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 14 February 2015

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,

thus a weighting surface is constructed by using a distance decay function based on these control points; then the population in each source zone is disaggregated into each cell within that source zone based on the ratio of the weighting value of that cell to the sum of weighting values of that source zone. Their model is pycnophylactic by design as the population in each source zone is kept constant during the interpolation procedures. Zhang and Qiu (2011) concluded that their intelligent, point-based areal interpolation with a specific distance decay coefficient outperformed the dasymetric method using land use as ancillary data and provided comparable results as the network length method. However, Langford (2013) found different results when he adopted the point-based method to interpolate population in the city of Cardiff. The point-based method was outperformed by most intelligent methods for areal interpolation applied across two spatial resolutions in his study. Thus, Langford (2013) pointed out that the choice of the point ancillary data and mathematical distance decay functions should be carefully selected with respect to the population distribution of the study area. He further tested the point-based method using bus stops rather than school locations as control points, and found better results than when using school locations; however these results were still not as good as for the other intelligent methods. More recently, Bakillah, Liang, Mobasheri, Jokar Arsanjani, and Zipf (2014) proposed a new framework using the pre-classified land use land cover categories and OpenStreetMap points-of-interest successively. In their method, the census population was first disaggregated from 7 districts in the city of Hamburg, Germany to each land use and land cover category. The next step of their study further disaggregated the population within each land use land cover category to each cell using Zhang and Qiu's (2011) method with selected OpenStreetMap points-of-interest as control points. Their results indicate that this increased the accuracy when compared against interpolation procedures that simply use the same OpenStreetMap points-of-interest as control points.

Usage of school locations, bus stops, and selected OpenStreetMap points-of-interests as control points in population distribution modeling has one logical pitfall. Population is actually located near these control points but not at them; however, cells containing the control points receive the most population, which tends to propagate errors, especially for small area population estimation. In contrast, this research evaluates a new form of zero-dimensional control points, geo-located Twitter data, for use in areal interpolation. The assumption in this study is that geo-located twitter messages (tweets) sent out during the night indicate where people actually live and the population density is more scattered in areas where no tweets are sent out during the night.

Study area and data

The study area is a nine-town region within Hartford County, Connecticut. The region is characterized by different types of land use with various population densities, such as medium urban fabric in Manchester, dense urban center within Harford, and forest areas at the periphery. According to the 2010 census, the region had a total population of 396,435. The nested hierarchy of census tract and block groups are used as source and target zones, respectively. The geography files were downloaded from the US Census Bureau's 2010 Topologically Integrated Geographic Encoding and Reference (TIGER/Line) Shapefiles Main Page (http://www.census.gov/geo/www/tiger/tgrshp2010/tgrshp2010.html). While the attribute population data were extracted from 2010 Census Summary File 1 (SF1) Table P012, and downloaded through the US Census Bureau's web data retrieval interface, American FactFinder (http://factfinder2.census.gov). The raw tabulation data for the counts of total population and the 46 age-sex subgroups (23 age groups for each gender) at source tract and target block group levels were aggregated to five groups: total population, 18–29 years, 30–49 years, 50–64 years and 65 years and over. The geography boundary files and the refined demographic data were then joined in ArcGIS 10.2.

The geo-located tweets were captured in near real-time using the streamR package (Barbera, 2014) in R environment. Based on the previous research, only about 1% of all tweets are geo-tagged (Crampton et al., 2013). Due to the small fraction of geo-located tweets, the raw Twitter dataset consists of 17 days of geo-located tweets in Connecticut within two time intervals, One from March 6th to March 13th 2014 and the other from April 8th to April 17th 2014, in order to identify as many residential places as possible, but also to mitigate bias of tweets sent during each day. Because the main purpose of this research is to use the locations of geo-located tweets as control points to interpolate nighttime residential population, we first selected the geo-located tweets that were sent between 6 PM and 8 AM from the raw Twitter dataset. We also observed that the data contained a considerable number of active users who sent more than one tweet in nearby locations. Only one of these nearby tweets was kept and all others within a three hundred meters radius were eliminated. The remaining 5177 night geo-located tweets were used as control points to model the population distribution. In order to test the age bias among users, total population was estimated first; then age-specific populations were estimated to determine if the control twitter data were more spatially correlated with the distribution of certain age groups of the population. Finally, twitter data were combined together with other ancillary data to test whether the integrated control data could further improve the accuracy of interpolation results.

