Using Address Histories in Health Research: Challenges and Recommendations for Research

Amy E. Hughes
U Texas at Dallas
U Texas Southwestern Medical Center
Amy.Speer@UTSouthwestern.edu

Sandi L. Pruitt
U Texas Southwestern Medical Center
Harold C. Simmons Comprehensive Cancer Center
Sandi.Pruitt@UTSouthwestern.edu

ABSTRACT
Longitudinal address histories are underutilized in Health GIS research; we describe some of the challenges limiting use of these data. We make suggestions for research efforts that can help other researchers access the information contained within address histories.

Categories and Subject Descriptors
H.2.8 [**Information Systems Applications**]: Database Applications—*Spatial databases and GIS*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Health, Medical information systems*

General Terms
Reliability, Standardization, Human Factors, Measurement

Keywords
Health, GIS, Mobility, Address history

1. BACKGROUND
Residential address serves as a record of exposure to infectious disease or environmental exposures in traditional spatial epidemiology, most famously through John Snow's application [1]. However, longitudinal residential address histories (hereafter address histories) can also serve as a longitudinal record of exposure to causal or confounding mechanisms (e.g. social factors) of other health outcomes such as latent or non-contagious disease [2]–[5]. As such, longitudinal address histories provide invaluable information for public health and GIS researchers across infectious and non-infectious health domains. However, although they represent rich data sources for health research, complete (i.e., lifetime) address histories are underutilized; researchers typically choose one address at a single time point as proxy for participant location. Address history data present a myriad of unique challenges for researchers. We review some of these challenges, including data availability, quality, access and security, geocoding, and analysis. We conclude with recommendations for future research necessary to facilitate the widespread, thoughtful use of address history data, which will improve our understanding of the complex relationships between geography and health.

HealthGIS'14, November 04-07 2014, Dallas/Fort Worth, TX, USA
Copyright 2014 ACM 978-1-4503-3136-4/14/11…$15.00
http://dx.doi.org/10.1145/2676629.2676638

2. CHALLENGES
Health GIS researchers may underutilize address histories because they present significant challenges requiring thoughtful, often highly complex, solutions. We briefly review several key challenges below.

2.1 Data Availability
Complete address histories, detailing location across the lifecourse, can be difficult to find and assemble; incomplete address histories (i.e., address history for the previous n years) become easier to collect as the number of spanning years decreases. Researchers may collect address histories as primary data, but this is uncommon, perhaps due to budget or time constraints. Researchers can also collect address histories via secondary, linked data sources using a number of approaches. We briefly describe challenges arising from each approach.

In primary data collection, researchers may retrospectively or prospectively collect address histories from participants. Retrospective primary collection of address histories requires a survey or interview setting; prospective primary collection of address histories requires repeated future contact with participants (e.g. in a prospective cohort study). The collection of a complete address history via interviews or surveys requires the design of a detailed process capable of capturing a varying number of addresses that cannot be over-written. The task of recording a complete address history may easily overwhelm research participants, and may require human or program interaction to walk through the steps of systematically recording an address history. This process can be both time- and labor-intensive for researchers. In addition, research participants may have trouble remembering correct addresses for all previous residences, or may feel negatively towards such a survey and refuse to participate for a number of reasons (e.g., privacy concerns and respondent burden). Recording address histories in a prospective study requires the ability to successfully contact each participant regularly, and ideally with every move; a daunting task given that contact information may change with each address change.

Researchers can also collect residential address histories via incidental or secondary sources, such as commercial sources (e.g., research companies like LexisNexis) or administrative datasets (e.g., Department of Motor Vehicles car registration records). The use of either presents notable disadvantages; purchasing complete address histories from commercial sources may be prohibitively expensive, while administrative records are not typically designed to be historical records and may not store address history data (i.e. data are over-written) [6]. Perhaps more importantly, use of secondary sources presents data integration challenges. Administrative or commercial organizations do not collect address histories for the purpose of research; the overlap between research study participants and secondary data source participants may be relatively small, restricting the number of data points in the final dataset. In other words, it may not be possible to locate secondary sources for all research participants. Merging primary research data with secondary address history data may result in differential match rates of participants based on another patient characteristic (i.e., the secondary address source contains 50% of participants with low socioeconomic status, and 95% of participants with high socioeconomic status). Finally, integration of secondary address histories requires the ability to link the two datasets, which introduces additional data availability and matching challenges because both datasets must contain a common unique identifier for all patients. Absence of a common unique identifier requires matching across multiple attributes (e.g., social security number, driver's license number, birthdate); some of these administrative data attributes may not be available in the research dataset. Challenges regarding the feasibility of matching secondary address history data sources to primary health data sources have been noted previously in the health GIS community [7], and unfortunately, remain too under-studied and unresolved.

