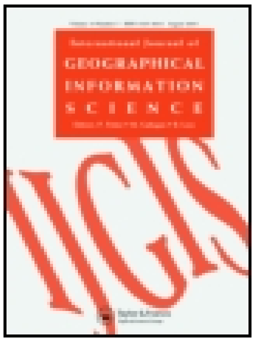


Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

# Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data

Yao Yao, Xiaoping Liu, Xia Li, Jinbao Zhang, Zhaotang Liang, Ke Mai & Yatao Zhang

**To cite this article:** Yao Yao, Xiaoping Liu, Xia Li, Jinbao Zhang, Zhaotang Liang, Ke Mai & Yatao Zhang (2017): Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2017.1290252

**To link to this article:** http://dx.doi.org/10.1080/13658816.2017.1290252

Published online: 08 Feb 2017.

Submit your article to this journal

Article views: 98

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tgis20

**Download by:** [Sun Yat-Sen University] **Date:** 16 February 2017, At: 03:16

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2017
http://dx.doi.org/10.1080/13658816.2017.1290252

# Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data

Yao Yao [a], Xiaoping Liu[b], Xia Li [b], Jinbao Zhang [a], Zhaotang Liang [a], Ke Mai [a] and Yatao Zhang [a]

[a]School of Geography and Planning, Sun Yat-sen University, Guangzhou, Guangdong province, China; [b]School of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, Guangdong province, China

**ABSTRACT**

Fine-scale population distribution data at the building level play an essential role in numerous fields, for example urban planning and disaster prevention. The rapid technological development of remote sensing (RS) and geographical information system (GIS) in recent decades has benefited numerous population distribution mapping studies. However, most of these studies focused on global population and environmental changes; few considered fine-scale population mapping at the local scale, largely because of a lack of reliable data and models. As geospatial big data booms, Internet-collected volunteered geographic information (VGI) can now be used to solve this problem. This article establishes a novel framework to map urban population distributions at the building scale by integrating multisource geospatial big data, which is essential for the fine-scale mapping of population distributions. First, Baidu points-of-interest (POIs) and real-time Tencent user densities (RTUD) are analyzed by using a random forest algorithm to down-scale the street-level population distribution to the grid level. Then, we design an effective iterative building-population gravity model to map population distributions at the building level. Meanwhile, we introduce a densely inhabited index (DII), generated by the proposed gravity model, which can be used to estimate the degree of residential crowding. According to a comparison with official community-level census data and the results of previous population mapping methods, our method exhibits the best accuracy (Pearson R = .8615, RMSE = 663.3250, p < .0001). The produced fine-scale population map can offer a more thorough understanding of inner city population distributions, which can thus help policy makers optimize the allocation of resources.

**ARTICLE HISTORY**
Received 29 August 2016
Accepted 29 January 2017

**KEYWORDS**
Population mapping; social sensing; geospatial big data; gravity model; random forest

## 1. Introduction

Fine-scale population distribution data, especially at the building level, play an important role in many fields, for example migrant population monitoring, resource allocation optimization and the analysis of city structures (Wu and Murray 2005, Lu *et al*. 2006,

**CONTACT** Xiaoping Liu liuxp3@mail.sysu.edu.cn

© 2017 Informa UK Limited, trading as Taylor & Francis Group

Bhaduri *et al*. 2007b, Gaughan *et al*. 2013, Langford 2013, Bakillah *et al*. 2014, Deville *et al*. 2014). As the most populous country in the world, China can only acquire detailed population maps via national censuses every decade. Their lengthy intervals and high costs prevent national censuses from reflecting population changes in a timely manner, especially at fine spatial resolutions (Zhou and Ma 2005). Thus, the use of general multisource GIS data sets is becoming more important to disaggregate and obtain fine-scale census data (Ural *et al*. 2011, Langford 2013, Bakillah *et al*. 2014, Deville *et al*. 2014, Stevens *et al*. 2015).

Many studies in the 1990s handled the rasterization problems of population mapping (Jones 1990, Deichmann 1996). The earliest studies focused on the spatial interpolation of populations, with several data sets and approaches proposed in China, including the spatial interpolation of population point data and statistical data and the dasymetric mapping method, which integrates auxiliary data (Langford *et al*. 1991, Eicher and Brewer 2001, Mennis 2003, Holt *et al*. 2004, Langford 2007). This framework has helped produce some global population data sets from typical data sources.

The Gridded Population of the World (GPW, version 2 and version 3), which offers gridded population data at a resolution of 2.5 arc-minutes, is a widely used global data set that adopts a spatially weighted method to refine population census data (Ciesin 2004). Based on the GPW, 30 arc-second resolution data sets called the Global Rural Urban Mapping Project (GRUMP) were produced and further developed by incorporating global land use classification data (Ciesin 2005) in 1990, 1995 and 2000. Currently, Landscan (http://web.ornl.gov/sci/landscan/) is the most prevalent spatial population data set and is the community standard for global population distributions. Landscan is updated annually (Bhaduri *et al*. 2007a). Compared to other large-scale population data at a global level, the spatial resolution of Landscan population data has increased to 1 km. However, all the population data sets that were mentioned above are at the global scale and can only be used to measure macroscale population changes. Therefore, the aforementioned methods are not suitable for studies on population behavior in cities, as the microscale is needed.

The strong correlation between remote sensing observations and large-scaled population distributions has been revealed with the rapid development of remote sensing and geographical information system (GIS) technology (Zha *et al*. 2003, Lu *et al*. 2006). Currently, the most popular approaches of extracting fine-scale population distributions still use remote sensing products, such as impervious surfaces and nighttime light data (Azar *et al*. 2010, Ural *et al*. 2011, Gaughan *et al*. 2013, Stevens *et al*. 2015, Yao *et al*. 2016). Azar *et al*. (2010) built a linear model between impervious surfaces and the population distribution and then obtained a refined population distribution map by extracting impervious surfaces in Haiti from Landsat images. Gaughan *et al*. (2013) proposed a spatial weighting logistical regression model that was based on Landsat-derived settlement maps and land cover data to map the population distribution in Southeast Asia at a spatial resolution of 100 m.

In light of these previous studies, Stevens *et al*. (2015) then used a random forest algorithm (RFA) to establish a nonparametric predictive model that could downscale census data and map fine-scale population distributions in Kenya, Vietnam and Cambodia. However, these studies were applied to countries with low economic development levels and simple urban functional structures. On the other hand, some attempts have been made to refine the scales of population distribution maps. For example, a few studies focused on mapping population distributions at the scale of buildings by using residential building footprints and census data to build empirical weighting models (Lwin and Murayama 2009, Ural *et al*. 2011). However, applying these approaches to fine-scale urban population mapping in China remains difficult because of various urban spatial structures and complicated population distributions in Chinese cities, which are influenced by many social, economic and cultural factors. Under such circumstances, building population distribution models of Chinese cities with only remote sensing products as auxiliary data, i.e. a simple combination of building footprints and census data, remains impractical.

In recent years, mobile location based service (LBS) technology has developed rapidly in China and large amounts of multisource geospatial big data can be obtained from Internet websites or public services, including points of interest (POIs), global positioning system (GPS) trajectories of public transportation, mobile communications and check-in data. These data sets can be combined to identify information about urban structures, economic vitality and traffic congestion (Ratti *et al*. 2006, Yue *et al*. 2009, Sevtsuk and Ratti 2010, Sun *et al*. 2011, Jacobs-Crisioni and Koomen 2012, Loibl and Peters-Anders 2012, Tong 2012, Chang *et al*. 2014, Hu *et al*. 2014, Liu *et al*. 2015) that are related to various human activities at the microscale and effectively reflect the features of population distributions and human behaviors. Deville *et al*. (2014) proposed a power law fitting model that used mobile phone base station data and census data to obtain dynamic population distribution maps of Portugal and France at the spatial scale of the radio coverage of the base stations.