Other population control data used in this research include remotely sensed land cover data, road buffers, and residential parcels. The remotely sensed land cover data set was classified by the Center for Land Use Education and Research (CLEAR) at the University of Connecticut at a 30 × 30 m spatial resolution based on Landsat Thematic Mapper satellite imagery acquired *circa* 2006. The data were downloaded from CLEAR's website (http://clear.uconn.edu). The “Developed” category, which contains most residential areas, was extracted and clipped to the study area as the final control data. The road networks were acquired from the US Census Bureau's 2010 TIGER/Line Shapefiles Main page (http://www.census.gov/cgi-bin/geo/shapefiles2010/main) and clipped to the study area, and secondary roads, local neighborhood roads, rural roads, and city streets were selected from the original road networks. Buffer areas were constructed around these roads as a constraint on the population distribution. The parcel data were obtained from the Connecticut Capitol Region Council of Governments (CRCOG). This data layer was also first clipped to the study area, and then 94,235 total residential parcels were extracted as the populated control zones. Finally, all four control data sets and the source and target zones layers were registered to the Connecticut State Plane Coordinate System (NAD83) for further analysis.

Research design

In the first set of analyses, the Twitter data set, remotely sensed developed land cover, road networks, and parcels are used separately as control data to estimate population counts from the source tract level to target block groups level. For the twitter data set, two different weighting surfaces are constructed first using locations of all geo-tagged nighttime tweets as control points. The two distance decay functions used to produce the surfaces are as follows:

$$W_{si} = \left(1 - \frac{\lambda_{si}}{\lambda_{s\ \max}}\right)^{q} \quad (1)$$

$$W_{si} = e^{\left(1-\frac{\lambda_{si}}{\lambda_{s\ \max}}\right)^{q}} \quad (2)$$

where λ_{si} is the distance of cell i in source zone s to the nearest control point, $\lambda_{s\ \max}$ is the maximum value of λ_{si} for all cells within source zone s, and q is the decay parameter that controls the distance decay degree. Equation (1) is the linear distance decay function while Equation (2) is the nonlinear one. Then the populations in each source tract zone are disaggregated to grid cells in that source zone using the following equation:

$$\widehat{D}_{si} = \frac{P_s}{\sum_{i=1}^{N_s} W_{si}} \times W_{si} \quad (3)$$

where $\widehat{D}_{si}$ is the estimated density value for cell i in source zone s, P_s is the population count in source zone s, N_s is the cell count in source zone s, W_{si} is the weighting value for cell i in source zone s. Finally the estimated population in each target block group is obtained by aggregating all cell values in that zone. For the remotely sensed data, the developed areas are integrated into a binary dasymetric model as the control variable to estimate the population count of block groups from census tract counts using the equation:

$$\widehat{P}_t = \frac{P_s}{A_{sc}} \times A_{tc} \quad (4)$$

where $\widehat{P}_t$ is the estimated population count in target zone t, P_s is the observed population count of the source zone s that the target zone nests within, A_{sc} is the intersection area between source zone and control zone, and A_{tc} is the intersection area between the target zone and the control zone. The road networks and parcel data sets are used in the similar manner as the remotely sensed data with road buffer areas and residential parcels as control data in each binary dasymetric model.

In the second analysis, every two of the four control datasets are combined together to infer the population count in target block groups. There are six models in this set of areal interpolations: Twitter and land cover data, Twitter and road networks, Twitter and parcel data, land cover and road networks, land cover and parcel data, and road networks and parcel data. For the areal interpolation models that used twitter data in conjunction with another data set, only the grid cells of the weighting surface that spatially intersected with the other control data were selected to calculate the population density value in Equation (3). For other methods in this set of analysis, the intersection areas of the two control data are used as the new control variable to calculate the population count in each target zone using the Equation (4).