2.2 Data Quality

Residential address histories suffer from quality issues. Data gathered retrospectively may suffer from recall bias. A research participant may not be able to remember (1) all the places he or she has lived in a given time period, (2) all the correct addresses corresponding to those places, or (3) how long the participant lived at each of those places. Time elapsed since the participant last remembered or used their address, the level of detail requested, and the total amount of information needed all affect recall bias [8]; therefore, recall bias in address histories compounds with longer address histories, more detailed or complete data collection, and with participants who move frequently. Address histories gathered through regular follow-up visits suffer from observation bias; a researcher cannot know if the participant has moved between observed addresses if the only two datum known to the researcher are address and associated date. This incomplete capture problem is of special concern when study participants have high level of residential mobility (e.g. when two moves occur within an observation interval). In this case, a regularly scheduled (i.e., yearly) data collection protocol fails to capture every within-interval move except the last. Secondary address history datasets suffer from this capture problem, and there is no way to check for complete capture of an address history without comparing the secondary data to additional sources of address history data. For example, address history obtained from a historical electronic medical record (EMR) may only include addresses obtained during incidental (i.e. opportunistic) clinic visits. Checking data accuracy and completeness would require parallel primary data collection (i.e. interviewing participants about address histories and comparing results to secondary data), for at least a subset of the participants. Such triangulation of address history data across multiple modes of data collection would require significant time and expense.

2.3 Data Security and Sensitivity

Address histories contain sensitive information. Many participants in an address history survey may be concerned about the confidentiality of these data. In a public health setting, address histories obtained from covered entities (e.g. health care systems) contain protected health information (PHI) because the addresses and some geocodes thereof, are considered patient identifiers. Health privacy laws [9] severely restrict access to such data, requiring researcher credentialing and IRB-approved data storage and transfer methods. Furthermore, participants of an address history survey may be reluctant to give address information for a variety of reasons, for example, concern about losing eligibility for programs dependent on residence. For example, a participant in an address history survey who utilizes a low-income health clinic restricted to county residents may neglect to report out-of-county addresses for fear of losing health clinic services.

2.4 Geocoding

Geocoding address histories presents a unique set of challenges. The accuracy of different geocoding schemes has been a frequent topic of research within the last 10 years. Positional error of geocoding schemes varies widely and can be dependent upon type of reference data selected [10]–[13]. Similarly, match rates vary with type of reference data; some reference data require an address to match exactly with reference data while others allow inference to dictate matches [13].

Geocoding of address histories sometimes requires the development of home-grown geocoding techniques, which present additional challenges including but not limited to the (a) choice of reference data; (b) collection of reference data; (c) prohibitively high cost of reference data; (d) details of fuzzy-matching algorithms used to create match scores for candidate addresses; (e) match score thresholds; and (f) match score ties, etc. Home-grown techniques have low start-up costs, but require significant coding and validation efforts [14].

Some popular pre-packaged geocoding schemes are not appropriate for use in studies aiming to conduct research at a given map resolution. The reference data scale affects results of analysis [15], and ZIP-code level geocoding does not often add the most useful information when performing analysis at the individual level [16]. Even employed at the correct resolution, some widely-available geocoding schemes introduce spatial or locational error into analyses through the use of algorithms that do not match addresses exactly [10]. These algorithms assign geographic coordinates based on the relative location of the participant's address compared to the location of beginning and ending address numbers of street blocks. In addition, commercially pre-packaged geocoding processes (e.g. ArcMap, Google Maps API) often store reference data on outside networks and/or use the internet to access the reference data and geocode the addresses. Because this process may introduce vulnerabilities to PHI, these processes are not always appropriate for health researchers given IRB-mandated data security and storage requirements.

