Spatial and Spatio-temporal Epidemiology 6 (2013) 7–16

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Random property allocation: A novel geographic imputation procedure based on a complete geocoded address file

Scott R. Walter [a,b,*], Nectarios Rose [a,1]

[a] Centre for Epidemiology and Evidence, New South Wales Ministry of Health, Sydney, Australia
[b] Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, University of New South Wales, Sydney, Australia

ARTICLE INFO

Article history:
Received 4 April 2012
Revised 26 February 2013
Accepted 17 April 2013
Available online 3 May 2013

Keywords:
Geocoding
Imputation
Spatial epidemiology
Australia

ABSTRACT

Allocating an incomplete address to randomly selected property coordinates within a locality, known as random property allocation, has many advantages over other geoimputation techniques. We compared the performance of random property allocation to four other methods under various conditions using a simulation approach. All methods performed well for large spatial units, but random property allocation was the least prone to bias and error under volatile scenarios with small units and low prevalence. Both its coordinate based approach as well as the random process of assignment contribute to its increased accuracy and reduced bias in many scenarios. Hence it is preferable to fixed or areal geoimputation for many epidemiological and surveillance applications.

1. Introduction

Geographical health research plays a key role in monitoring disease, informing public health policy and in understanding the epidemiology of disease. The use of geographic information systems (GIS) in public health research and surveillance is becoming more widespread with the increasing availability of geocoded health data. In order to draw valid inferences from geocoded data about spatial aspects of health it is important to have address information on each individual under study that is as accurate and complete as possible. Geocoding is the process of matching an address with a longitude and latitude. In practice, often a certain proportion of records will have incomplete address information and cannot be assigned accurately to an exact position in space (Nuckols et al., 2004; Oliver et al., 2005). This is particularly an issue in rural areas where many addresses do not conform to a standard format and the sparseness of dwellings can lead to greater positional error (Bonner et al., 2003; Cayo and Talbot, 2003). There may also be spatial, temporal (Goovaerts, 2012) or other factors predictive of address incompleteness that can create bias in analyses. Ignoring such data may lead to non-detection of outbreaks or genuine clusters in spatio-temporal coordinate-based surveillance. Omission of incomplete data will result in underestimation of areal counts and if the missingness is not at random, bias may be introduced which can affect areal surveillance and epidemiological methods such as point event and count-based models.

Rather than exclude records with missing address information, coordinates can be imputed at street, locality or postcode level depending on available information. If only property number is missing, coordinates can be assigned to the centre of the street with minimal impact on areal analyses. If, however, only locality information is present, a number of imputation methods exist for assigning incomplete addresses to either a point of longitude and latitude or to a spatial unit of interest. When using such geoimpu-

* Corresponding author. Address: Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, University of New South Wales, Level1, AGSM Building, Kensington, NSW 2052, Australia. Tel.: +61 (0) 2 9385 0510; fax: +61 (0) 2 9385 8280.

E-mail addresses: scott.walter@unsw.edu.au (S.R. Walter), nrose@doh.health.nsw.gov.au (N. Rose).

[1] Address: Centre for Epidemiology and Evidence, New South Wales Ministry of Health, LMB 961, North Sydney, NSW 2059, Australia. Tel.: +61 (0) 2 9391 9193; fax: +61 (0) 2 9391 9676.

1877-5845/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sste.2013.04.005

tation methods, it is difficult to quantify the extent to which results may be biased and how this bias may vary depending on the data completeness and the type of analysis or surveillance method employed.

Many epidemiological and surveillance methods are based on small-area counts of health outcomes. The estimation of small area rates, which is based on counts, provides a simple illustration of the bias that can result from the choice of geo-imputation method and is relatively straightforward to simulate, hence its use as the main outcome in this study. For such aggregation of data, obtaining unbiased areal counts is more important than imputing individual records to the correct area. Methods that assign all inexact addresses in a locality to a single area have been shown to be preferable in terms of individual-level accuracy, but can create artificial clustering (Hibbert et al., 2009). On the other hand, methods that assign incomplete addresses to multiple areas via some random process tend to better approximate the spatial distribution of disease at the expense of individual-level accuracy. There are many well known areal interpolation techniques (Flowerdew and Green, 1994; Goodchild and Lam, 1980; Gregory and Ell, 2005; Langford et al., 1991) as well as fixed methods assigning cases deterministically to a point such as a centroid, however, few methods impute coordinates via a random process.