However, mapping populations with mobile phone data has some obvious disadvantages. First, mobile phone base stations have variable effective transmitter powers, which create inconsistencies between generated Thiessen polygons and actual radio coverage. Second, a previous study indicated that the low correlation between the caller volume and underlying population reveals the inadequacy of treating the distribution of mobile subscribers as a representation of the distribution of an entire population (Kang *et al*. 2012). Moreover, collecting mobile phone data in China is nearly impossible because of the government's personal privacy policy. To overcome the limits of policies, Bakillah (2014) used volunteered geographic information (VGI) to map the population distribution at the building level in Hamburg and produced a satisfactory mapping result. However, Bakillah's method only relied on POIs and fine land use/land cover data (LULC) and did not consider the spatial heterogeneity of the population distribution when computing the population inside buildings. To our knowledge, different geospatial big data can capture different aspects of the ground truth, especially for actual population distributions (Liu *et al*. 2015). However, none of the above studies could allocate population distributions at the building level by using multisource geospatial big data because of a lack of an effective model. We suggest that a model that can effectively fuse information from multisource geospatial data, including official survey data and big data, can better reveal the actual population distribution at a fine scale.

In our study, we designed a framework to map the population distribution at the building level by integrating multisource geospatial data. The first step was to obtain the peak points of population density in census units and preliminary population disaggregation data by computing the correlation coefficients between each POI and population density category. In the second step, a nonlinear fitting model was built to map the population distribution while considering the spatial heterogeneity and several geospatial big data sets were introduced into the fitting model as auxiliary data sets. In the last step, the population and nonresidential area in buildings were computed by using our proposed iterative gravity model. Finally, this model was used to generate fine-scale population distribution maps at the building level in five central districts in Guangzhou, which is the largest city in southern China.

## 2. Study area and data

Our study case was conducted in five central urban districts (Yuexiu, Liwan, Tianhe, Haizhu and Baiyun) in Guangzhou, Guangdong Province. As the most important districts over the past two thousand years, these five districts have the highest population densities in Guangdong Province and serve as the political, cultural and economic centers of Guangzhou and present complex urban morphologies with a variety of land use types, including residential, commercial, industrial and education (Liu *et al*. 2014). Moreover, the living conditions in our study area vary from urban villages to luxury housing and working environments also exhibit considerable diversity, which creates complex patterns among population behaviors. Meanwhile, Guangzhou City, the political, economic and cultural center of Guangdong Province and one of the country's most important economic development centers, contains a large portion of the migrant population. As far as we know, the increasing residential and migrant populations will have environmental and ecological effects on cities and create serious challenges for the government regarding the configuration of education, transportation, medical facilities and other resources (Chen *et al*. 2013, Aunan and Wang 2014).

Based on statistical data from the Sixth National Population Census of China in 2010, the total area of all five districts in our study area is 984.8 km$^2$, approximately 13.61% of the total area of Guangzhou. The total number of administrative community-level units is 1278, while the total number of street-level units is 101, and the recorded permanent resident population in 2014 was 6.9496 million (http://data.gzstats.gov.cn/gzStat1/chaxun/ndsj.jsp), comprising 58.75% of the total resident population of Guangzhou. In our study area, Yuexiu has the highest resident population density (approximately 29,600 persons/km$^2$), followed by Liwan (approximately 16,200 persons/km$^2$), which is the oldest district and the political and cultural center of Guangzhou. Haizhu has a population density of 13,600 persons/km$^2$ contains multiple universities and factories, serving as both the educational and industrial district. The population density of Tianhe is ranked fourth among the five districts (approximately 10,300 persons/km$^2$). Tianhe is home to the Central Business District (CBD) of Guangzhou and its growth rate has been the fastest over the past 10 years, making this district the most concentrated commercial zone in Guangzhou. Panyu District has the smallest resident population density in the study area (approximately 3200 persons/km$^2$). This district is a mixed zone that contains commercial, industrial and agricultural functions. The study area, population density and land cover data are illustrated in Figure 1.

Several auxiliary spatial data sets were also applied in this study. In addition to the basic GIS data sets of Guangzhou, the POIs in our study were provided by Baidu Map Services (http://map.baidu.com), which is the most used and largest web map service

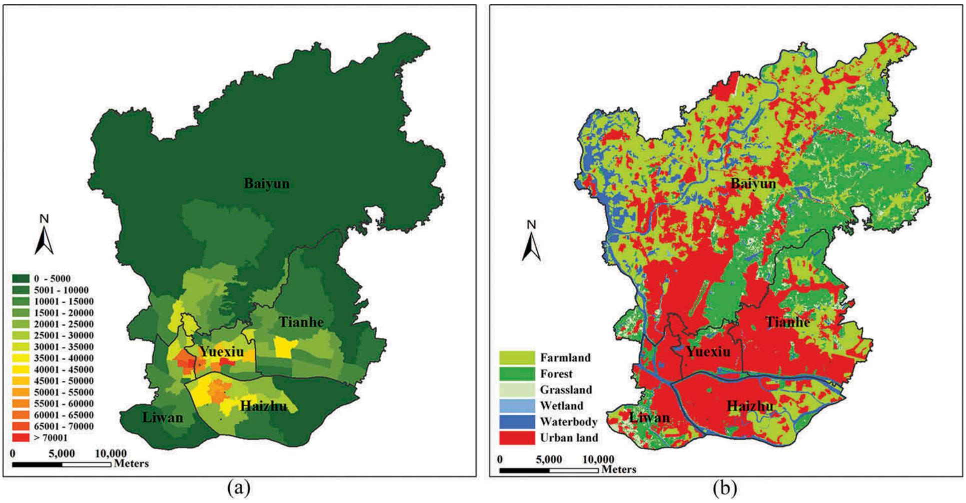


**Figure 1.** Case study area: Liwan district, Haizhu district, Yuexiu district, Tianhe district and Baiyun district, Guangzhou, Guangdong province, PRC. (a) Population density (unit: person/km$^2$) in the five central districts of Guangzhou within the street level census unit (data source: China's Sixth National Population Census). (b) Land cover data at a spatial resolution of 30 m (data source: Chinese Research Institute of Surveying and Mapping).

provider in China. With the help of the application programming interfaces (APIs) that were provided by Baidu Map, we extracted the population-related POIs from our study area, approximately 237,402 records in total, including business establishments, commercial sites, educational facilities (e.g. kindergartens, primary schools, middle schools), residential communities, clinical facilities and scenic locations. The density of urban road networks is directly related to the level of urbanization, especially in developing countries (Poston Jr and Yaukey 2013). Road data from the study area in 2015 were downloaded from the OpenStreetMap (OSM) website (http://openstreetmap.org).

The realtime Tencent user density (RTUD) is the most important data type in this study and was provided by Tencent (http://www.qq.com), one of the largest Internet companies in both China and the world. The RTUD records the locations of smart phone users who were using Tencent applications, such as Tencent Mobile App QQ (a messenger-like software), WeChat (a mobile chat software), Soso Maps (a web mapping services and navigation software) and some other mobile applications that provide LBS services. According to the Tencent Data Report of WeChat Users (Co. 2015), the average daily number of total activities from WeChat accounts has reached approximately 570 million, more than one-third of the total population in China. Furthermore, the total number of Tencent users has reached 808 million, and 60% of Tencent users range in age from 15 to 29 years old. According to 'Big data white paper of Tencent Co. in 2016 (http://bigdata.qq.com), the ratio of Tencent users to the total population has exceeded 93% in China's first-tier cities, such as Beijing, Shanghai and Guangzhou. Figure 3 shows that the average per-hour number of online users among the total population of the study area is approximately 28.99% during the day. The Tencent Company is the most important Internet service portal in China, so the distribution of Tencent users can be seen as a type of bias sampling of