Aside from home-grown address locator development, resolution mis-matches, and locational error, geocoding address histories presents the critically important and complex question of how best to incorporate time into the geocoding process. The geocoding process itself can incorporate time through date-conditional use of reference data from appropriate time periods, e.g., geocoding an address dated in 2009 to the 2009 parcel reference data. In this case, the researcher would account for the changing urban environment and economic landscape but would need to further account for spatial inconsistencies in the geocoded data across time periods. The process of associating a point (in this case, an address) to an area (in this case, a portion of street or a parcel) always requires the use of a centroid point. The location of the centroid changes with changes in boundaries of the polygon and changes in boundaries occur over time through changes to streets, parcels, etc. These boundary changes create spatial inconsistencies in the georeferenced data across time periods. Neglecting to incorporate time into the geocoding process itself offers the researcher the benefit of easy identification for moves across time periods, but the benefit comes at the cost of losing information about the changing urban environment (i.e., street, parcel, or other administrative boundaries) and economic landscape (i.e., housing values tied to parcel information, business developments).

2.5 Analysis

Analysis follows the successful geocoding of address histories. Challenges relating to analysis of address histories include characterizing addresses as moves or non-moves, managing the addresses that were not geocoded, and data interpretation.

The ability to determine moves or non-moves depends heavily on the researcher's decision to account for the changing urban structure during the geocoding process. If researchers choose to ignore the time dimension in the geocoding process, then this differentiation is straight-forward. Because a geocoded area is represented by a single set of coordinate points, researchers can simply look for changes in the assigned coordinate points of the addresses. However, if researchers choose to incorporate time into the geocoding process, the differentiation between moves and non-moves requires some type of processing of coordinate points (e.g., buffering) or of addresses (e.g., further fuzzy matching based on street address strings). For example, a researcher may choose to define non-movers through a small GIS buffer when using a cadastral (parcel) geocoding scheme in a county with relatively homogenous parcel sizes.

Address histories are longitudinal datasets, meaning that many addresses are associated with one participant in a time-ordered succession. Some of the addresses for a participant may geocode successfully while others may not, generating within-participant missing data. Along the same lines, there is no guarantee that an address history is complete; a participant could easily have resided at a place not recorded within the data. This is particularly true for secondary data address history sources. For example, if researchers collect address histories from the electronic medical record (EMR) via billing address confirmation at clinic check-in, there is significant potential for the EMR to omit an address from the record for a participant with extremely infrequent clinic visits. This situation generates missing data points for the patient; however, the researcher cannot know the extent of this type of missing data. Suppression of participants with missing data and restrictive assumptions, respectively, are two methods for dealing with these particular issues. The challenge of missing data, likely omnipresent in address histories, complicates the interpretation of address histories.

Interpretation of address histories is heavily dependent on which address is collected. Residential address histories are generally included in analyses to account for social, geographical environment, or built environment exposures, but a person does not stay at home at all times. In fact, many people spend most of their waking hours outside the home. Work addresses can capture similar exposures, but inclusion of work address history alone provides an incomplete picture of exposures as well [17]–[20], and furthermore, these data are not relevant for all populations (e.g. the elderly). When address histories are used as a marker of exposure, the type of address collected can affect how closely the address history reflects that exposure.

Data exploration of address histories also presents challenges. Understanding of the associations in spatial datasets usually begins with the creation of visualizations through the use of exploratory spatial data analysis. However, meaningful visualizations of address histories are rare and can be difficult to create. Address histories can represent large amounts of data, and visualizations quickly become overcrowded.