The third and fourth analyses followed this pattern. In the third analysis, every three of the four control data are used together for the areal interpolation procedures resulting in four different combinations of the four control data sets, and in the fourth analysis all four control data sets were used all at once to calculate the population counts in target zones. To evaluate the use of the various control data layers for the interpolation of different population subgroups, total population was estimated first, and then the population for each subgroup was estimated. This sequence was repeated for each of the four different sets of analysis.

Comparison of results

The ground truth against which performances of areal interpolation methods will be evaluated is the population counts at block group level. Performances are measured by two relative errors, adjusted root mean square error (Adj-RMSE) and adjusted mean absolute error (Adj-MAE), as described by Lin et al. (2013). Absolute errors, such as root mean square error (RMSE) and mean absolute error (MAE), are more affected by the values of an attribute variable, making them less useful for comparing results of different interpolated variables, especially when magnitude changes are involved (e.g., population between 18 and 29 years are part of total population, resulting in lower RMSE and MAE values from this factor alone). The two relative errors, Adj-RMSE and Adj-MAE, are computed respectively as the RMSE score scaled by each known target zone population and the MAE scores scaled by the same target zone population. These values are less affected by the magnitude of interpolated variable and are more suited for cross-variable comparisons. The difference between the Adj-RMSE and Adj-MAE is that the former considers the variance of error magnitude while the latter considers the absolute value of that error. For those procedures using road buffers and/or geo-located tweets, both variables are kept constant with a distance decay parameter q equals to two and a road buffer distance equals to 100 feet.

Table 1 compares results of total population interpolation derived from various sets of control variables. Areal interpolation with geo-located tweets (GT) performs the worst, with an Adj-RMSE value equaling 0.501 and an Adj-MAE score of 0.293, while the combination of only developed land (DL), geo-located tweets, and residential parcels (RP) provides the best results, with an Adj-RMSE score of 0.245 and an Adj-MAE value of 0.168. Usually a combination of control data layers results in increased accuracy (e.g. combination of DL and GT has a better result than that derived from either DL only or GT only); however, that is not true for all cases. Although GT data alone produces the worst areal interpolation results, it is the one of best enhancements, because all areal interpolations perform better when GT was added to others to make an integrated control layer. RP is the other control that has the same enhancement effect. Alternatively, RB is the worst enhancement, as the areal interpolation results deteriorate sharply when it is added. When RB is added to DL and RP, GT and RP, and DL, GT, and RP, respectively, the Adj-RMSE value for each situation increases from 0.273 to 0.314, from 0.279 to 0.281, and from 0.245 to 0.290, respectively. A clue as to the most likely explanation is that RP may be conflict with RB when DL, GT or both are already included in the control data. DL as an enhancement returns modest

Table 1
Errors for total population estimation with different control data sets.

Control data	Adj-RMSE	Adj-MAE
DL	0.467	0.251
GT	0.501	0.293
RB	0.438	0.249
RP	0.336	0.234
DL, GT	0.460*	0.243*
DL, RB	0.397	0.212
DL, RP	0.273	0.182
GT, RB	0.393	0.220
GT, RP	0.279	0.206
RB, RP	0.303	0.206
DL, GT, RB	0.383*	0.201*
DL, GT, RP	0.245*	0.168*
DL, RB, RP	0.314	0.201
GT, RB, RP	0.281	0.200
DL, GT, RB, RP	0.290*	0.188*

Note: Adj-RMSE = adjusted root mean square error; Adj-MAE = adjusted mean square error; DL = developed land; GT = geo-located tweets; RB = road buffers; RP = residential parcels. Both linear and nonlinear distance decay functions were evaluated for geo-located tweets, however, only the function with the better results are presented in the table. Numbers with * on the up right corner are derived from non-linear distance decay function.

levels of improvement, and when DL is added to RB and RP, and RB, RP and GT, the Adj-RMSE values for each situation increases from 0.303 to 0.314, and from 0.281 to 0.290 respectively.