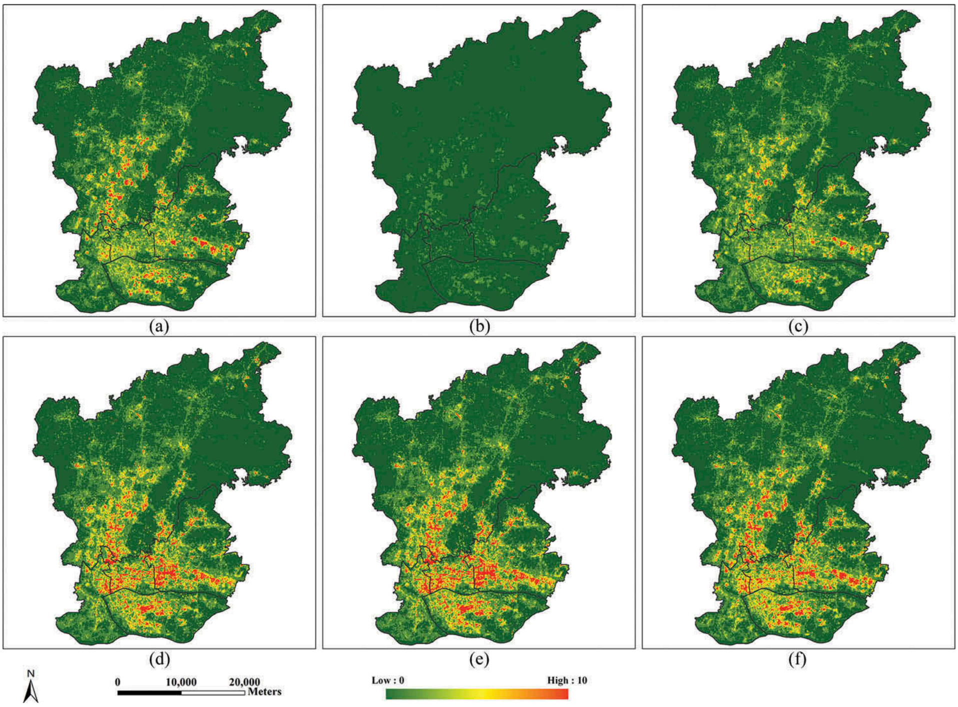
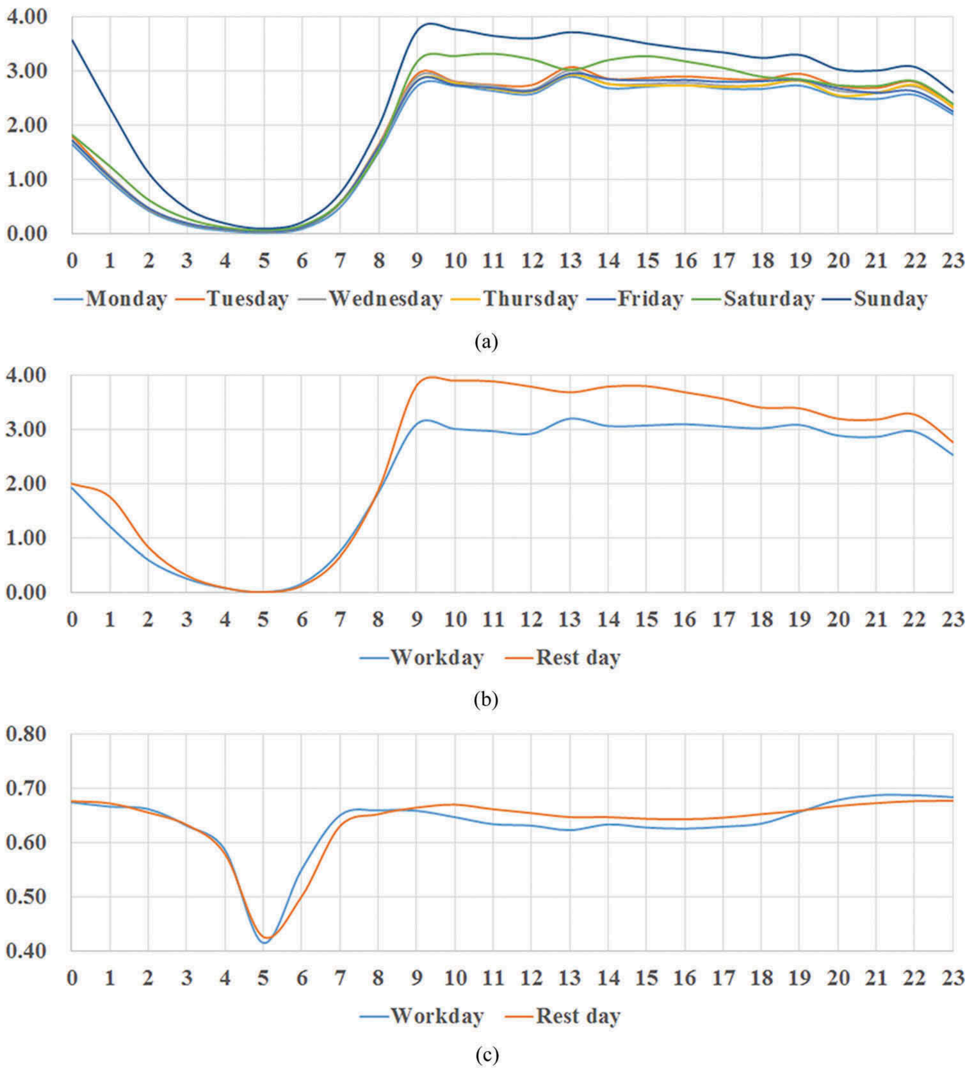


**Figure 2.** Real-time Tencent user density data (RTUD, unit: person, spatial resolution: 25 m) on 16 June 2015 (Thursday) in the study area of Guangzhou: (a) 00:00 am, (b) 04:00 am, (c) 08:00 am, (d) 12:00 pm, (e) 16:00 pm and (f) 20:00 pm.

the general population dynamic distribution. Recently, easy go, a public map service of Tencent WeChat, offered a public query service of a location's congestion degree. We designed a web crawler to fetch RTUD data from the study area from 14 June to 28 June 2015. After coordinate correction and rasterization, we produced Multi-band RTUD data at a spatial resolution of 25 m (Figure 2).

Figure 3(a) shows the temporal changes in the average total number of RTUDs in the study area in one week. (1) Obvious periodic oscillations of 24 h occurred for the temporal RTUD data. (2) The total number of active Tencent users steadily ranged from 1.6 million to 2.0 million during daytime hours (9:00 am to 18:00 am). Additionally, we observed a slight difference between rest days and work days during the daytime: the previous population density was likely greater that the migrant population from outside, who usually enter the city over the weekend. (3) During the night on work days and rest days (21:00 pm to 01:00 am the next day), the number of users was steadily distributed between 1.2 million and 1.8 million (Figure 3 (b)), which had the strongest correlation with census data at the community level (Figure 3(c)). Therefore, we introduced a time series-based data compression method from a previous study (Liu *et al*. 2015) to pre-process the RTUD data. We smoothed and compressed the entire RTUD time series and performed mean filtering on the work day and rest day data to reduce the data size and computational work without missing too much information. Additionally, we only applied the nighttime RTUD to



(a)

(b)

(c)

**Figure 3.** Temporal changes over hours (*x* axis) in the study area: (a) average total number of Tencent users in one week (*y* axis), (b) average total number of Tencent users on work days and rest days (*y* axis) and (c) Pearson's correlation coefficient between the number of Tencent users and census data at the community level (*y* axis).

build the population distribution model because of the huge amount of outside population during the daytime.

Moreover, the spatial distribution of buildings has a relationship with the urban inhabitant distribution (Ural *et al.* 2011). The Guangzhou Urban Planning Bureau provided 579,112 building-attribute records in the study area, including residential community, urban village, government, educational, commercial, shopping, enterprise and public institution classes, with attribute information such as property type, location, area and floor height. Then, we split the building data into residential and working classes (Figure 4).