3. RECOMMENDATIONS

The Health GIS community needs a clear research agenda to better understand and overcome the challenges inherent in the use of address history data. Based on our experiences using a large database of address history data gleaned from electronic medical records (EMR), we have formulated the following 4 recommendations, designed to propel this research forward. We feel that these recommendations will help guide the field toward the development, implementation, and dissemination of best practices of address history data collection, management and quality control, geocoding, and analysis. We present the following recommendations aimed at overcoming challenges in these four areas:

3.1 Efficient Data Collection Methods

Development of efficient primary data-collection methods for address histories, with a focus on respondent burden. In a more broad sense, the practicality and utility of collecting address histories for use in health GIS is an issue previously singled out for research attention [7]. The call for research regarding feasibility of address history collection does not explicitly motivate the development of methods to collect the data efficiently. Development of methods to relieve survey-taker burden could include, for example, introduction of interactive map tools to help participants place previous addresses, comparisons of the success rates of different address history survey types, or development of easy and secure communication methods between researchers and participants for follow-up visits after an address change. Methods that relieve participant burden could address the challenge of data availability by simplifying the collection of address histories. The development of specific methods, like interactive map tools that help participants locate previous addresses and communication methods that allow participants to easily contact researchers after a move, could also address the data quality issues by eliminating incorrect addresses from an address history and limiting missing addresses.

Extending one suggested example [21] of an efficient data collection method, development of user-friendly interactive map tools, could greatly benefit health GIS researchers. A map tool used in a face-to-face survey environment could help a participant remember or infer the location of a previous address through the use of simple buffers; a participant could use the knowledge that his or her old house was close to a certain set of businesses or public amenities to narrow down and then locate his or her old house. This particular method, once validated, could be used to help participants record their address histories online at their leisure in their home, making data collection easier for both the participants and the researchers.

3.2 Best Practices for Secondary Data

Creation of best practices for use of secondary address history data, covering the collection of primary address history surveys for subsamples of the secondary data to ensure data quality, validity, and completeness. Two iterations of best practices for geocoding created and adopted by the North American Association of Central Cancer Registries (NAACR) provide extensive information regarding geocoding processes, but do not address validity concerns for secondary data sources [14], [22]; these guides offer an informed starting point for further development of best practices for secondary address data.

The development of these best practices can help researchers more easily assess, evaluate, and interpret the quality of secondary data. Furthermore, standards will facilitate researchers' capacity to evaluate the similarities, differences, and trade-offs between information gleaned from primary vs. secondary address history data collection.

3.3 Choices for Historic Reference Data

Development and dissemination of affordable, widely available quality reference data. A number of sources for reference data exist. Currently available sources offer the ability to geocode to different spatial resolutions, and data costs range from free to prohibitively expensive. Accurate updated commercial sources, like NAVTEC or TeleAtlas data, comprise the expensive end of the cost spectrum at minimum costs of $10,000; less accurate and sometimes less available government data, like TIGER line files and central appraisal district records comprise the free end of the cost spectrum [23].

The most widely-used reference data in the United States is the street network depicted in the US Census TIGER line files [14]. These files are now updated regularly, but the node topology of these files has been known to be inaccurate in some areas; OpenStreetMap [24] is an innovative grass-roots movement to catalog and fix these errors.

Historical reference data are not widely available. This data is most useful for studies spanning a timeframe that sees areas undergoing significant construction or natural disasters resulting in changes to the physical addresses within the data [14]. Currently, the best way to obtain historical data is for researchers to stockpile consecutive releases of data updates. However, this approach may not capture changes in physical construction but rather correction of previously released errors. Furthermore, collecting earlier versions of released and updated data can be difficult. Dissemination of this historical data can help researchers utilize the spatiotemporal information contained in an address history.

3.4 Quantification of Spatial Uncertainty

Development of methods to evaluate and improve the spatial uncertainty inevitably introduced through data processing, e.g., addresses recorded with typographical errors or incomplete geocoding, when interpreting address history data. Spatial uncertainty has been identified as a NIH priority, reflected in the recent Program Announcement (PA-11-238) that calls for innovative research identifying sources of spatial uncertainty in public health data. Efforts should be made to leverage this available funding to address spatial uncertainty in address history data.

