidemiologic studies have crudely defined exposure based solely on proximity of a residential address to an exposure of interest. For instance, McLaughlin and associates¹⁰ used a 25-km radius around nuclear facilities in Canada to classify those residing within that circle as exposed and those outside as unexposed. Clearly, geocoding has sufficient spatial resolution to distinguish differences on this scale. In another study, Croen et al¹¹ investigated maternal residential proximity to hazardous waste sites and congenital malformations. Maternal residence within 1 mile (1.6 km) of a hazardous waste site was defined as exposed. Again, even with the higher required spatial resolution to define exposure, geocoding should have sufficient positional accuracy to classify exposure appropriately.

There are exposures, however, with great spatial variation over relatively small distances. When these exposures are being considered, the study may require more precise methods to locate subjects and exposure sources to produce valid risk estimates. Electromagnetic fields, for instance, require high spatial resolution to estimate exposure accurately. The intensity of magnetic fields decreases exponentially with distance. In many epidemiologic studies that investigate electromagnetic fields and cancer, residential proximity to power lines and electricity transmission equipment has been measured in meters rather than kilometers or miles as in the previous examples. For example, in their meta-analysis of residential proximity to electricity transmission and distribution equipment and childhood cancer, Washburn et al¹² used less than 50 m to define the exposed group. Based on the present validation of geocoding, there is some question whether the positional accuracy of geocoding is sufficient. Use of geocoding in situations where high spatial resolution is required may lead to extensive nondifferential misclassification of exposure, thereby greatly reducing the validity of the risk estimates.

The generalizability of this study may be limited by regional variation in the completeness of the Dynamap/2000. These results should be comparable to regions where the Dynamap/2000 has completeness similar to that of Erie and Niagara Counties. Furthermore, as the US Bureau of the Census and GDT continue to develop and improve the TIGER/line and the Dynamap/2000 databases, regional variation will diminish. However, recent improvements in the

completeness of these updated street databases may not alleviate the concerns about their positional accuracy because post-1990 updates have not been as carefully assembled as the pre-1990 updates.

Finally, as GIS becomes more commonly used in epidemiologic research, the need to assess geocoding methods and reference themes will become more important, especially when high spatial resolution is required to classify a study subject's exposure accurately. Overall, this study indicates that these tools are sufficiently accurate for some—but not all—epidemiologic studies. Consequently, care should be taken in the interpretation of results, taking into account sources of error in positional accuracy for geocoded addresses that may affect exposure classification.

ACKNOWLEDGMENTS

We thank Dominica Vito and Julie LaFalce for their efforts in collecting these data.

REFERENCES

1. Krieger N, Waterman P, Lemieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health*. 2001;91:1114–1116.
2. Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiol Rev*. 1999;21:143–161.
3. Bellander T, Berglund N, Gustavsson P, et al. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environ Health Perspect*. 2001;109:633–639.
4. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. *Environ Health Perspect*. 1997;105:598–605.
5. Howe HL. Geocoding NY State Cancer Registry. *Am J Public Health*. 1986;76:1459–1460.
6. Anonymous. *Getting to Know ArcView GIS*. 3rd ed. Redlands, CA: Environmental Systems Research Institute, Inc.; 1999.
7. Liadis JS. *GPS TIGER Accuracy Analysis Tools (GTAAT) Evaluation and Test Results*. United States Bureau of the Census, Division of Geography, TIGER Operations Branch, 2000.
8. Drummond WJ. Address matching: GIS technology for mapping human activity patterns. *J Am Planning Assoc*. 1995;61:240–251.
9. Hofmann-Wellenhof B, Lichtenegger H, Collins J. *Global Positioning Systems: Theory and Practice*. New York: Springer-Verlag; 1997.
10. McLaughlin JR, Clarke EA, Nishri ED, Anderson TW. Childhood leukemia in the vicinity of Canadian nuclear facilities. *Cancer Causes Control*. 1993;4:51–58.
11. Croen LA, Shaw GM, Sanbonmatsu L, et al. Maternal residential proximity to hazardous waste sites and risk for selected congenital malformations. *Epidemiology*. 1997;8:347–354.
12. Washburn EP, Orza MJ, Berlin JA, et al. Residential proximity to electricity transmission and distribution equipment and risk of childhood leukemia, childhood lymphoma, and childhood nervous system tumors: systematic review, evaluation, and meta-analysis. *Cancer Causes Control*. 1994;5:299–309.