On the Wrong Side of the Tracts? Evaluating the Accuracy of Geocoding in Public Health Research

Nancy Krieger, PhD, Pamela Waterman, MPH, Kerry Lemieux, MS, Sally Zierler, DrPH, and Joseph W. Hogan, PhD

ABSTRACT

Objectives. This study sought to determine the accuracy of geocoding for public health databases.

Methods. A test file of 70 addresses, 50 of which involved errors, was generated, and the file was geocoded to the census tract and block group levels by 4 commercial geocoding firms. Also, the "real world" accuracy of the best-performing firm was evaluated.

Results. Accuracy rates in regard to geocoding of the test file ranged from 44% (95% confidence interval [CI] = 32%, 56%) to 84% (95% CI = 73%, 92%). The geocoding firm identified as having the best accuracy rate correctly geocoded 96% of the addresses obtained from the public health databases.

Conclusions. Public health studies involving geocoded databases should evaluate and report on methods used to verify accuracy. (*Am J Public Health.* 2001;91:1114–1116)

Geocoding and geographic information systems (GIS) technology are increasingly used in public health research and practice.[1–4] At issue is determining where people reside and using this information to understand population distributions of health, whether in relation to neighborhood socioeconomic conditions,[5–8] proximity to environmental health hazards,[4,9] or spatial distributions of cases.[4,10]

Geocoding of public health surveillance systems is increasingly carried out by state health departments, as is geocoding of databases for study populations established for large-scale epidemiologic investigations.[4,11–13] Indeed, the US government's report *Healthy People 2010* set a goal of geocoding 90% "of all major National, State, and local health data systems to promote nationwide use of [GIS] at all levels."[14(p23-10)] Likewise, the National Cancer Institute recently prioritized "geographic-based research on cancer control and epidemiology."[15]

As is true for any type of measurement, geocoding is not an error-free process. Addresses can be incorrectly recorded (e.g., misspelled street name) or correctly recorded but assigned the wrong geocode (e.g., latitude, longitude, census block group, or census tract errors). Public health articles on geocoding methodology, however, have focused chiefly on the capabilities and user-friendliness of geocoding software rather than the accuracy of geocoding itself.[4,16–18] In light of our own need to select an appropriate geocoding source for a project involving approximately 1 million records from 2 state health departments,[19] we developed and tested a protocol designed to evaluate the accuracy, cost, timeliness, and quality of customer services of commercial geocoding firms.

Methods

Design

Our protocol involved 3 components: (1) selecting geocoding firms for evaluation; (2) sending them a test file with previously evaluated addresses, many with known problems; and (3) sending the most accurate firm an additional test file of addresses randomly selected from public health databases. We initially identified 5 firms for evaluation on the basis of an Internet search for firms focused strictly on geocoding (as opposed to commercial applications to enable companies to evaluate potential markets for their products) and recommendations from staff at the 2 state health departments involved in our study (Massachusetts Department of Public Health and Rhode Island Department of Health). We eliminated 1 firm because it served only as a "broker" for one of the other identified firms.

To aid in assessing each firm's services, we developed a standard set of questions concerning data processing protocol (including method of geocoding, preferences for data transfer media, and desired address format), confidentiality policies, previous success rates, and estimated cost and time to geocode our project's anticipated number of records (1.5 million records in 6 batches of 250 000 addresses each). All firms indicated that, as standard practice, they used the most recent street address data available to geocode records. Three firms relied on at least 2 data sources, including US census Topographically Integrated Geographic Encoding and Reference System (TIGER) files, US postal service data, and their own GIS software, and one firm did not disclose its data source.

Nancy Krieger, Pamela Waterman, and Kerry Lemieux are with the Department of Health and Social Behavior, Harvard School of Public Health, Boston, Mass. Sally Zierler and Joseph W. Hogan are with the Department of Community Health, Brown University Medical School, Providence, RI.

Requests for reprints should be sent to Nancy Krieger, PhD, Department of Health and Social Behavior, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115 (e-mail: nkrieger@hsph.harvard.edu).

This brief was accepted July 19, 2001.

After obtaining answers to our specified questions and evaluating the quality of customer services (e.g., whether our telephone calls or e-mails were returned promptly or slowly, whether the customer representative was knowledgeable or ill informed about the company's services), we sent each firm the same test file to geocode. Of the 70 addresses included in this file, 50 were "incorrect" street addresses with known errors generated by a geocoding specialist at the Massachusetts Cancer Registry. These "incorrect" street addresses were modified versions of existing addresses altered to include common errors (e.g., out-of-range address numbers, abbreviated or misspelled street names, and correct towns but incorrect zip codes). Census tracts of the correct versions of these 50 addresses had previously been identified by the geocoding specialist using census block group maps and street guides.

We supplemented the 50 "incorrect" addresses with 20 telephone book addresses selected at random from the local telephone book and identified their tract numbers using the US Bureau of the Census geocoding Web site.[20] Accuracy of matching was assessed in relation to correct identification of the census block group geocodes of each address via street address information.