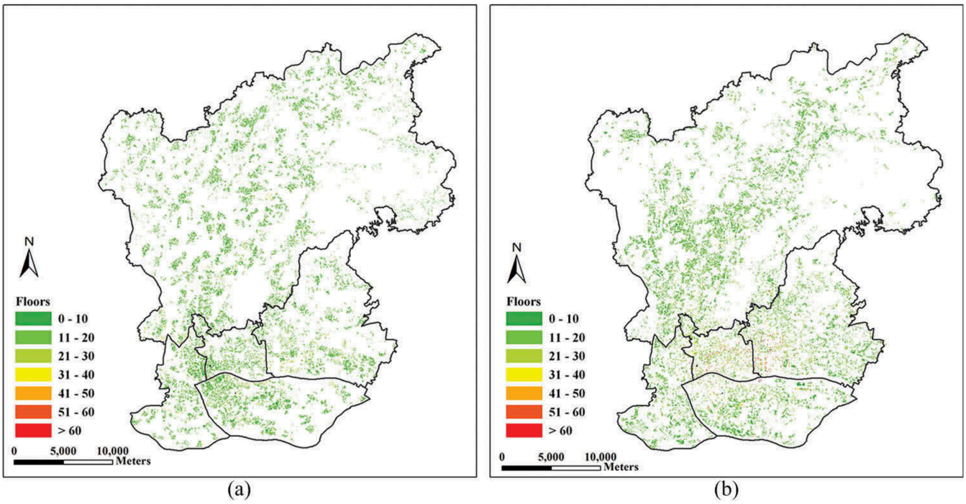


**Figure 4.** Extracted (a) residential and (b) working building footprints in the study area of Guangzhou.

## 3. Methodology

The purpose of our study was to downscale census data into population distributions at the building level while considering the spatial heterogeneity of population distributions (see flowchart in Figure 5). In this study, we took three steps to estimate the population distribution in every building. (1) We selected POIs that could indicate high population density and map the preliminary population disaggregation in each census unit. (2) We built a nonlinear population model by integrating the RFA and several sources of geospatial big data to indicate the spatial heterogeneity in the study area. (3) We calculated the population distributions at the building level by using the proposed iterative building-population gravity model and compared the reliability of the results with census data and results from other state-of-the-art methods.

The models described below were implemented by our research team using C++ on Windows 8.1 (×64), linking with open-source libraries including GDAL (http://www.gdal.org), CGAL (http://www.cgal.org/) and Shark (http://image.diku.dk/shark/).

### 3.1. *Extracting peak points of population density*

'High-density indicator' (HDI) points are POIs that have strong correlations with population density (Bakillah *et al*. 2014). In previous studies, HDIs were determined by calculating the correlation between the population size and the POI counts in each census unit (Bakillah *et al*. 2014). This straightforward method can provide convenient HDI points but does not consider the existence of meaningless high-frequency POIs, such as the name tags of roads and districts. To tackle this problem, we used term frequency-inverse document frequency (TF-IDF) values as a better alternative to POI counts. TF-IDF is a statistical method that is extensively used in natural language processing. By evaluating the significance degrees of different words, TF-IDF can filter out common words and retain the most important and distinct words as the subject classification criteria. In our study, we took census data as the

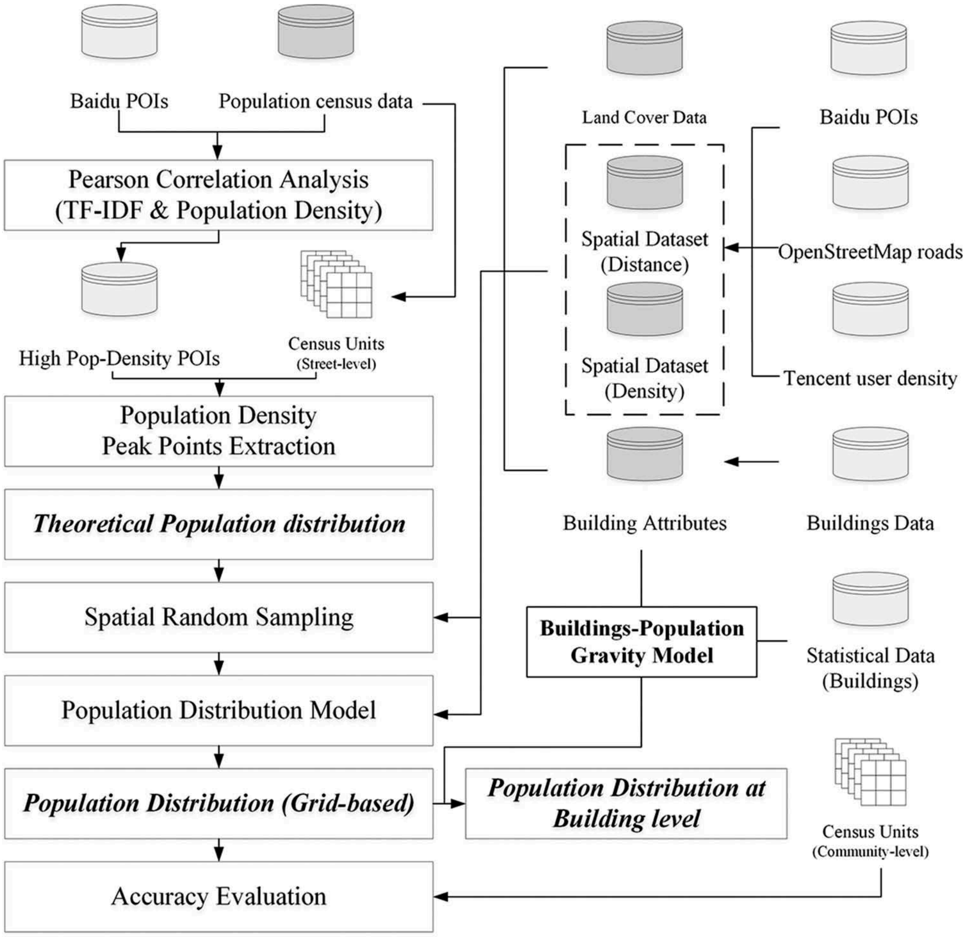

Baidu POIs
Population census data
Pearson Correlation Analysis (TF-IDF & Population Density)
High Pop-Density POIs
Census Units (Street-level)
Population Density Peak Points Extraction
Theoretical Population distribution
Spatial Random Sampling
Population Distribution Model
Population Distribution (Grid-based)
Accuracy Evaluation
Land Cover Data
Baidu POIs
Spatial Dataset (Distance)
OpenStreetMap roads
Spatial Dataset (Density)
Tencent user density
Building Attributes
Buildings Data
Buildings-Population Gravity Model
Statistical Data (Buildings)
Population Distribution at Building level
Census Units (Community-level)

**Figure 5.** Flowchart of the proposed method for mapping populations at the building scale.

corpus data $D$, which comprised statistical units $d$. The total number of $d$ in $D$ was $K$. Thus, the TF-IDF values of the $i$th type of POI in the $j$th census unit could be specified as:

$$\begin{cases} \mathrm{TF}_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \\ \mathrm{IDF}_i = \log\frac{|D|}{\left|\left\{j{:}\mathrm{POI}_i \in d_j\right\}\right|} \\ \mathrm{TF} - \mathrm{IDF}_{ij} = \mathrm{TF}_{ij}\cdot\mathrm{IDF}_i \end{cases} \quad (1)$$

Where $\mathrm{TF}_{ij}$ denotes the term frequency of the $i$th type of POI in the $j$th census unit and $\mathrm{IDF}_i$is the implementation result of the document frequency inversion of the $i$th type of POI. $|D|$represents the total number of census units and $\left|\left\{j : \mathrm{POI}_i \in d_j\right\}\right|$is the number of census units that contain the $i$th type of POI.

After computing TF-IDF vectors for each type of POI, the Pearson correlation coefficient between the$i$th type of POI and population density could be obtained as:

$$\rho_{\mathrm{POI}_i.\mathrm{PD}} = \frac{N\sum_j \mathrm{TI}_{ij}\cdot \mathrm{PD}_j \;-\; \sum_j \mathrm{TI}_{ij}\cdot \sum_j \mathrm{PD}_j}{\sqrt{N\sum_j \mathrm{TI}_{ij}^2 \;-\; \left(\sum_j \mathrm{TI}_{ij}\right)^2}\cdot\sqrt{N\sum_j \mathrm{PD}_{ij}^2 \;-\; \left(\sum_j \mathrm{PD}_{ij}\right)^2}} \tag{2}$$

Where $\mathrm{TI}_{ij}$ is the TF-IDF value of the $i$th type of POI in the $j$th census unit and $\mathrm{PD}_j$denotes the population size of the $j$th census unit.

By sorting $\rho_{\mathrm{POI}_i.\mathrm{PD}}$ in descending order, the correlation strengths between different types of POIs and population density could be clearly shown. The HDI points were already produced, so the peak points (PPs) of the point density in every census unit were calculated in the next step:

(a) In a given unit, if a specific POI category had the highest correlation with the population density, then the peak point could be defined as the centroid of this category. For example, the most relevant category in our study was the clinical facility category, so if clinical facility POIs were present in the unit, the centroid would be designated as the unit's PP.

