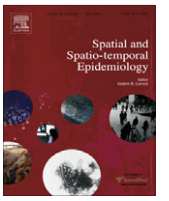
Spatial and Spatio-temporal Epidemiology 3 (2012) 39–54

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

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



The effect of administrative boundaries and geocoding error on cancer rates in California

Daniel W. Goldberg [a,*], Myles G. Cockburn [b]

[a] University of Southern California, Spatial Sciences Institute, Los Angeles, CA, USA
[b] University of Southern California, Department of Preventive Medicine, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Available online 10 February 2012

Keywords:
Geocoding
ZIP codes
Disease rates

ABSTRACT

Geocoding is often used to produce maps of disease rates from the diagnosis addresses of incident cases to assist with disease surveillance, prevention, and control. In this process, diagnosis addresses are converted into latitude/longitude pairs which are then aggregated to produce rates at varying geographic scales such as Census tracts, neighborhoods, cities, counties, and states. The specific techniques used within geocoding systems have an impact on where the output geocode is located and can therefore have an effect on the derivation of disease rates at different geographic aggregations. This paper investigates how county-level cancer rates are affected by the choice of interpolation method when case data are geocoded to the ZIP code level. Four commonly used areal unit interpolation techniques are applied and the output of each is used to compute crude county-level five-year incidence rates of all cancers in California. We found that the rates observed for 44 out of the 58 counties in California vary based on which interpolation method is used, with rates in some counties increasing by nearly 400% between interpolation methods.

1. Introduction

Geocoding, or the process of translating textual information most commonly in the form of postal addresses into geographic locations, is typically one of the first geocomputational processes applied to enable spatially-based health science investigations (Rushton et al., 2006; Nuckols et al., 2004). This process provides health scientists with the ability to place individuals and groups within a spatio-temporal context from which questions related to aspects such as geographic barriers to health services can be posed and investigated (Henry et al., 2011; Boscoe et al., 2011). The use of geocoding within the health sciences has a long history ranging from disease surveillance and outbreak monitoring (Krieger et al., 2002; Gumpertz et al., 2006; Abe and Stinchcomb, 2008; Bell et al., 2006; Boscoe et al., 2004; Boulos, 2004) to epidemiological investigations into the role and impacts that environmental exposures have on human health (Brody et al., 2004; McConnell et al., 2006; Zandbergen and Green, 2007; Reynolds et al., 2004; Rull et al., 2009; van Wiechen et al., 2004; Ferguson et al., 2004; Mazumdar et al., 2008).

Throughout this long history of using geocoding tools and geocoded data for health science research, numerous researchers have identified many limitations in using these data in scientific studies. These limitations are typically broken along two axes: (1) spatial accuracy of the geographic location computed for any particular subject – the distance between the output computed and the true location (Mazumdar et al., 2008; Zhan et al., 2006; Zandbergen, 2008; Whitsel et al., 2006; Ward et al., 2005; Schootman et al., 2007; Goldberg and Cockburn, 2010a; Fulcomer et al., 1989; Bonner et al., 2003); and (2) match rate achieved when geocoding a large set of records – the number of records capable of being assigned output geocodes (Zhan et al., 2006). Each of these issues has been investigated on numerous occasions with researchers consistently finding both to be non-randomly distributed

* Corresponding author.
E-mail addresses: dwgoldbe@usc.edu (D.W. Goldberg), myles@med.usc.edu (M.G. Cockburn).

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2012.02.005

across space, time, and population-specific characteristics. Ignoring either of these issues typically results in studies containing geographic bias, potentially invalidating the results (Zandbergen and Green, 2007; Schootman et al., 2007; Oliver et al., 2005; Krieger et al., 2002; Bichler and Balchak, 2007). The reason for this can be seen in the standard data pipeline utilized in scientific practices displayed in Fig. 1. The geocoding process and the geocoded data it produces are the underlying source upon which all subsequent analyses are performed and conclusions are drawn. Therefore, any errors in the production of these data are propagated throughout the rest of the scientific pipeline.