One of the main purposes of this paper is to investigate the effect of age bias among Twitter users on the interpolation results. Table 2 shows the errors for interpolation of each age group population with different control data sets. The age-specific population 18–29 years always has the largest error values compared against the other age groups whatever control layer is used. Rather than comparing the errors associated with the different age groups, we therefore compare the decreasing percentage of errors when adding the GT data as an enhancement to other control layers for each age group in order to examine the age bias effects on areal interpolation results. When GT was added to DL, RB, and RP respectively to make an integrated control data with two variables, the Adj-RMSE values decreased for all subgroups except those aged 65 and over (see Table 2). When GT was added to DL and RB, DL and RP, and RB and RP to make an integrated control data with three variables, the same patterns was found for Adj-RMSE values. When GT was added to DL, RB, and RP to make an integrated control data with all four variables, the Adj-RMSE values decreased 13.99% for those aged 18–29 years, 12.15% for those aged 30–49, 1.65% for the age group 50–64 years, but again increased for those 65 years and over. Twitter adoption by internet users ages 18–29 years was 31%, 30–49 years was 19%, 50–64 years was 9%, and 65 years and over was 5% according to Pew Research Center's Internet Project August Tracking Survey in 2013 (Duggan & Smith, 2014). Furthermore, the fact that internet users are also skewed towards a younger population makes the percentage of the subgroup populations 18–29 years and 30–49 years even more higher than that of the subgroup populations 50–64 years and 65 years and over. Thus the enhancement effect of GT for each subgroup population is consistent with the user percentage of that age group, which means that a higher percentage of Twitter users usually leads to better enhancement of GT for that age subgroup.

Fig. 1 shows the spatial distribution of adjusted absolute error for interpolation of each group population with GT, RP, and combination of GT and RP as control data respectively. The reason for choosing these three control datasets is that GT is the main focus of this paper, RP is the best single control dataset, the combination of GT and RP shows the enhancement effect of GT on RP at local scale. Overall, the error pattern at the target zone level is consistent with that of global errors shown in Table 1. Error maps in the first row, derived from GT alone, usually have the most units shaded with a dark color (in the web version) for each population group and display a more dispersed pattern with more dark units locating in the rural periphery due to the fact that a small fraction of population in this area use Twitter. Error maps in the last row, derived from combination of GT and RP have the least dark units and alleviate the rural effect for interpolation with just GT as control data.

The assumption that geo-located Twitter and the other control layers reflect the population distribution differs for field versus object models. The field-based model assumes the distribution of population continuously varies across the study area by constructing a weighting surface with geo-located tweets as control data, while the object-based model constraints the distribution of population spatially by selecting out the residential polygons based on the location of road buffers, developed land, or residential parcels with population counts attached as an attribute variable. The field-based model underperforms against the object-based model using any of the control layers; however, the combination of the field-based model and the object-based model provides the best interpolation results. Thus the optimal population distribution model would be a discontinuous surface at one level by excluding non-residential areas, yet continuously varying within a residential area. Based on the results, the residential parcel layer was the best control data for delineating the populated area. Using night-time, geo-located tweets as control points also has the problem of age bias among Twitter users, which is shown by the higher error values for those 65 years and over by adding geo-located night tweets as an additional control layer. Collecting tweets with a longer period of time may alleviate this problem. The location of night-time, geo-located tweets are not necessarily guaranteed to be at a residence. People may work overtime or be at a recreational venue during that time. To overcome this problem, more sophisticate techniques need to be developed to refine the night time tweets, such as the text mining technique used by Ghosh and Guha (2013) to analyze the textual content of tweets.

Conclusions

This paper evaluates the effectiveness of Twitter data as single ancillary information or an integrated one combined with other

Table 2
Errors for specific age group population estimation with different control data sets.