(b) Otherwise, if the most relevant POI category was absent in a given unit, then we would assign its PPs based on the centroid of a following category.

(c) If no POIs were located in a unit, we would define the centroids of the artificial surfaces in land cover data as replacements of PP.

## 3.2. *Preliminary population disaggregation*

Insufficient sampling can lead to over-fitting problems in model construction. To prevent such sample problems, Deville (2014) proposed a buffer method that establishes a disc-sampling zone that is centered at PPs to ensure adequate samples in census units. Although Deville's method ensures sample abundance, it does not consider the spatial heterogeneity of the population distribution, which means that the sampling error can easily accumulate. With this understanding of the spatial heterogeneity, we assumed that the population number in every census unit satisfied a Gaussian distribution, whose peak lay at the PPs. In our model, the population aggregation data were first adjusted to support the hypothesis. In the next step, the sampling procedure was performed. The calculations of the population distribution probability $f_{unit_i}(x,y)$ are specified in Equation (3). In particular, $f_{unit_i}(x,y)$ denotes the population distribution probability at location (*x*, *y*) in the *i*th census unit.

$$\begin{cases} f_{unit_i}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{-\frac{v_{unit_i}(x,y)}{2(1-\rho^2)}\right\} \\ v_{unit_i}(x,y) = \left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \end{cases} \tag{3}$$

In Equation (3), both $X$ *and* $Y$ obey a normal distribution, in which $X\tilde{N}(\mu_X,\sigma_X^2)$ and $Y\tilde{N}(\mu_Y,\sigma_Y^2)$. $\rho$ is the correlation coefficient between $X$ and $Y$. $\mu_X$and $\mu_Y$ are set to the location of a PP in the $i$th census unit. $\sigma_X$ and $\sigma_Y$ are equal to $\underset{(x,y)}{\mathrm{argmax}}[Dist(loc(x,y),PP_i)]$, while $Dist(loc(x,y),\mathrm{PP}_i)$ is the distance from location (*x*, *y*) of a PP in the $i$th census unit. Thus, the estimated population size $\mathrm{PD}(x, y)$ at location ($x$, $y$) can be calculated as:

$$\begin{cases} \mathrm{PD}_i(x,y) = \mathrm{PD}_i \cdot f_{unit_i}(x,y) \\ \mathrm{PD}(x,y) = \sum\limits_i \mathrm{PD}_i(x,y) \end{cases} \tag{4}$$

### 3.3. *Mapping the HSR population distribution with an RFA-based fitting model*

In the kernel density analysis, the search radius was set to 500 m (Chen *et al*. 2016, Hu *et al*. 2016). Buildings can be categorized into two groups based on the building floor number data and used to produce two auxiliary grids. One group is residential buildings and the other is work buildings. The number of building floors can provide valuable insight into actual buildings and their usable areas based on the grid size. In our study, OSM data were applied to generate the distance to roads, which could be input into the nonlinear fitting model to simulate the population distribution. We utilized a bilinear interpolation for data whose spatial resolutions did not perfectly correspond to others.

After pre-processing all the data, a training data set $D$ was generated. First, a buffer zone was drawn to define the sampling range, setting the PPs in every census unit as buffer centers and producing the corresponding buffer discs (radius $=r_s$). Second, mathematical statistics were calculated based on the buffers to obtain the total size of the population $y$, the summation of each spatial variable $X = [x_1, x_2, \ldots, x_n]$, and the total area of residential buildings $A_{Bu}$, where $n$ denotes the total number of spatial variables, followed by saving the produced sample variables (location, buffer size, $X$, $A_{Bu}$, $y$) and building data in a training data set $D$. The final step was recursion. The sampling process was repeatedly executed by increasing the buffer radius $r_s$ until the threshold radius $r_e$ was reached. The growth size was set to a fixed length $\Delta r$. Finally, the sample size of the training data set $D$ was increased to $K = \left(\frac{(r_e - r_s)}{\Delta r} + 1\right) \cdot N_u$, where $N_u$ is the total number of census units.

Because of the existence of the Collective Household (Hukou) Policy in China, the Hukous of some individuals are registered in accordance with their work unit (Danwei) instead of their household. Thus, in some areas with numerous government organizations and state-owned enterprises, census population estimates greatly outweigh the actual inhabitant population (Chan and Zhang 1999, Zhu 2007, Fan 2008). We employed the three-sigma rule to avoid fitting errors from these outliers (Grafarend 2006), assuming that the per capita housing area of each sample is $\lambda = y/A_{Bu}$, the average per capita housing area of all the samples in training data set $D$ is $\bar{\lambda}_D$, and the standard deviation is $\sigma_\lambda$. If $\lambda_i$ of the $i$th sample does not fall into the range $\left[\bar{\lambda}_D - 3\sigma_\lambda, \bar{\lambda}_D + 3\sigma_\lambda\right]$, the $i$th sample is regarded as an outlier and removed. The training data set was thoroughly refined through this process.

To address the spatial correlation variables, a RFA-based nonlinear and nonparametric fitting model was introduced in this study to fit spatial variables and population density. We assumed that $X_{ij}(i \in [1, M], j \in [1, N])$ and $Y_i(i \in [1, M])$ are the average values of each spatial variable and population density in the training data set, respectively. $M$ is the total number of training samples, and $N$ denotes the total number of spatial variables. Then, an RFA-based nonlinear fitting model was built to fit these spatial variables $X$ to the population densities $Y$. RFA is a nonlinear and nonparametric fitting model (Breiman 2001, Biau 2012, Fakhraei *et al.* 2014). RFA randomly takes $m \cdot n$-dimension $(m \ll M, n \ll N)$ samples depending on the training data's spatial dimensions through the bagging method. $C$ trees are trained by these selected sample data without pruning operations. RFA does not use all the variables to split nodes; instead, only some of the variables are randomly selected to make decisions. With this approach, the correlation of each decision tree can be reduced, enhancing the classification accuracy of each decision tree. During the node-splitting process, $\log(M+1)$ variables were randomly selected to participate in the calculation, satisfying the randomness requirement and forming the random forest. After averaging the results $\tilde{Y}_i$ from every decision tree based on the total tree number $C$, we obtained the fitted result $\bar{Y}$, which was calculated according to the equation $\bar{Y} = \sum_{i=1}^{c} \tilde{Y}_i / C$.

Furthermore, RFA is an aggregation of decision tree classifiers. A new sub-data set is generated by extracting random samples from the original training data set via the bagging method (Biau 2012). During random feature selection, individual decision trees are constructed from each training sub-data set and these decision trees are not pruned during the growth process, so we can obtain an out-of-bag (OOB) based estimation error report for each decision tree. The generalization error of RFA can be calculated by averaging the errors of the decision trees via OOB estimation. The RFA-based fitting model from previous studies overcame the multiple correlative problems among spatial variables, especially in higher-dimensional fitting situations (Palczewska *et al.* 2014). To obtain the contribution weights $w_i (i \in [1, N])$ of each spatial variable during the training process, random noise was added to each type of normalized spatial training variable over time.

During the next step, these noisy training data sets were input into the previously trained fitting model with statistical population densities. The average errors $E_i (i \in [1, N])$ can still be calculated even when the random noise is added to spatial variables of the $i$th type. Assuming that $\bar{E}$ is the original training error without any noise, the contribution weights $w_i (i \in [1, N])$ of the $i$th spatial variable can be computed by Equation (5) (Fakhraei *et al.* 2014, Palczewska *et al.* 2014):

$$w_i = \frac{var(E_i - \bar{E})}{\sum_{i=1}^{N} var(E_i - \bar{E})} \tag{5}$$

After creating the RFA-based fitting model, a spatial-differentiation-sensitive population distribution map was finally produced. The Chinese government only updates the total population number annually but without providing detailed information at the administrative distribution scale, but we can utilize these model-generated data for population adjustment. Suppose that the RFA simulation result of the total population is $\mathrm{pop}_{\mathrm{RFA}}$ and that the real population from official data is $\mathrm{pop}_{\mathrm{real}}$; then, the adjustment coefficient of the simulation result would be $c = \mathrm{pop}_{\mathrm{RFA}} / \mathrm{pop}_{\mathrm{real}}$. The fine-scale population distribution can be fitted by adjusting the simulation result.