On the basis of the test file results, we selected the best firm and further evaluated its "real world" accuracy by using a random sample of 150 addresses derived from public health databases. Half of the addresses were randomly selected from Massachusetts death certificate data (with addresses from Boston, however, oversampled to facilitate sight verification of geocoding accuracy), and half were randomly selected from Rhode Island birth certificate data. Of these 150 addresses, 8 were not geocodable because (1) street names were not included in relevant street guides or atlases or (2) numbers were out of range for streets sufficiently long to cross block group boundaries.

Using the remaining 142 randomly selected addresses, we compared the geocodes assigned by the firm with those obtained through (1) block group maps produced by the US Bureau of the Census[21] and (2) the US Bureau of the Census geocoding Web site, which geocodes only to the census tract level.[20] Finally, we "sight verified" a convenience sample of 10 addresses located in economically diverse areas of Boston.

Statistical Analysis

We estimated the percentages of addresses correctly geocoded by each geocoding resource and the binomial confidence intervals for these percentages.[22] We summarize this information along with data on cost and quality of services.

Results

Test File

The selected geocoding firms differed considerably in the areas of accuracy, cost, timeliness, and customer service (Table 1). Geocoding accuracy rates among the different firms, for the full set of 70 test addresses, ranged from 44% to 84%. Accuracy rates ranged from 36% to 80% for the 50 "incorrect" addresses and from 65% to 100% for the 20 telephone book addresses. Estimated costs for the proposed scope of work (i.e., geocoding 1.5 million records) ranged from $8800 to $15 800.

Timeliness of geocoding the test file likewise varied markedly, ranging from 5 hours to 21 days; actual time exceeded estimated time for 2 of the 4 firms. Quality of customer service, ascertained before accuracy of geocoding was evaluated, also was mixed, ranging from informative, responsive, and friendly to uninformative, unresponsive, and rude. Taking into consideration accuracy, cost, quality of customer service, and turnaround time, we selected company A (Table 1), which provided the best service for the lowest cost, for our next phase of evaluation.

Real World Accuracy

Relative to the US Bureau of the Census block group maps, company A correctly geocoded, to the block group level, 96% of the 142 geocodable addresses randomly selected from the death and birth certificate data. Similarly, at the tract level, the accuracy rate of company A was high (95%) and equivalent to that of the US Bureau of the Census geocoding Web site (94%). In addition, all 10 "sight verified" addresses were assigned to the correct census tract and block group. Finally, the cost of geocoding the project's selected databases (containing nearly 1 million records) was $9114, only 4% higher than the cost company A had initially estimated for the job.

Discussion

Our results, based on addresses for a geocoding project involving 2 New England states, indicate that accuracy and cost of geocoding can vary dramatically across commercial geocoding firms. We accordingly recommend that all public health projects involving geocoding evaluate and report on methods to verify the accuracy of their geocoding methodology.

In addition, future research should evaluate both (1) variability in accuracy of geo-

TABLE 1—Evaluation of 4 US Commercial Geocoding Firms, 1999: Estimated Costs and Time, Quality of Customer Service, and Accuracy of Geocoding Test File

		Time to Geocode Test File			Addresses Geocoded Correctly, % (95% Confidence Interval)		
	Estimated Cost,[a] $	Estimated	Actual	Qualitative Evaluation of Customer Service[b]	Total (n = 70)	Incorrect Addresses (n = 50)	Telephone Book Addresses (n = 20)
Company A	8800	7–14 d	5 d	☺☺	84 (73, 92)	80 (66, 90)	95 (75, 100)
Company B	15800	≤7 d	21 d	☹	76 (64, 85)	68 (53, 80)	95 (75, 100)
Company C	13541	7 d	5 h	😐	74 (62, 84)	64 (49, 77)	100 (83, 100)
Company D	13485	7–14 d	19 d	☹☹	44 (41, 65)	36 (23, 51)	65 (41, 85)

[a] Based on 6 submissions of 250 000 records (total: 1.5 million records).

[b] ☺ = returned telephone calls or e-mails promptly; friendly, personable; displayed impressive knowledge of geocoding processes and techniques; asked unprompted questions about project; offered unprompted helpful observations or tips for project. 😐 = returned phone calls or e-mails, reasonably pleasant, did not appear particularly knowledgeable of geocoding process, displayed minimal interest in project. ☹ = very slow to respond to telephone calls or e-mails (required repeated attempts), brusque or rude, displayed no perceptible interest in project. Number of icons reflects intensity of evaluation. Number of different sales representatives contacted at each company: Company A = 2, Company B = 1, Company C = 2, Company D = 3.

coding using a larger database comprising a national sample of addresses and (2) the impact of such variability (including random as well as systematic error) on analyses using geocoded public health databases. By improving the rigor of geocoding methodology for public health databases, public health researchers and practitioners will expand possibilities for ascertaining the impact of social and environmental conditions on the public's health. □

Contributors

N. Krieger planned the study, directed data analysis, participated in sight verification of addresses, and wrote the paper. P. Waterman contacted the geocoding firms and, with K. Lemieux, carried out the work to verify accuracy of geocoding of the test file and the addresses from the public health databases. S. Zierler and J. W. Hogan assisted in conceptualizing the study, guiding data analysis, and interpreting the data.