Primary address history data is difficult to collect, secondary address history data is difficult to verify, and the geocoding process can introduce error into data analyses. Spatial uncertainty, coupled with these—and other—difficulties, can dissuade researchers from collecting these data. Without a clear understanding of the role of spatial uncertainty in analyses of georeferenced address histories, the benefits of having the data may not outweigh the data collection and management costs. Sources of spatial uncertainty and methods for adjusting for spatial uncertainty in address histories should be developed. Researchers will be more likely to justify additional costs required to analyze address histories given the availability of validated methods to evaluate and improve spatial certainty in these data.

Our recommendations are designed to move the field toward more widespread use of address histories in health research by fostering a greater understanding of the process of extracting robust geographical information from the address history. Our recommendations are aimed at enabling health researchers, focused on health problems and not trained in GIS, to quickly make sound decisions during this process, allowing for greater focus on the research question of interest. Increased utilization of address histories will inspire novel applications of address histories in health research. The body of literature at the intersection of health and GIS will grow in new, exciting directions through the development and refinement of applications of spatiotemporal methods to georeferenced longitudinal health data.

4. SUMMARY

The landscape in which research utilizing address histories is most useful—public health, or, more broadly, policy evaluation—is littered with studies bound by methodological limitations. In these fields, causal research must rely on observational studies that require identifying conditions or data. With the introduction of novel, previously hard-to-obtain spatiotemporal data, accurately geocoded and correctly interpreted address histories can dramatically improve the prospects of non-experimental causal analysis. The numerous complexities and challenges posed by address history data provide unique opportunities for thoughtful and substantive contributions to this causality conversation. Development of methods for handling such complexities and challenges, such as those we present here, will facilitate use of these data in new ways, resulting in better understanding of the complex relationships between geography and health.

5. ACKNOWLEDGMENTS

This work was supported by Cancer Prevention Research Institute of Texas (CPRIT R1208), National Cancer Institute (5U54CA163308), Agency for Healthcare Research and Quality R24 HS 22418-01, and National Center for Advancing Translational Sciences, UT Southwestern Center for Translational Medicine (U54 RFA-TR-12-006).

6. ADDITIONAL AUTHOR INFORMATION

Amy E. Hughes is a PhD student at The University of Texas at Dallas, GIS Department, 800 W. Campbell Rd GR2.822-7, Richardson, TX 75080-3021; and an intern at The University of Texas Southwestern Medical Center, Department of Clinical Sciences, 5323 Harry Hines Blvd E1.410D, Dallas, TX 75390-9169.

Sandi L. Pruitt is an Assistant Professor at the University of Texas Southwestern Medical Center, and a member of the Harold C. Simmons Comprehensive Cancer Center in Dallas, TX.

7. REFERENCES

[1] J. Snow, *On the mode of communication of cholera*. John Churchill, 1855.

[2] M. N. Bates, O. A. Rey, M. L. Biggs, C. Hopenhayn, L. E. Moore, D. Kalman, C. Steinmaus, and A. H. Smith, "Case-Control Study of Bladder Cancer and Exposure to Arsenic in Argentina," *Am. J. Epidemiol.*, vol. 159, no. 4, pp. 381–389, Feb. 2004.

[3] J. G. Brody, A. Aschengrau, W. McKelvey, R. A. Rudel, C. H. Swartz, and T. Kennedy, "Breast cancer risk and historical exposure to pesticides from wide-area applications assessed with GIS.," *Environ. Health Perspect.*, vol. 112, no. 8, pp. 889–897, Jun. 2004.

[4] J. Selander, M. E. Nilsson, G. Bluhm, M. Rosenlund, M. Lindqvist, G. Nise, and G. Pershagen, "Long-Term Exposure to Road Traffic Noise and Myocardial Infarction:," *Epidemiology*, vol. 20, no. 2, pp. 272–279, Mar. 2009.

[5] A. Wang, S. Costello, M. Cockburn, X. Zhang, J. Bronstein, and B. Ritz, "Parkinson's disease risk from ambient exposure to pesticides," *Eur. J. Epidemiol.*, vol. 26, no. 7, pp. 547–555, Jul. 2011.

[6] G. M. Jacquez, M. J. Slotnick, J. R. Meliker, G. AvRuskin, G. Copeland, and J. Nriagu, "Accuracy of Commercially Available Residential Histories for Epidemiologic Studies," *Am. J. Epidemiol.*, vol. 173, no. 2, pp. 236–243, Jan. 2011.