### 3.4. *Mapping population distributions at the building level*

The final step was to allocate the population at the building level and calculate the vacancy rate of housing. The iterative-gravity model was the core of this step. In our study, a seed-growing algorithm was applied in the gravity model. First, we set the centroids of residential buildings as the initial growing seed positions. The corresponding entities of growing seeds were not buildings but pixels in the population distribution map. These initial growing seeds were the first parameters that were input into the gravity model. Assuming that the *k*th pixel of the *i*th census unit is denoted by $\text{loc}_{ik}(x_{ik}, y_{ik})$, the *j*th residential building in the *i*th census unit is denoted by its centroid location $\text{loc}_{ij}(x_{ij}, y_{ij})$. Then, the gravity force between $\text{loc}_{ik}(x_{ik}, y_{ik})$ and $\text{loc}_{ij}(x_{ij}, y_{ij})$ can be specified mathematically by:

$$\begin{cases} \text{Gravity }(i, j, k, t+1) = \dfrac{F_{ij} \cdot A_{ij} - \left[\text{pop}_{ij}(t) + \text{pop}_{ik}\right] \cdot \bar{A} - c_i}{\left(\sqrt{\left(x_{ij} - x_{ik}\right)^2 + \left(y_{ij} - y_{ik}\right)^2}\right)^{\beta}} j, k \in i \\ j' = \underset{j}{\text{argmax}} \left[\text{Gravity}(i, j, k, t+1)\right] \end{cases} \tag{6}$$

Where $A_{ij}$ and $F_{ij}$ denote the floor surface area and the total floors of building $\text{loc}_{ij}(x_{ij}, y_{ij})$, respectively. $\text{pop}_{ij}(t)$ is the residential population in building $\text{loc}_{ij}(x_{ij}, y_{ij})$ after *t* iterations. The constant $\bar{A}$ is the per capita housing area of the census unit. In this study, we specified $\bar{A}$ as $34.4m^2$ according to the government statistical data of Guangzhou. The variable $c_i$ is the simulated value of the average area of housing vacancies in the *i*th census unit and an adjustment factor. The initial value of $c_i$, marked as $c_i(init)$ in Equation (7), was set as:

$$c_i(\text{init}) = \frac{\sum_n F_{ij} \cdot A_{ij} - \text{pop}_{\text{total}} \cdot \bar{A}}{n_i} \tag{7}$$

$c_i$ is mainly an adjustment factor. The existence of $c_i$. is indispensable to judge whether the population allocation result is satisfactory. $\beta$ is the distance decay parameter. The power law form of distance decay functions may better reveal the effect of the inherent distance of spatial interactions (Palczewska *et al.* 2014). Liu *et al.* (2015) suggested that $\beta$ should be within [1, 2] based on his studies on individual movements at the urban scale. After consulting related studies, we set 1.5 as an appropriate value of $\beta$.

Thus, the population at location $\text{loc}_{ik}(x_{ik}, y_{ik})$ was allocated to building $j'$, which had the maximum attraction. After assigning the current seed, the allocation process of the neighboring seed was initiated.

As mentioned above, we had to validate that the population allocation result was reasonable. An initial value $c_i$ was established to calculate the per capita housing area in the *i*th census unit and compared to $\bar{A}$. Assuming that the total number of residential buildings in the *i*th census unit is $n_i$ and the total population is $\text{pop}_{ij}$, the per capita housing area of the *j*th residential building in the *i*th census unit $\bar{H}_{ij}$ and the per capita housing area in the *i*th census unit $\bar{H}_i$ can be calculated as:

$$\begin{cases} \bar{H}_{ij} = \frac{F_{ij} \cdot A_{ij} - c_i}{\text{pop}_{ij}} \\ \bar{H}_i = \sum_n \bar{H}_{ij} / n_i \end{cases} \tag{8}$$

The iteration precision is denoted as $\alpha$. If the deviation between $\bar{H}_i$ and $\bar{A}$ is beyond $\alpha$, the allocation results during this time are considered failures. The iteration would continue until reaching the threshold precision. In this study, we adopted the Dichotomy Method of numerical approximation, and the optimal population allocation in each region was reached through repeated adjustment of $c_i$. Finally, fine-scale population distributions could be estimated at the building level, and the densely inhabited index (DII) could be simultaneously calculated with Equation (9), which can reveal the development of the real estate market and the status of the regional economy.

$$\text{DII} = \frac{c_i \cdot n_i}{\sum_j F_{ij} \cdot A_{ij}} \tag{9}$$

### 3.5. *Accuracy assessment*

The accuracy of our proposed model was assessed using the Pearson correlation coefficient ($R_p$), coefficient of determination ($R^2$) and root mean square error (RMSE) statistics (Bhaduri *et al*. 2007a, Azar *et al*. 2010, Aunan and Wang 2014, Deville *et al*. 2014, Stevens *et al*. 2015), comparing the predicted values with community-level census and official household survey data.

## 4. Results

### 4.1. *Population disaggregation with POIs and census data*

Table 1 lists the Pearson's correlation coefficient between each POI and population density category in descending order, where the top three categories are clinical facilities, residential communities and education. The peak points of the population density were produced using the approach described in section 3.1. Preliminary population disaggregation data were obtained with the above normally distributed spatial probability model. During the next step, the population disaggregation data were used as input training data to map the population distribution.

### 4.2. *Spatial sampling and population mapping*

As we built the RFA-based population-fitting model, we selected 26 categories of spatial variables as inputs, including Tencent RTUD, Baidu POIs, OSM roads and basic GIS data. The buffer radii of the sampling discs were set from 50 to 1000 m in steps of 25 m, and the centers of the discs were assigned to the locations of PPs when sampling spatial data. During this process, 4000 training samples were selected, 207 of which were outliers, which were excluded in the following data-cleaning step. We implemented the RFA-based fitting model with 100 decision trees, and the percentages of the training

**Table 1.** Pearson correlation coefficient between the TF-IDF of POIs and population density (The top 3 most important factors are rendered in bold).

| Categories of POIs | Correlation | Categories of POIs | Correlation |
|---|---|---|---|
| **Clinic facility** | **.7874** | Entertainment | .5309 |
| **Residential community** | **.6756** | Government | .5077 |
| **Education** | **.6561** | Hotels | .3765 |
| Restaurant | .6536 | Virtual landmark | .3535 |
| Automobile service | .6396 | Scenic spots | .3512 |
| Traffic Facility | .6262 | Financial network-point | .1977 |
| Life service | .6261 | Business Building | .0719 |
| Road label | .6215 | Location label | .0323 |
| Shopping | .5539 | Mountain | −.0049 |
| Corporation | .5425 | Greenland | −.3814 |

data set and out-of-bag data set were set to .6 and .4, respectively, for cross-validation. Additionally, we examined the influences of different sampling methods, including Deville (2014)'s method, random sampling method and proposed method, Table 5 shows the comparison results. And we analysis the meaning of the sampling radius sizes set in the proposed method, as illustrated in Figure 9.

Next, the fitting model of the spatial variables and census population was generated, in which the OOB average error was 1.5911. Figure 6 shows the mapping result of the simulated population distribution when downscaled from statistical data at the street level and Table 2 illustrates the respective contribution weights of all 25 spatial input variables. Significance statistics, including Pearson correlation coefficients and a standard coefficient of determination analysis, were used to compare the simulated population distribution and government census data within each unit.

Table 2 demonstrates that the average nighttime Tencent user density was most similar to the actual population distribution (13.80%), which indicates that Tencent RTUD can be well applied to simulate nighttime population distribution and population dynamics. The POI densities of life services, educational facilities and clinical facilities were also three important factors in the population down-scaling model via RFA, comprising more than one third (34.67%) of the contribution to the simulated result. This result was similar to the correlation between POIs and population (Table 1). Residential buildings considerably influenced the inhabitant distribution, with up to a 9.85% contribution. Conversely, working buildings contributed almost nothing (0.55%). In addition to the above-mentioned spatial variables, the distance to roads from Opensteetmap.org had a strong relationship with the inhabitant distribution, comprising 4.64% of the contribution. The road density had a close relationship with human activity because it affects the accessible range. Thus, the distance to each district center obviously played an important role.