Control data	18–29 years		30–49 years		50–64 years		65 years and over	
	Adj-RMSE	Adj-MAE	Adj-RMSE	Adj-MAE	Adj-RMSE	Adj-MAE	Adj-RMSE	Adj-MAE
DL	1.006	0.387	0.611	0.292	0.408	0.247	0.634	0.358
GT	1.011	0.410	0.635	0.331	0.454	0.292	0.703*	0.394
RB	0.943	0.399	0.568	0.287	0.360	0.229	0.579	0.337
RP	0.704	0.374	0.414	0.264	0.275	0.204	0.553	0.337
DL, GT	0.977	0.353	0.588	0.279*	0.414*	0.246*	0.656*	0.364*
DL, RB	0.905	0.360	0.523	0.246	0.323	0.195	0.564	0.323
DL, RP	0.630	0.329	0.352	0.212	0.213	0.150	0.507	0.292
GT, RB	0.868	0.342	0.510	0.252	0.339*	0.215*	0.585*	0.334*
GT, RP	0.577	0.320	0.336	0.227	0.257*	0.188*	0.550*	0.332*
RB, RP	0.666	0.354	0.379	0.232	0.237	0.169	0.541	0.320
DL, GT, RB	0.869	0.320	0.499	0.232*	0.321*	0.194*	0.581*	0.329*
DL, GT, RP	0.533	0.295	0.299	0.195*	0.206*	0.146*	0.519*	0.299*
DL, RB, RP	0.693	0.348	0.395	0.228	0.242	0.165	0.540	0.312
GT, RB, RP	0.567	0.321	0.333	0.216	0.236*	0.165*	0.560*	0.325*
DL, GT, RB, RP	0.596	0.319	0.347	0.213*	0.238*	0.163	0.555*	0.317*

Note: Adj-RMSE = adjusted root mean square error; Adj-MAE = adjusted mean square error; DL = developed land; GT = geo-located tweets; RB = road buffers; RP = residential parcels. Both linear and nonlinear distance decay functions were evaluated for geo-located tweets, however, only the function with the better results are presented in the table. Numbers with * on the up right corner are derived from non-linear distance decay function.

