Spatial and Spatio-temporal Epidemiology 3 (2012) 69–82

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets

P.A. Zandbergen [b,*], T.C. Hart [a], K.E. Lenzer [b], M.E. Camponovo [b]

[a] Department of Criminal Justice, University of Nevada, Las Vegas, NV, United States
[b] Department of Geography, University of New Mexico, NM, United States

ARTICLE INFO

Article history:
Available online 11 February 2012

Keywords:
Geocoding quality
Positional accuracy
Error propagation

ABSTRACT

The quality of geocoding has received substantial attention in recent years. A synthesis of published studies shows that the positional errors of street geocoding are somewhat unique relative to those of other types of spatial data: (1) the magnitude of error varies strongly across urban–rural gradients; (2) the direction of error is not uniform, but strongly associated with the properties of local street segments; (3) the distribution of errors does not follow a normal distribution, but is highly skewed and characterized by a substantial number of very large error values; and (4) the magnitude of error is spatially autocorrelated and is related to properties of the reference data. This makes it difficult to employ analytic approaches or Monte Carlo simulations for error propagation modeling because these rely on generalized statistical characteristics. The current paper describes an alternative empirical approach to error propagation modeling for geocoded data and illustrates its implementation using three different case-studies of geocoded individual-level datasets. The first case-study consists of determining the land cover categories associated with geocoded addresses using a point-in-raster overlay. The second case-study consists of a local hotspot characterization using kernel density analysis of geocoded addresses. The third case-study consists of a spatial data aggregation using enumeration areas of varying spatial resolution. For each case-study a high quality reference scenario based on address points forms the basis for the analysis, which is then compared to the result of various street geocoding techniques. Results show that the unique nature of the positional error of street geocoding introduces substantial noise in the result of spatial analysis, including a substantial amount of bias for some analysis scenarios. This confirms findings from earlier studies, but expands these to a wider range of analytical techniques.

1. Introduction

Geocoding is the process of assigning an XY coordinate pair to the description of a place by comparing the descriptive location-specific elements to those in reference data. The most common type of geocoding is address geocoding where the input consists of street addresses. The quality of geocoding has received substantial attention in recent years and it has been recognized that errors in geocoding need to be understood in order to determine the robustness of spatial analysis techniques which employ the results of geocoding. The general purpose of this paper is threefold: (1) to synthesize the existing knowledge on the nature of positional errors in geocoding; (2) to present a framework for modeling the effect of these errors on the spatial analysis of geocoded datasets; and (3) to present several case-studies that illustrate the implementation of this framework.

[*] Corresponding author. Address: Department of Geography, Bandelier West Room 111, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131, United States. Tel.: +1 505 277 3105.
E-mail address: zandberg@unm.edu (P.A. Zandbergen).

1877-5845/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2012.02.007

2. Background

2.1. Geocoding foundations

The geocoding process consists of translating an address entry, searching for the address in the reference data, and delivering the best candidate or candidates as a point feature on the map. Techniques involved in geocoding borrow from various academic fields, most notably, information theory, decision theory, probability theory, and phonetics. While geocoding applications are diverse and span many different fields, there are several common problems associated with geocoding that have traditionally caused poor match rates and positional error in the resulting spatial datasets (e.g., Rushton et al., 2006; Goldberg et al., 2007).

One of the main challenges to accurate geocoding is the availability of good reference data. This includes a set of geographic features to match against as well as robust address characteristics that enable matching address records to feature locations. This requires a sturdy address model to organize the reference data components. Several common address models exist, each with a particular set of supporting materials and characteristic errors. Commonly used address models include street networks, parcels, and address points which have been reviewed by Zandbergen (2008a, 2009). Street networks have historically been the most widely employed address data model, especially in the US. Address geocoding is accomplished by first matching the street name, then the segment that contains the house numbers and finally by placing a point along the segment based on linear interpolation within the range of house numbers. Many different reference datasets are available for this type of geocoding.

Geocoding against parcels makes it possible to match against individual plots of land (or rather, the centroids of those polygons) rather than interpolating against a street centerline. Parcel geocoding typically results in much lower match rates, but is now becoming more widespread given the development of parcel level databases by many jurisdictions in the US (Rushton et al., 2006). To overcome the limitations of parcels for geocoding, address points have emerged as an alternative address data model. Address points typically represent the locations of all addressable structures within a jurisdiction and are created from a combination of primary field data collection (GPS, field surveys) and secondary data interpretation (parcels, imagery, building footprints). In the US, address point geocoding is not yet in very widespread use. However, many local governments have started to create address point databases and several commercial geocoding firms provide address point geocoding for selected coverage areas.

2.2. Geocoding quality

A substantial body of literature has emerged on the quality of datasets obtained through address geocoding. The overall quality of any geocoding result can be characterized by the following components: completeness, positional accuracy, concordance with geographic units, and repeatability. Completeness is the percentage of records that can reliably be geocoded, also referred to as the match rate. Positional accuracy indicates how close each geocoded point is to the actual location of the structures of interest. Concordance is the degree to which geocoded locations are assigned to the correct geographic unit of interest. Repeatability indicates how sensitive the geocoding results are to variations in the reference data input, the matching algorithms of the geocoding software, and the skills and interpretation of the analyst.

The focus of the current paper is the positional accuracy of geocoded locations, defined as the Euclidean distance between the geocoded point location and the actual location of the structure associated with the address. Different components contribute to the error, including: (1) match to an incorrect street segment; (2) incorrect placement along the street segment; (3) incorrect offset from the street segment; and (4) positional error in the street segment. In most empirical studies, these components are not addressed separately and the measured error is therefore the aggregate effect of all four components. Several empirical studies in recent years have determined the positional accuracy of street geocoding, as reviewed by Zandbergen (2009) (Table 1). Despite differences in the design of the various studies, several general observations can be made as follows:

1. *The magnitude of positional errors varies strongly along urban–rural gradients.* Based on the review of published studies by Zandbergen (2009) using median values the "typical" positional error for residential addresses ranges from 2201 m. This is a very broad range and much of this can be attributed to differences across urban–rural gradients. For example, Cayo and Talbot (2003) found a median error of 38 m for urban areas, 78 m for suburban areas and 201 m for rural areas. Several other studies have found similar differences, confirming a clear general trend that geocoding is much more accurate in urban areas compared to rural areas.
2. *The distribution of the magnitude of positional errors of street geocoding does not follow a normal distribution.* Formal testing by Zandbergen (2008b) has shown that the distribution approximates a log-normal distribution when the distribution of the direction of errors is uniform. In a similar study, Zimmerman et al. (2007) have shown that mixtures of bivariate *t* distributions with two or three components are required to characterize the distribution of the magnitude of positional error when the distribution of the direction is strongly influenced by the gridded nature of the street network.
3. *The direction of positional errors of street geocoding* (*i.e.*, the angle in degrees of the line connecting the actual structure with the geocoded location) *is not random and is closely related to the local properties of the street network.* Studies to date reveal mixed results with several finding no significant difference from a uniform distribution (Cayo and Talbot, 2003; Strickland et al., 2007) while others finding a significant difference (Schootman et al., 2007; Zimmerman et al., 2007). However, aggregate statistics for direction ignore local