As indicated on several occasions by numerous authors, the internal intricacies of a geocoding system are known to greatly influence the quality of the results that can be obtained, affecting both spatial accuracy and match rates (Boscoe et al., 2004; Zhan et al., 2006; Whitsel et al., 2006; Ward et al., 2005; Schootman et al., 2007; Zandbergen, 2009; Wu et al., 2005; Gatrell, 1989). Most geocoding systems in use today for health science applications can be considered black boxes, where information or details describing the internal processing algorithms and/or data sources used in the process are not described. This most often stems from commercial reasons – to protect one's product line, ensure long term utilization of the system, and maintain a competitive advantage, it behooves a geocoding service or software provider to release none but the most general of details about the platform. In practice, this means that few details other than the reference data used are typically available.

This lack of technical detail and/or transparency of a vendor's geocoding process would not be a problem if geocoding was a simple, straightforward, error free process. However, this is simply not the case as evidenced by the broad range of academic disciplines that have contributed to our understanding and the development of geocoding techniques including but not limited to geography and geographic information science (Zandbergen, 2008; Wu et al., 2005; Tobler, 1972; O'Reagan and Saalfeld, 1987), computer science (Goldberg and Cockburn, 2010a; Bakshi et al., 2004; Goldberg et al., 2010; Goldberg and Cockburn, 2010b), and mathematics and statistics (Christen et al., 2004; Christen and Churches, 2005; Jaro, 1989, 1984). Published research reports describe several competing geocoding techniques (see Goldberg et al., 2007 for a review), and prior work has described how a one-size-fits-all geocoding approach is simply not appropriate (Zhan et al., 2006; Whitsel et al., 2006; Ward et al., 2005; Schootman et al., 2007). To begin, a single input data set may exhibit a variety of characteristics known to perform better with particular geocoding techniques due to the nature of geographic features and addressing systems in a region (Bonner et al., 2003; McElroy et al., 2003; Kravets and Hadden, 2007). For example, if a data set simultaneously contains both urban and rural records, records for urban addresses where parcels are small will perform better using parcel reference layers, while rural records may perform better using street centerline reference files because parcels are large (Schootman et al., 2007; Oliver et al., 2005; Krieger et al., 2002; Bichler and Balchak, 2007). Likewise, different techniques and/or reference data sources may be appropriate given the time period covered by an input data set to be geocoded, with historical records perhaps favoring historical reference data layers (Brody et al., 2004; Smith and Crane, 2001; Rull and Ritz, 2003; Rose et al., 2004; Kennedy et al., 2003; Brody et al., 2002). Finally, some techniques may simply be non-applicable for a particular region because the required data sets just do not exist (Zandbergen, 2008; Stage and von Meyer, 2005). For example, if a digital parcel file is not available in a specific county, parcel-level geocoding approaches are just not an option. Similarly, if a specific area of a rural county has not yet upgraded to be E-911 compliant, it may be the case that Rural Route addresses are the only street address-like information available, despite the well-known limitations of using these data for geocoding (Mazumdar et al., 2008; Goldberg et al., 2007; Zimmerman et al., 2007; Vieira et al., 2008; Cayo and Talbot, 2003).

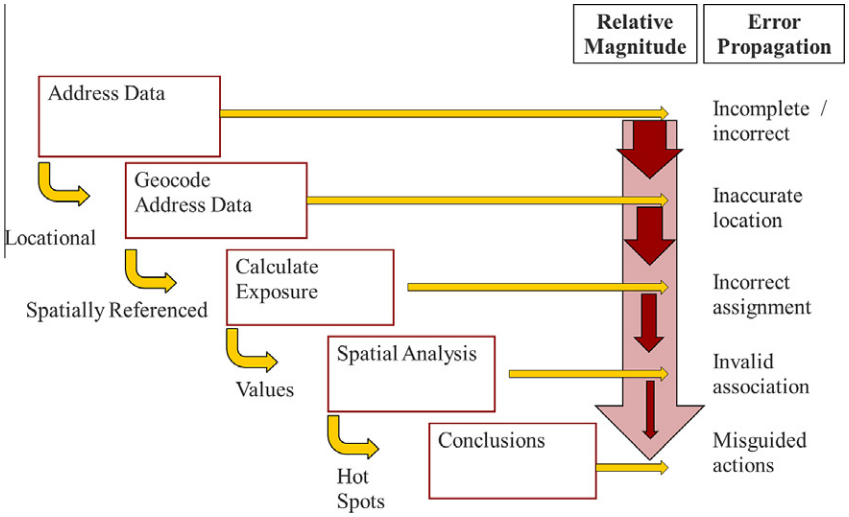


Fig. 1. Geocoded data pipeline and error propagation in environmental exposure studies.