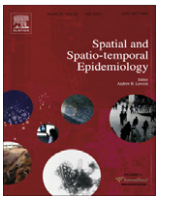
Spatial and Spatio-temporal Epidemiology 3 (2012) 7–16

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste



A research agenda: Does geocoding positional error matter in health GIS studies?

Geoffrey M. Jacquez *

BioMedware, Inc., 121 W, Washington, 4th Floor – TBC, Ann Arbor, MI 48104, USA

ARTICLE INFO

Article history:
Available online 14 February 2012

Keywords:
Geocoding error
Positional uncertainty
Disease clustering
Environmental exposures

ABSTRACT

Until recently, little attention has been paid to geocoding positional accuracy and its impacts on accessibility measures; estimates of disease rates; findings of disease clustering; spatial prediction and modeling of health outcomes; and estimates of individual exposures based on geographic proximity to pollutant and pathogen sources. It is now clear that positional errors can result in flawed findings and poor public health decisions. Yet the current state-of-practice is to ignore geocoding positional uncertainty, primarily because of a lack of theory, methods and tools for quantifying, modeling, and adjusting for geocoding positional errors in health analysis.

This paper proposes a research agenda to address this need. It summarizes the basics of the geocoding process, its assumptions, and empirical evidence describing the magnitude of geocoding positional error. An overview of the impacts of positional error in health analysis, including accessibility, disease clustering, exposure reconstruction, and spatial weights estimation is presented. The proposed research agenda addresses five key needs: (1) a lack of standardized, open-access geocoding resources for use in health research; (2) a lack of geocoding validation datasets that will allow the evaluation of alternative geocoding engines and procedures; (3) a lack of spatially explicit geocoding positional error models; (4) a lack of resources for assessing the sensitivity of spatial analysis results to geocoding positional error; (5) a lack of demonstration studies that illustrate the sensitivity of health policy decisions to geocoding positional error.

1. Introduction

"It is an unfortunate reality that even though a broad range of literature exists specifically geared to exposing how minor error in geocoding accuracy can affect results based on detailed spatial models, recent research initiatives continue to employ geocoded data without regard for how the accuracy can introduce possible inconsistencies or bias into the results" – (Goldberg et al., 2007).

Perhaps one of the foremost problems in measurement in the analysis of geographically referenced health data is that of geocoding positional error. Geographic models underpin concepts such as disease clustering, environmental exposure assessment, neighborhood context in health disparities analysis, accessibility to restaurants and parks in studies of overweight and obesity, and local availability to health care and screening facilities. But while geocoding – the process of converting text-based addresses into geographic coordinates – is a fundamental process in diverse disciplines including health (Boulos, 2004; Rushton et al., 2006), criminal justice (Zandbergen and Hart, 2009), political science (Haspel and Knotts, 2005) and computer science (Hutchinson and Veenendall, 2005), the sensitivity of spatial analysis results to positional error – the difference between a true location and that returned from the geocoded address – is not routinely addressed. In health analysis, recent studies have demonstrated that the strength of the odds relationship between disease exposures modeled at geocoded locations

* Tel.: +1 734 913 1098x203; fax: +1 734 913 2201.
E-mail address: jacquez@biomedware.com

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2012.02.002

declines with decreasing geocoding accuracy, and that "estimated measures of positional accuracy must be used in the interpretation of results of analyses that investigate relationships between health outcomes and exposures measured at residential locations" (Mazumdar et al., 2008). Yet the state-of-practice in health analysis is to ignore geocoding positional accuracy entirely.

