Applied Geography 58 (2015) 41–47

Contents lists available at ScienceDirect

Applied Geography

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

Evaluating geo-located Twitter data as a control layer for areal interpolation of population

Jie Lin [a, *], Robert G. Cromley [b]

[a] Department of Earth Sciences, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang 310027, PR China
[b] Department of Geography, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Available online

Keywords:
Areal interpolation
Geo-located Twitter
Remotely sensed imagery
Volunteered geographic information

ABSTRACT

Control data are critical for improving areal interpolation results. Remotely sensed imagery, road network, and parcels are the three most commonly used ancillary data for areal interpolation of population. Meanwhile, the open access geographic data generated by social networks is emerging as an alternative control data that can be related to the distribution of population. This study evaluates the effectiveness of geo-located night-time tweets data as ancillary information and its combination with the three commonly used ancillary datasets in intelligent areal interpolation. Due to the skewed Twitter user age, the other purpose of this study is to test the effect of age bias control data on estimation of different age group populations. Results suggest that geo-located tweets as single control data does not perform as well as the three other control layers for total population and all age-specific population groups. However, the noticeable enhancement effect of Twitter data on other control data, especially for age groups with a high percentage of Twitter users, suggests that it helps to better reflect population distribution by increasing variation in densities within a residential area delineated by other control data.

Introduction

In many applied studies it is often necessary to estimate population counts for different types of area delineations. For example, risk assessment requires that the population located in a risk zone such as a toxic plume or a potential storm surge area. Population counts are available from the US census for different census administrative units but not for the multitude of potential area units (e.g., service delivery areas, hydrologic study areas, zones created by analytical tools of a GIS) that geographers and other scientists might encounter. To solve this estimation problem, different types of areal interpolation methods have been proposed. Researchers have continued to develop new procedures and strategies for improving areal interpolation results over the past few decades (Qiu & Cromley, 2013). These efforts focus not only on developing better analytical methods but also exploring better control data for transferring attribute data collected for one areal delineation of geographic space (the set of source zones) to another (the set of target zones). Control data, which are spatially correlated with the distribution of the attributes being estimated in areal interpolation, are used to improve the quality of the estimation. Over time the number and variety of control layers used in areal interpolation has increased as new and different technologies have become available. The purpose here is to evaluate another new technology, geo-located Twitter, as a potential control layer in areal interpolation.

Twitter, since its launch in 2006, has experienced an exponential growth rate, and its popularity continues to grow (Leetaru, Wang, Cao, Padmanabhan, & Shook, 2013). In addition, Twitter began supporting tweet-based location in August 2009. Unlike user-based location, which is a city or neighborhood manually selected by a user from predefined list of locations supported by Twitter.com and stored in the user's profile, this tweet-based location is derived from an imbedded GPS in a mobile device or triangulation from cell or Wi-Fi signals without users' manual interruptions. Each geo-tagged tweet has a latitude and longitude indicating the location where the tweet was created. These tweet-based locations provide the user's current location to four decimals, meaning they can capture a precise street address such as a residential house. Thus this spatial information attached to each geo-tagged tweet, especially those sent during the night, offers potential control data for the estimation of the spatial distribution of population.

* Corresponding author. Tel.: +86 571 87952453.
E-mail address: jielin@zju.edu.cn (J. Lin).

http://dx.doi.org/10.1016/j.apgeog.2015.01.006
0143-6228/

The rest of this paper is organized as follows: The Nature of Control Data section discusses the different nature and spatial dimensionality of control data that has been used in areal interpolation methods. The Study Area and Data section briefly describes the study area and control data involved in this research. The Research Design section describes how the geo-located tweets data set and its combination with other data sets were used as control data to formalize the solution of areal interpolation of population. The Comparison of Results section presents and compares the areal interpolation results derived from various control variables for different age group populations. Finally, conclusions and future work are discussed in the Conclusions section.

The nature of control data

Spatially, control data can be divided into two-dimensional control zones, one-dimensional control lines, and zero-dimensional control points (Zhang & Qiu, 2011). Remotely sensed imagery is the most commonly used two-dimensional ancillary data for population distribution modeling. Lo (1986) categorized and discussed four different population estimation models based on how remotely sensed imagery is used: measurements of built-up urban areas, dwelling units-counting, areal measurement of different land cover categories, and direct modeling. The methods in the first category use allometric growth models to describe the relationship between population and urban size. Lo and Welch (1977) estimated population of selected Chinese cities from their built-up area derived from Landsat images. Other reported population related research belong to this category include using low-resolution nighttime satellite imagery provided by the Defense Meteorological Satellite Program's Operational Linescan System (DMSP/OLS) (see Lo, 2001; Sutton, 1997; Zhuo et al., 2009) and synthetic aperture radar (SAR) (Henderson & Xia, 1997).