Furthermore, the densities of life services and educational facilities had much stronger effects on the population density among the Baidu POIs, comprising 13.66% and 12.40%, respectively. Meanwhile, the density of corporations comprised 0.90% of the density distribution, which indicates that areas with high densities of commercial and business facilities may have a low correlation with nighttime population densities (from 21:00 pm to 1:00 am the next day; Figure 6(d)). Moreover, Guangzhou, which developed from a southern trade port, is now a metropolis because of rapid economic development, in which fisheries and port-based trade

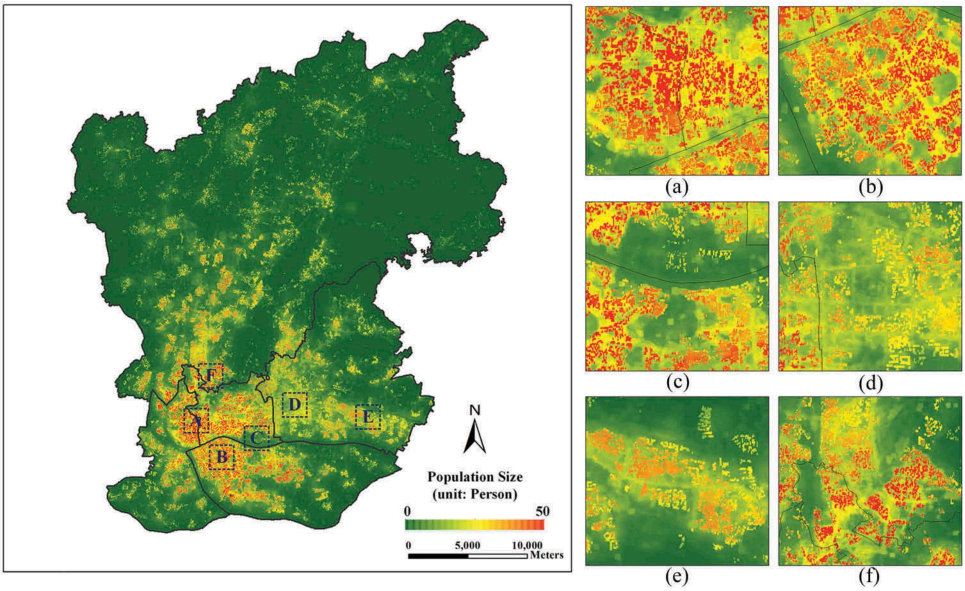


**Figure 6.** Simulated population distribution map at a local scale of 25 m via RFA. (a) Zhongshan Seventh Road (old city center), (b) Jiangnan Road (shopping center), (c) Dongshan Lake Park (north, leisure park) and Sun Yat-sen University (south, university), (d) Tianhe Sports Center (CBD), (e) Tangxia Village (urban village) and (f) Guangzhou Railway Station and Sanyuanli Village (urban village).

**Table 2.** Contributions of different spatial variables.

| ID | Type | Data source | Spatial variables | Weights |
|---|---|---|---|---|
| 1 | User Density | Tencent RTUD | Average night-time user density (21 pm – 1am +1day) | 13.80% |
| 2 | Density | Baidu POIs | Life service | 13.66% |
| 3 | Density | Baidu POIs | Education facility | 12.40% |
| 4 | Floors and Area | Official statistics | Residential building | 9.85% |
| 5 | Density | Baidu POIs | Clinic facility | 8.61% |
| 6 | Density | Baidu POIs | Residential facility | 8.46% |
| 7 | Distance | Openstreetmap | Road | 4.64% |
| 8 | Density | Baidu POIs | Hotel | 3.90% |
| 9 | Density | Baidu POIs | Business building | 3.26% |
| 10 | Density | Baidu POIs | Restaurant | 2.50% |
| 11 | Density | Baidu POIs | Government facility | 2.48% |
| 12 | Density | Baidu POIs | Location label | 2.45% |
| 13 | Density | Baidu POIs | Financial network-point | 2.21% |
| 14 | Distance | GIS data | Railway | 2.15% |
| 15 | Distance | GIS data | Waterway | 1.49% |
| 16 | Distance | GIS data | Slope | 1.21% |
| 17 | Density | Baidu POIs | Virtual landmark | 1.21% |
| 18 | Density | Baidu POIs | Shopping | 1.07% |
| 19 | Density | Baidu POIs | Automobile service | 1.03% |
| 20 | Density | Baidu POIs | Road tag | 1.02% |
| 21 | Density | Baidu POIs | Corporation | 0.90% |
| 22 | Density | Baidu POIs | Entertainment | 0.79% |
| 23 | Density | Baidu POIs | Working building | 0.55% |
| 24 | Density | Baidu POIs | Scenic spot | 0.27% |
| 25 | Density | Baidu POIs | Greenland | 0.07% |
| 26 | Density | Baidu POIs | Mountain | 0.00% |

matter much less than in the past. Thus, the distance to waterways and railways was not strongly related to the inhabitant distribution, comprising only 1.49% and 2.15% of the contribution, respectively.

Figure 6 shows the fine-scale population distribution. The population census data were downscaled from street-level statistical data to the pixel level at a spatial resolution of 25 m. The center of the old city at the intersection of the Liwan District and Yuexiu District (Figure 6(a)) had the densest population, with 0.22 persons/m$^2$; the density in the shopping center of Haizhu District was also dense (Figure 6(b)). Dongshan Lake Park (Figure 6(c)) is a scenic spot that is surrounded by numerous residential communities. These populated centers with high resident densities are sparsely distributed around the park. Luxury communities in the eastern CBD (Figure 6(d)) had a residential population density of only 0.13 persons/m$^2$, which was only half of that in old cities. Figure 6(e,f) shows that the city villages with the largest scales (Tangxia Village and Sanyuanli Village) had the highest population densities, with means of 0.16 and 0.21 persons/m$^2$, respectively. Thus, the population map from the RFA-based fitting model had a reasonable performance and expressed the spatial heterogeneity in the population distribution. Additionally, the map demonstrated the diverse attractions in different urban areas in the context of urban functions. Moreover, we took some photos in these areas and uploaded them to a website (http://www.geosimulation.cn/gis4win/doc/sta_pop.pdf) because investigating these areas' inner spatial structures with only satellite images was difficult.

### 4.3. *Mapping population distributions at the building level*

The study area contained 460,960 residential buildings in the 'residential community' or 'urban village' categories. The total residential area was 67,419,427 m$^2$. Figure 7, for which the building-population gravity model specified by Equation (6) was used, shows the estimated housing area per capita of all the residential buildings at the center of the old city in Guangzhou, and Table 3 shows the estimated housing area per capita of each building category in different districts. The results indicate that residents who lived in Yuexiu and Liwan, the centers of the old city, had the smallest average housing areas, while residents in Baiyun had the largest housing area because of a number of new residential buildings in the suburban Baiyun District. By repeatedly adjusting $c_i$ in Equation (6), we produced a reasonable population distribution at the building level and computed the densely inhabited index.