The availability of georeferenced data in health analysis is expanding rapidly, due to several technological and policy trends. First, there is increased availability of user-generated, location-enabled health data as segments of the population become comfortable with sharing information through smart phones, web-browsers and other means; and as search engine keywords and social media are used to assess near real-time trends in health-related symptoms, medications, and outcomes (Ginsberg et al., 2009; Wilson and Brownstein, 2009; Seifter et al., 2010). The confluence of crowd sourcing (e.g. "reflexive consumerism," where patients review hospitals and professionals on the web) and volunteer geographic information (VGI, where individuals report activities at their location) is enabling significant advances in disaster response, epidemiology and exposure assessment science (Goodchild and Glennona, 2010; Adams, 2011). For example, by coupling technologies for near real-time sensing of pollutants with location-enabled devices such as mobile phones, VGI is being used to validate model-based high spatial resolution exposure estimates. This makes possible validation of individual-level exposure estimates as a person goes about their daily activities (Jacquez and Meliker, 2010; Stevens and D'Hondt, 2010).

Second, the US health care system and the Department of Health and Human Services are investing heavily in interoperable electronic health records expected to revolutionize health care and disease control and surveillance. Recent national legislation such as the Health Information Technology for Economic and Clinical Health (HITECH) Act and the Affordable Care Act (ACA) include provisions requiring the collection of detailed electronic data in standardized format for insurance and care equity purposes (Weissman and Hasnain-Wynia, 2011). Many of the data records for these systems include personal identifiers – names, addresses, and related health information – that can be used to construct georeferenced databases on patients, providers and health-related resources such as screening facilities.

Third, advances in spatio-temporal epidemiology facilitate reconstruction of geocoded residential histories of patients (Jacquez et al., 2011). The feasibility of developing reliable geospatial data retrospectively for large, epidemiological studies has been demonstrated, and revisiting completed studies using spatial epidemiological methods is now possible (Robinson et al., 2009). In an era of fiscal constraints expensive, large epidemiological studies are less likely to be funded. Application of spatiotemporal epidemiology to completed case-control, cohort and longitudinal studies holds enormous promise for gaining new insights into disease causation that leverages our nation's existing investments in health research.

Despite the burgeoning of georeferenced health data cited in the preceding paragraphs, positional uncertainty is rarely accounted for in geospatial health analysis, even though it can lead to erroneous results sufficient to lead to incorrect conclusions and flawed health policy decisions, as detailed in Section 3.

What is missing is a detailed understanding of empirical geocoding error distributions, theory underpinning the sources and propagation of such error through health decision making; models of positional error, and how it may be accounted for in geohealth analyses; tools for making such theory and models accessible to health practitioners; and resources such as databases for which empirical geocoding errors are known.

This paper proposes a research agenda to address these needs, and is organized as follows. Section 2 describes the basics of the geocoding process, the assumptions on which geocoding is based, and empirical data describing the magnitude of street geocoding positional error. This sets the stage for Section 3, which presents impacts of positional error in health analysis, including accessibility evaluation, disease clustering, exposure reconstruction, and spatial weights estimation. Section 4 details important knowledge gaps and proposes a research agenda for advancing our ability to make informed and accurate decisions using uncertain geospatial health data.

2. Geocoding process, assumptions, sources and magnitude of positional error

2.1. Overview of geocoding process

Geocoding is the process of taking address information and converting it into geographic coordinates useful for health analysis. Several approaches have been proposed, including deterministic and probabilistic address matching, among others. All involve the input of an address to be geocoded, normalization and standardization of that address into an acceptable format typically comprised of an address number, street name, city or town name, state and ZIP code, and an iterative comparison of that address to a reference data set (e.g. streets database) from which the geographic coordinates can be calculated. This calculation typically is by interpolation along a street segment for which the geographic coordinates of the beginning and end points are known, and/or areal interpolation within a parcel, ZIP code, or city polygon. Further details on the geocoding process are provided elsewhere (Goldberg, 2008).

When considering accuracy two aspects of the geocoding process are of interest: *completeness* (e.g. the proportion of addresses that successfully geocoded) and *positional accuracy* (e.g. how closely the geocoded coordinates correspond to the true coordinates). This paper is concerned with positional accuracy, and its impact on the results of geospatial health analyses.

2.2. Validity of geocoding assumptions and positional error

What assumptions of the geocoding process, when violated, introduce positional error? First, geocoding assumes that all of the addresses in the address range exist and can occupy space along the street segment (e.g. 600 through