The methods in the second category first identify residential-building footprints from high resolution remotely sensed imagery (Webster, 1996) or analogue aerial photography (Wu, Wang, & Qiu, 2008), and then multiply the area of delineated residential building by the number of occupants per area unit derived from statistical model to produce a population estimate (Lu, Im, Quackenbush, & Halligan, 2010). Recently, area has been replaced by the volume of a residential building with building height derived from Light Detection and Ranging imagery (LiDAR) in population estimation (Lu et al., 2010; Sridharan & Qiu, 2013). The volume-based method eases the heterogeneous relationship between population and delineated residential buildings due to the high-rise nature of urban areas.

The third category includes intelligent methods that integrate land cover data in the population estimation, ranging from a simple binary dasymetric method (Fisher & Langford, 1996; Flowerdew & Green, 1989; Holt, Lo, & Hodler, 2004) to more sophisticated procedures such as those that are logical extensions of Wright's (1936) original dasymetric mapping (see Eicher & Brewer, 2001; Langford, 2006; Mennis, 2003; Mennis & Hultgren, 2006; Reibel & Agrawal, 2007) and various forms of statistical analysis (Cromley, Hanink, & Bentley, 2012; Langford, Maguire, & Unwin, 1991; Lin, Cromley, & Zhang, 2011; Lo, 2008; Schroeder and Van Riper, 2013; Yuan, Smith, & Limp, 1997). An alternative approach in this category did not use the land cover categories as the correlate but instead integrated an impervious surface fraction derived from Thematic Mapper (TM) imagery into a cokriging method to interpolate population density using the spatial correlation and cross-correlation between population and this fraction (Wu & Murray, 2005).

Direct modeling approaches in the fourth category estimate the population distribution within a specified automated digital image analysis framework. Iisaka and Hegedus (1982) used a multiple regression model to predict population count with the mean radiances of values on each of the four Landsat Multispectral Scanner (MSS) as explanatory variables. Li and Weng (2005) used stepwise regression to develop models for estimating population density in Indianapolis, Indiana with spectral signatures, principal components, vegetation indices, fraction images, textures, and temperature derived from Landsat ETM+ as predictive indicators. Liu, Clarke, and Herold (2006) used Ikonos satellite images to estimate urban population distribution based on image texture within a linear regression framework. However, their results suggested that image texture was not good enough for predicting urban population distribution.

Because remotely sensed imagery can be easily integrated into any geographic information system (GIS), its use as control data for improving population estimation accuracy will continue. However, Sadahiro (2000) argued that remotely sensed data are not always readily available or are expensive. Furthermore, pre-classifying these imagery data require an understanding of multispectral signatures and image classification techniques, which may be beyond the scope of many GIS analysts (Langford, 2013). Although pre-classified land use/land cover data for the United States is freely available from Multi-Resolution Land Characteristics Consortium (MRLC) (http://www.mrlc.gov/) and updated every five years, it provides only an Anderson Level I-like classification, which distinguishes only among the broadest land cover types; thus, it has some inadequacies (Lin, Cromley, Civco, Hanink, & Zhang, 2013).

An alternative two-dimensional control zone layer is road buffer areas (Mrozinski & Cromley, 1999). Later, Langford (2007) facilitated dasymetric-based population interpolation by extracting buildings through the identification of specific color indices within raster scan maps to construct population control zones. Maantay, Maroko, and Herrmann (2007) and Tapp (2010) also have used digitized cadastral units to delineate the two-dimensional control zones. These different controls have also been used to enhance remote sensing-derived land cover data by incorporating local road buffer or parcel data layers in a multi-layer, multi-class dasymetric framework (Lin et al., 2013; Su, Lin, Hsieh, Tsai, & Lin, 2010). Goodchild, Anselin, and Deichmann (1993) demonstrated a framework to subjectively and interactively design control zones with homogeneous population density based on prior knowledge in the absence of ancillary data, followed by estimating densities with various statistical methods, the target zone population can then be derived by integrating control zones intersected with it. The results illustrated the main purpose of their research – that a substantial improvement in areal interpolation results can be achieved by a small amount subjective information given by a user who is familiar with the study area.

With respect to one-dimensional ancillary data, road networks are the main source of control information. Xie (1995), for example, used network information to allocate population in target zones using the length of road, a road's feature class code, or the number of houses associated with the line segments of a road network as weights. Of these weights, the one based on the feature class code provided the best estimation. A more recent work that used street network data as control lines was conducted by Reibel and Bufalino (2005) to interpolate population in Los Angeles, California.

Noting that the two-dimensional control zones-based and one-dimensional control lines-based intelligent approaches usually requires complex topological overlay operations, which increases computational burden, especially when large quantity of data need to be processed, Zhang and Qiu (2011) employed school locations as control points in a point-based method for areal interpolation. Their method assumes that schools tend to locate at an approximate population center to best serve the residents around them,