Acknowledgments

Work on this project was supported by National Institutes of Health grant 1 R01 HD36865-01.

The study was approved by the Harvard School of Public Health Human Subjects Research Committee and the institutional review boards of Brown University, the Massachusetts Department of Public Health, and the Rhode Island Department of Health.

We acknowledge the following individuals for their substantial contributions to our study: Daniel J. Friedman, assistant commissioner, Massachusetts Department of Public Health; Jay Buechner, chief, Office of Health Statistics, Rhode Island Department of Health; Mary Mroszcyczk, geocoding/special projects coordinator, Cancer Registry, Massachusetts Department of Public Health; Elaine Trudeau, registrar of vital records, Massachusetts Department of Public Health; Roberta Chevoya, state registrar of vital records, Rhode Island Department of Health; Jarvis Chen, research assistant, Harvard School of Public Health; and Annie Gjelsvick, research assistant, Brown University.

References

1. Thrall GI. The future of GIS in public health management and practice. *J Public Health Manage Pract.* 1999;5:75–82.
2. Richards TB, Croner CM, Rushton G, Brown CK, Fowler L. Geographic information systems and public health: mapping the future. *Public Health Rep.* 1999;114:359–373.
3. Moore DA, Carpenter TE. Spatial analytic methods and geographic information systems: use in health research and epidemiology. *Epidemiol Rev*. 1999;21:143–161.
4. Williams RC, Howie MM, Lee CV, Henriques WD, eds. Geographic information systems in public health: proceedings of the Third National Conference, San Diego, 1998. Available at: http://www.atsdr.cdc.gov/gis/conference98. Accessed September 25, 2000.
5. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *Am J Public Health*. 1992;82:703–710.
6. Krieger N, Quesenberry C Jr, Peng T, et al. Social class, race/ethnicity, and incidence of breast, cervix, colon, lung, and prostate cancer among Asian, Black, Hispanic, and White residents of the San Francisco Bay Area, 1988–92 (United States). *Cancer Causes Control*. 1999;10: 525–537.
7. Chen FM, Breiman RF, Farley M, Plikaytis B, Deaver K, Cetron MS. Geocoding and linking data from population-based surveillance and the US census to evaluate the impact of median household income on the epidemiology of invasive *Streptococcus pneumoniae* infections. *Am J Epidemiol.* 1998;148:1212–1218.
8. O'Campo P, Xue X, Wang MC, Caughy M. Neighborhood risk factors for low birthweight in Baltimore: a multilevel analysis. *Am J Public Health*. 1997;87:1113–1118.
9. Bouton PB, Fraser M. Local health departments and GIS: the perspective of the National Association of County and City Health Officers. *J Public Health Manage Pract.* 1999;5:33–41.
10. Wilkinson SL, Gobalet JG, Majoros M, Zebrowski B, Olivas GS. Lead hot zones and childhood lead poisoning cases, Santa Clara County, California, 1998. *J Public Health Manage Pract.* 1999;5:11–12.
11. Krieger N, Chen JT, Ebel G. Can we monitor socioeconomic inequalities in health? A survey of US health departments' data collection and reporting practices. *Public Health Rep.* 1997; 112:481–491.
12. MacDorman MF, Gay GA. State initiatives in geocoding vital statistics data. *J Public Health Manage Pract.* 1999;5:91–93.
13. Diez-Roux AV, Nieto FJ, Caulfield L, Tyroler HA, Watson RL, Szklo M. Neighbourhood differences in diet: the Atherosclerosis Risk in Communities (ARIC) Study. *J Epidemiol Community Health*. 1999;53:55–63.
14. *Healthy People 2010: Conference Edition.* Washington, DC: US Dept of Health and Human Services; 2000.
15. National Cancer Institute. Geographic-based research in cancer control and epidemiology. Available at: http://www.grants.nih.gov/grants/guide/pa-files/PAS-00–120.html. Accessed July 14, 2000.
16. Thrall SE. Geographic information system (GIS) hardware and software. *J Public Health Manage Pract.* 1999;5:82–90.
17. Rushton G. Methods to evaluate geographic access to health services. *J Public Health Manage Pract.* 1999;5:93–100.
18. Tempalski B, Kreiswirth B, Alcabes P. The importance of error in geocoding: a tuberculosis case study. Available at: http://www.atsdr.cdc.gov/gis/conference98. Accessed September 25, 2000.
19. Krieger N, Zierler S, Hogan JW, et al. Geocoding and measurement of neighborhood socioeconomic position. In: Kawachi I, Berkman LF, eds. *Neighborhoods and Health*. New York, NY: Oxford University Press Inc. In press.
20. US Bureau of the Census. The census tract street locator. Available at: http://tier2.census.gov/ctsl/ctsl.htm. Accessed May 31, 2000.
21. *Census of Population and Housing, 1990: Summary Tape File 3 Technical Documentation*. Washington, DC: US Bureau of the Census; 1991.
22. Daly L. Simple SAS macros for the calculation of exact binomial and Poisson confidence limits. *Comput Biol Med.* 1992;22:351–361.