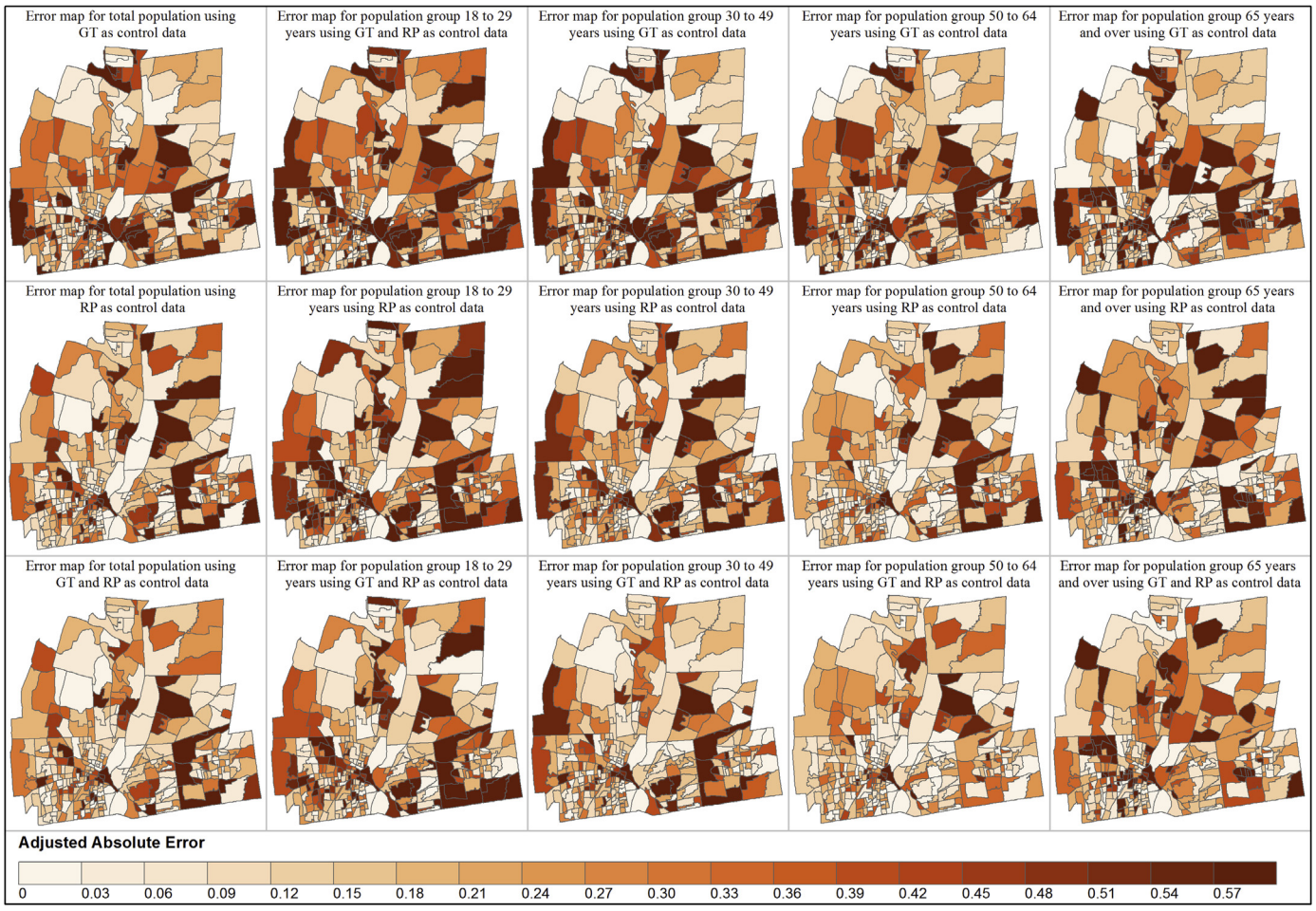

Fig. 1. Spatial distribution of adjusted absolute error for areal interpolation with GT, RP, and combination of GT and RP as control data.

control variables, such as remotely sensed developed land, road buffers, and residential parcels in the areal interpolation of population. Compared with the three other control variables, Twitter offers unprecedented accessibility to the public with almost any data published through the platform via Application Programming Interfaces (APIs) at no additional cost. The real time programmatic access enables tweets can be collected at any time resolution, thus the temporal discrepancy between the ancillary data and interpolated attribute variable can be easily overcome. However, the few studies devoted to the geographic information attached to tweets, mostly focused on the semantic tweets or network graph (Leetaru et al., 2013). This paper contributes to a growing literature about the geography of Twitter by correlating geo-located night time tweets with the distribution of residential population. Compared with other volunteered geographic information, such as the OpenStreetMap points-of-interest used by Bakillah et al. (2014) to infer population at building level, geographic locations are collected by mobile devices' geolocation features without any human interruption.

The comparative evaluation of ancillary datasets shows that the choice of specific ancillary data can significantly affect the performance of intelligent areal interpolation. The obvious enhancement effect of Twitter data on other control data under most circumstances suggests that it helps to better reflect population distribution by increasing variation in densities within a residential area delineated by other control data. Though Twitter data underperforms other more commonly used control data as a single ancillary layer of information, it is a promising ancillary dataset for estimating the distribution of population, especially age-specific population groups with large percentage of Twitter users. Another finding suggests that combining more control variables does not necessarily lead to improved areal interpolation performance. Some control data may weaken others, such as the addition of road buffers to residential parcels leads to a loss of estimation accuracy under most circumstances. Some control data also has little correlation with the interpolated variable, such as the addition of geo-located tweets for the interpolation of the elderly population that leads to loss of accuracy in most instances. The relationships between the control datasets and between each control layer and the interpolated population vary across different regions; thus experiments must be conducted first to determine which control variables to include for making integrated control layer.

Acknowledgements

We thank two anonymous reviewers for their constructive comments. This work is partially supported by the Fundamental Research Funds for the Central Universities (Grant 172210152).

References

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest.