[7] L. W. Pickle, M. Szczur, D. R. Lewis, and D. G. Stinchcomb, "The crossroads of GIS and health information: a workshop on developing a research agenda to improve cancer control," *Int. J. Health Geogr.*, vol. 5, no. 1, p. 51, 2006.

[8] S. S. Coughlin, "Recall bias in epidemiologic studies," *J. Clin. Epidemiol.*, vol. 43, no. 1, pp. 87–91, 1990.

[9] *Health Insurance Portability and Accountability Act of 1996*. 1996.

[10] M. R. Bonner, D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim, "Positional Accuracy of Geocoded Addresses in Epidemiologic Research:," *Epidemiology*, vol. 14, no. 4, pp. 408–412, Jul. 2003.

[11] M. Schootman, D. A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo, and G. Higgs, "Positional Accuracy and Geographic Bias of Four Methods of Geocoding in Epidemiologic Research," *Ann. Epidemiol.*, vol. 17, no. 6, pp. 464–470, Jun. 2007.

[12] M. R. Cayo and T. O. Talbot, "Positional error in automated geocoding of residential addresses," *Int. J. Health Geogr.*, vol. 2, no. 1, p. 10, Dec. 2003.

[13] P. A. Zandbergen, T. C. Hart, K. E. Lenzer, and M. E. Camponovo, "Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets," *Spat. Spatio-Temporal Epidemiol.*, vol. 3, no. 1, pp. 69–82, Apr. 2012.

[14] D. Goldberg, "A Geocoding Best Practices Guide," North American Association of Central Cancer Registries, Springfield, IL, 2008.

[15] D. A. Griffith, M. Millones, M. Vincent, D. L. Johnson, and A. Hunt, "Impacts of positional error on spatial regression analysis: a case study of address locations in Syracuse, New York," *Trans. GIS*, vol. 11, no. 5, pp. 655–679, 2007.

[16] N. Krieger, P. Waterman, J. T. Chen, M.-J. Soobader, S. V. Subramanian, and R. Carson, "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*, vol. 92, no. 7, pp. 1100–1102, 2002.

[17] S. N. Zenk, A. J. Schulz, S. A. Matthews, A. Odoms-Young, J. Wilbur, L. Wegrzyn, K. Gibbs, C. Braunschweig, and C. Stokes, "Activity Space Environment and Dietary and Physical Activity Behaviors: A Pilot Study," *Health Place*, vol. 17, no. 5, pp. 1150–1161, Sep. 2011.

[18] J. Vallée, E. Cadot, C. Roustit, I. Parizot, and P. Chauvin, "The role of daily mobility in mental health inequalities: The interactive influence of activity space and neighbourhood of residence on depression," *Soc. Sci. Med.*, vol. 73, no. 8, pp. 1133–1144, Oct. 2011.

[19] Y. Kestens, A. Lebel, M. Daniel, M. Thériault, and R. Pampalon, "Using experienced activity spaces to measure foodscape exposure," *Health Place*, vol. 16, no. 6, pp. 1094–1103, 2010.

[20] S. Inagami, D. A. Cohen, and B. K. Finch, "Non-residential neighborhood exposures suppress neighborhood effects on self-rated health," *Soc. Sci. Med.*, vol. 65, no. 8, pp. 1779–1791, Oct. 2007.

[21] B. Chaix, Y. Kestens, C. Perchoux, N. Karusisi, J. Merlo, and K. Labadi, "An interactive mapping tool to assess individual mobility patterns in neighborhood studies," *Am. J. Prev. Med.*, vol. 43, no. 4, pp. 440–450, 2012.

[22] L. Wiggins, "Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices," North American Association of Central Cancer Registries, Springfield (IL), Oct. 2002.

[23] J. N. Swift, D. W. Goldberg, and J. P. Wilson, "Geocoding best practices: review of eight commonly used geocoding systems," *Los Angel. CA Univ. South. Calif. GIS Res. Lab.*, 2008.

[24] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Pervasive Comput. IEEE*, vol. 7, no. 4, pp. 12–18, 2008.