Intuitively we expect that random methods should incorporate some measure of population density in order to approximate the spatial distribution of disease; the finer the resolution of population density information, the better the distributional approximation. The smallest spatial unit currently used in Australia for calculating populations covers about 200 households (Australian Bureau of Statistics, 2010), however, the availability of a complete geocoded address file allows incomplete addresses to be assigned to property coordinates that have been randomly selected from the address file. If one assumes the spatial distribution of properties closely mirrors the population distribution, then random assignment of incomplete addresses in this way accounts for the population distribution at the finest possible spatial resolution. We refer to this method as random property allocation.

The aim of the study was to compare the performance of this novel random allocation technique with four common geoimputation methods using small area rates as the outcome of interest. Through the use of simulation, these methods were compared for a range of disease prevalence values, spatial unit sizes and proportions of address incompleteness in order to provide evidence of an optimal choice of geoimputation method. In addition, this study demonstrates how a complete geocoded address file can be used to simulate spatial aspects of epidemiological scenarios.

2. Methods

2.1. Data sources

The Geocoded National Address File (G-NAF) is an index of all Australian property addresses and their corresponding longitude and latitude coordinates derived from government land records, as well as postal and electoral address data (MapData Sciences; PSMA Australia). There are close to 3.6 million individual addresses listed as being in the state of New South Wales (NSW). Boundary files from the Australian Bureau of Statistics (ABS) were used to assign addresses in the G-NAF to three sizes of geographical unit: 2006 Collection Districts (CD), 2007 Statistical Local Areas (SLA) and 2005 Area Health Services (AHS) (Fig. 1). The ABS has defined a hierarchy of spatial units, including CDs and SLAs, for use in census administration (ABS, 2001). There are about 12,000 CDs in NSW ranging in area from 0.002 to 14,000 square kilometres, and there are about 200 SLAs ranging in area from 4 to 93,000 square kilometres. The eight AHSs are administrative regions defined by the NSW Department of Health and based on 2005 units of census geography. For the sake of simplicity, CDs, SLAs and AHSs will be hereafter referred to as small, medium and large spatial units, respectively. These size-related terms also reflect the area of these units relative to the localities that form the basis of the imputation. In general, within each spatial unit type, the size of individual areas tends to be approximately proportional to remoteness.

This analysis assumes that incomplete addresses only have locality information from which to impute. There are just over 5,000 localities in NSW ranging in size from 0.003 to almost 18,000 square kilometres (Fig. 1). In general, localities are not nested neatly within the administrative spatial units described, nor vice versa.

2.2. Geoimputation methods

Random property allocation and four comparison geoimputation methods were used to assign incomplete addresses to each of the three sizes of spatial unit described above. Random property allocation, devised by Nectarios Rose, assigns each incomplete address to the coordinates of a property centroid within a given locality that has been randomly selected from properties in a complete geocoded address file. The random selection allows each address within a locality to have equal probability of selection, and multiple records with inexact addresses could be randomly allocated to the same property. Each imputed coordinate can then be assigned to its corresponding area of interest.

A simple alternative method for imputing coordinates is to assign all incomplete addresses to the locality centroid. We consider two definitions of the centroid in this study. The geographic centroid can be conceptualised as the centre of mass of the two dimensional polygon that represents a locality. Geographic locality centroids were generated using ArcGIS software (ESRI Inc., 2010). The population weighted centroid is a weighted mean of longitude and latitude coordinate values for all properties in a given locality, with each property also weighted by the number of people per household. Whilst the exact number of residents in each property is not known, it is possible to assign the average number of residents per property at census district level. Case coordinates imputed by either method of local-