International Journal of Geographical Information Science, 28(9), 1940–1963. http://dx.doi.org/10.1080/13658816.2014.909045.
Barbera, P. (2014). *streamR: Access to Twitter streaming API via R.* http://cran.r-project.org/web/packages/streamR/index.html Accessed November 2014.
Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., et al. (2013). Beyond the geotag: situating "big data" and leveraging the potential of the geoweb. *Cartography and Geographic Information Science, 40*(2), 130–139. http://dx.doi.org/10.1080/15230406.2013.777137.
Cromley, R. G., Hanink, D. M., & Bentley, G. C. (2012). A quantile regression approach to areal interpolation. *Annals of the Association of American Geographers, 102*(4), 763–777. http://dx.doi.org/10.1080/00045608.2011.627054.
Duggan, M., & Smith, A. (2014). *Social media update 2013*. Pew Research Center. http://pewinternet.org/Reports/2013/Social-Media-Update.aspx Accessed November 2014.
Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science, 28*(2), 125–138. http://dx.doi.org/10.1559/152304001782173727.
Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. *The Professional Geographer, 48*(3), 299–309. http://dx.doi.org/10.1111/j.0033-0124.1996.00299.x.
Flowerdew, R., & Green, M. (1989). Statistical methods for inference between incompatible zonal systems. In M. F. Goodchild, & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 239–247). London: Taylor and Francis.
Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40*(2), 90–102. http://dx.doi.org/10.1080/15230406.2013.776210.
Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*, 383–397.
Henderson, F., & Xia, Z. (1997). SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report. *IEEE Transactions on Geoscience and Remote Sensing, 35*(1), 79–85. http://dx.doi.org/10.1109/36.551936.
Holt, J. B., Lo, C. P., & Hodler, T. W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science, 31*(2), 103–121.
Iisaka, J., & Hegedus, E. (1982). Population estimation from Landsat imagery. *Remote Sensing of Environment, 12*(4), 259–272. http://dx.doi.org/10.1016/0034-4257(82)90039-6.
Langford, M. (2006). Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems, 30*(2), 161–180. http://dx.doi.org/10.1016/j.compenvurbsys.2004.07.001.
Langford, M. (2007). Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems, 31*(1), 19–32. http://dx.doi.org/10.1016/j.compenvurbsys.2005.07.005.
Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis, 45*(3), 324–344. http://dx.doi.org/10.1111/gean.12012.
Langford, M., Maguire, D., & Unwin, D. (1991). The areal interpolation problem: estimating population using remote sensing in a GIS framework. In I. Masser, & M. Blakemore (Eds.), *Handling geographical information: Methodology and potential applications* (pp. 55–77). London: Longman.
Leetaru, K. H., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: the geography of Twitter. *First Monday, 18*(5), 1–33. http://dx.doi.org/10.5210/fm.v18i5.4366.
Li, G., & Weng, Q. (2005). Using Landsat ETM+ imagery to measure population density in Indianapolis, Indiana, USA. *Photogrammetric Engineering and Remote Sensing, 71*(8), 947–958.
Lin, J., Cromley, R. G., Civco, D. L., Hanink, D. M., & Zhang, C. (2013). Evaluating the use of publicly available remotely sensed land cover data for areal interpolation. *GIScience and Remote Sensing, 50*(2), 212–230. http://dx.doi.org/10.1080/15481603.2013.795304.
Lin, J., Cromley, R. G., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS, 17*(1), 1–14. http://dx.doi.org/10.1080/19475683.2010.540258.
Liu, X., Clarke, K., & Herold, M. (2006). Population density and image texture: a comparison study. *Photogrammetric Engineering and Remote Sensing, 72*(2), 187–196.
Lo, C. P. (1986). *Applied remote sensing*. Harlow: Longman.
Lo, C. P. (2001). Modeling the population of China using DMSP operational linescan system nighttime data. *Photogrammetric Engineering and Remote Sensing, 67*(9), 1037–1047.
Lo, C. P. (2008). Population estimation using geographically weighted regression. *GIScience & Remote Sensing, 45*(2), 131–148. http://dx.doi.org/10.2747/1548-1603.45.2.131.
Lo, C., & Welch, R. (1977). Chinese urban population estimates. *Annals of the Association of American Geographers, 67*(2), 246–253.
Lu, Z., Im, J., Quackenbush, L., & Halligan, K. (2010). Population estimation based on multi-sensor data fusion. *International Journal of Remote Sensing, 31*(21), 5587–5604. http://dx.doi.org/10.1080/01431161.2010.496801.
Maantay, J. A., Maroko, A. R., & Herrmann, C. (2007). Mapping population distribution in the urban environment: the cadastral-based expert dasymetric system (CEDS). *Cartography and Geographic Information Science, 34*(2), 77–102. http://dx.doi.org/10.1559/152304007781002190.
Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42. http://dx.doi.org/10.1111/0033-0124.10042.
Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194. http://dx.doi.org/10.1559/152304006779077309.
Mrozinski, R. D., & Cromley, R. G. (1999). Singly- and doubly-constrained methods of areal interpolation for vector-based GIS. *Transactions in GIS, 3*(3), 285–301. http://dx.doi.org/10.1111/1467-9671.00022.
Qiu, F., & Cromley, R. (2013). Areal interpolation and dasymetric modeling. *Geographical Analysis, 45*(3), 213–215. http://dx.doi.org/10.1111/gean.12016.
Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review, 26*(5–6), 619–633. http://dx.doi.org/10.1007/s11113-007-9050-9.
Reibel, M., & Bufalino, M. E. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A, 37*(1), 127–139.
Sadahiro, Y. (2000). Accuracy of count data estimated by the point-in-polygon method. *Geographical Analysis, 32*(1), 64–89. http://dx.doi.org/10.1111/j.1538-4632.2000.tb00416.x.
Schroeder, J. P., & Van Riper, D. C. (2013). Because Muncie's densities are not Manhattan's: using geographical weighting in the EM algorithm for areal interpolation. *Geographical Analysis, 45*(3), 216–237. http://dx.doi.org/10.1111/gean.12014.
Sridharan, H., & Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis, 45*(3), 238–258. http://dx.doi.org/10.1111/gean.12010.
Su, M.-D., Lin, M.-C., Hsieh, H.-I., Tsai, B.-W., & Lin, C.-H. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *The Science of the Total Environment, 408*(20), 4807–4816. http://dx.doi.org/10.1016/j.scitotenv.2010.06.032.
Sutton, P. (1997). Modeling population density with night-time satellite imagery and GIS. *Computers, Environment and Urban Systems, 21*(3), 227–244. http://dx.doi.org/10.1016/S0198-9715(97)01005-3.
Tapp, A. (2010). Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science, 37*(3), 215–228. http://dx.doi.org/10.1559/152304010792194976.
Webster, C. J. (1996). Population and dwelling unit estimates from space. *Third World Planning Review, 18*(2), 155–176.
Wright, J. K. (1936). A method of mapping densities of population: with Cape Cod as an example. *Geographical Review, 26*(1), 103–110. http://dx.doi.org/10.2307/209467.
Wu, C., & Murray, A. T. (2005). A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems, 29*(5), 558–579. http://dx.doi.org/10.1016/j.compenvurbsys.2005.01.006.
Wu, S., Wang, L., & Qiu, X. (2008). Incorporating GIS building data and census housing statistics for sub-block-level population estimation. *The Professional Geographer, 60*(1), 121–135. http://dx.doi.org/10.1080/00330120701724251.
Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. *Computers, Environment and Urban Systems, 19*(4), 287–306. http://dx.doi.org/10.1016/0198-9715(95)00028-3.
Yuan, Y., Smith, R. M., & Limp, W. F. (1997). Remodeling census population with spatial information from Landsat TM imagery. *Computers, Environment and Urban Systems, 21*(3), 245–258. http://dx.doi.org/10.1016/S0198-9715(97)01003-X.
Zhang, C., & Qiu, F. (2011). A point-based intelligent approach to areal interpolation. *The Professional Geographer, 63*(2), 262–276. http://dx.doi.org/10.1080/00330124.2010.547792.
Zhuo, L., Ichinose, T., Zheng, J., Chen, J., Shi, P. J., & Li, X. (2009). Modelling the population density of China at the pixel level based on DMSP/OLS non-radiance-calibrated night-time light images. *International Journal of Remote Sensing, 30*(4), 1003–1018. http://dx.doi.org/10.1080/01431160802430693.