Figure 8 shows the densely inhabited index (DII) in each administrative unit at the street level. In the city center of Guangzhou, the DII was much higher than that in other regions (Figure 6(a), DII = 0.56), which suggests that the average housing area of residents in the old district was substantially smaller. This observation can be attributed to the commixture of old residential buildings and dense populations. Old residential buildings in this region were generally built as public houses by enterprises (Danwei) during the past century, and their design was usually low and narrow. With these dense populations, these old residential buildings were largely responsible for the high DII values in this region. Meanwhile, some emerging residential areas near city centers had much lower DIIs because of the high housing prices in busy regions (see Figure 8(b), DII = −0.22 and 8 C, DII = −0.01). Moreover, a large number of migrant laborers have

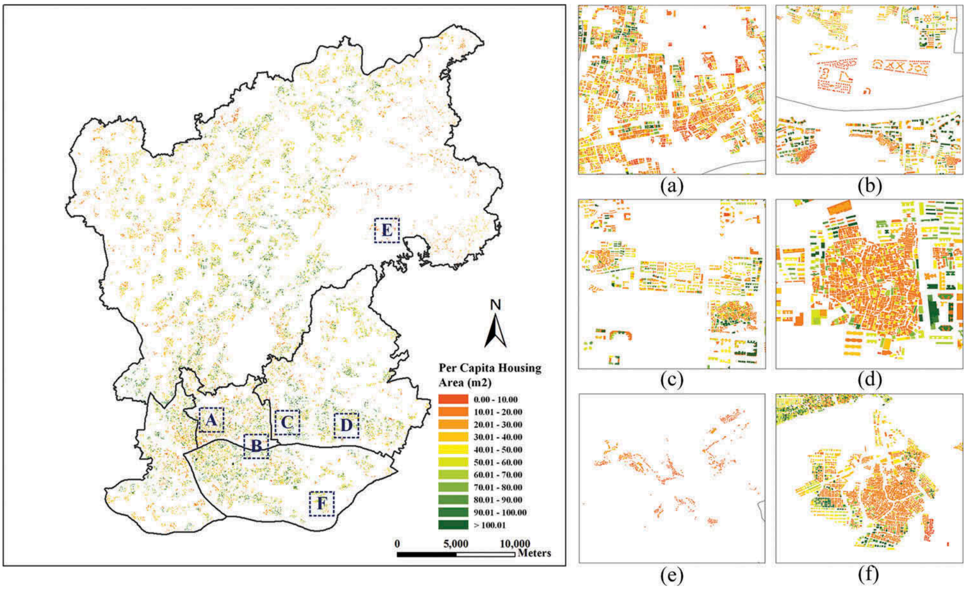


**Figure 7.** Housing area per capita at the building level from the proposed population-building gravity model. (a) Guangdong Province's government (city center of Guangzhou), (b) Er-sha Island (luxury residential area), (c) Flower City Square (CBD), (d) Tangxia Village (urban village), (e) Taihe Town (satellite town) and (f) Xiaozhou Village (rural area).

**Table 3.** Estimated housing area per capita (unit: m$^2$ per capita) in the study area.

| Districts\building type | Residential community | Urban village | Whole buildings |
|---|---|---|---|
| Yuexiu district | 18.8469 | 9.3940 | 16.5331 |
| Liwan district | 24.0715 | 15.9917 | 17.3849 |
| Haizhu district | 29.8832 | 20.6783 | 21.9969 |
| Tianhe district | 37.1856 | 28.2310 | 30.1035 |
| Baiyun district | 55.1041 | 35.2407 | 36.3031 |
| **Whole area** | **34.9973** | **30.5654** | **31.0923** |

flocked to Guangzhou for job opportunities because of this city's rapid development. To seek cheap and affordable residences, these transit populations inhabit villages both on the outskirts and in the downtown portions of the city, forming a unique phenomenon that is known as an 'urban village' (Liu *et al.* 2010).

As the largest urban village in Guangzhou, Tangxia Village is heavily populated by the local population and a migrant, low-income population who are also registered as a portion of the permanent resident population in the census data. The DII of this area was relatively high (DII = 0.19). Notably, the DII values in Taihe Town (Figure 8(e), DII = 3.68) and Xiaozhou Village were considerably higher than the normal value (Figure 8(f), DII = 7.00). The DII values of Taihe and Xiaozhou were both exceptionally larger than 1.00. The migration of labor-intensive enterprises from downtown to the outskirts of cities and the poor residential buildings and infrastructure may explain these outliers. Exorbitant rent in urban Guangzhou drove labor-intensive enterprises to relocate to suburban areas, resulting in significant increases in the number of registered residents because of the household registration system of the People's Republic of China. Meanwhile, the

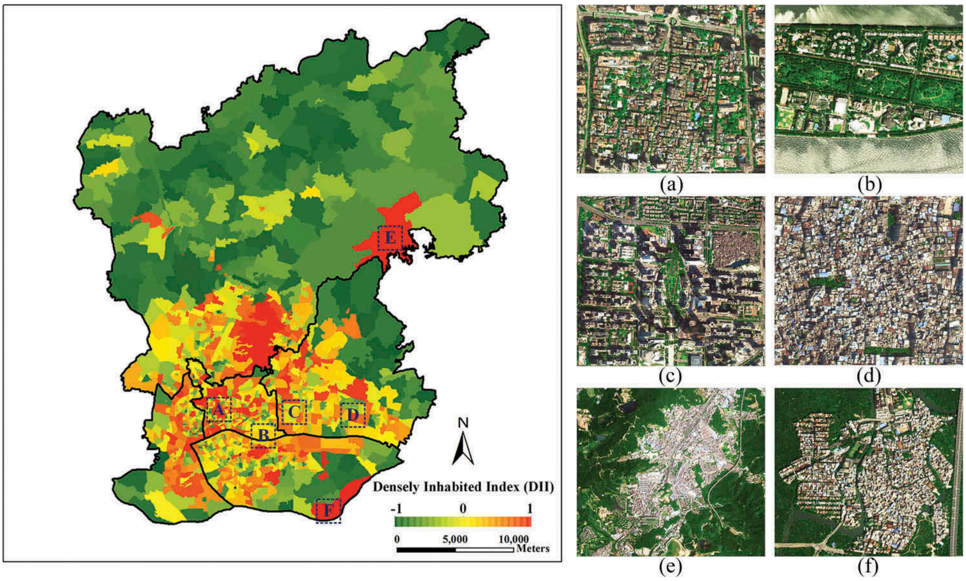


**Figure 8.** Densely inhabited index (DII) in each administrative unit at the street level from the simulated population distribution at the building level. (a) Guangdong Province's government building (city center of Guangzhou), (b) Er-sha Island (luxury residential area), (c) Flower City Square (CBD), (d) Tangxia Village (urban village), (e) Taihe Town (satellite town) and (f) Xiaozhou Village (rural area).

construction of infrastructure and residential buildings in these areas contrasted with the development in urban districts, which was also one of the main reasons for the large DII values.

According to the simulated results, 73.23% of the total population had a per capita living area no greater than that listed in official housing statistical data for the study area (34.4 m$^2$). Additionally, the per capita housing area in the suburban areas of each census unit increased significantly with lower population density and the appearance of new buildings. The average per capita housing area of residents in urban villages was 12.66% less than that of individuals in residential communities; in Yuexiu district, this proportion reached 50.16% (Figure 8 and Table 3). Based on housing statistical data, the estimated housing area per capita was generally in good agreement with the actual living condition of each administrative district. Our proposed gravity model thus provides reasonable simulation results at the building level..

### 4.4. *Accuracy evaluation by a comparison with community-level census data*

Our proposed method had the highest population mapping accuracy of the methods assessed (Table 4, Table 5). As expected, the areal weighting method performed poorly based on population mapping. The multiclass dasymetric mapping method achieved better results than binary dasymetric mapping, since the former is more realistic in considering the multiclass patterns of population distributions. The estimated accuracies of the binary dasymetric mapping and interpolation with cokriging were nearly identical, consistent

**Table 4.** Accuracy comparison of different population mapping methods.

| Methods | Pearson R | Standard $R^2$ | RMSE |
|---|---|---|---|
| Areal weighting | .4396 | .1932 | 873.6086 |
| Binary dasymetric mapping | .5783 | .3344 | 745.1776 |
| Multiclass dasymetric mapping | .6224 | .3874 | 644.3877 |
| Interpolation with cokriging | .4913 | .2413 | 751.2791 |
| ORNL Landscan data (2010) | .1062 | .0113 | 1047.2822 |
| POIs-based population model | .7015 | .4921 | 703.5676 |
| Proposed method | .8615 | .7422 | 663.325 |

**Table 5.** Accuracy assessment of the estimated population distributions among different sampling methods.

| Sampling methods | Pearson R | Standard $R^2$ | RMSE |
|---|---|---|---|
| Spatial sampling on population disaggregation map (Proposed method) | .8615 | .7422 | 663.3250 |
| Random sampling on population disaggregation map | .8141 | .6558 | 726.2745 |
| Spatial sampling on original census data (Deville's method) | .2931 | .0859 | 1101.7862 |

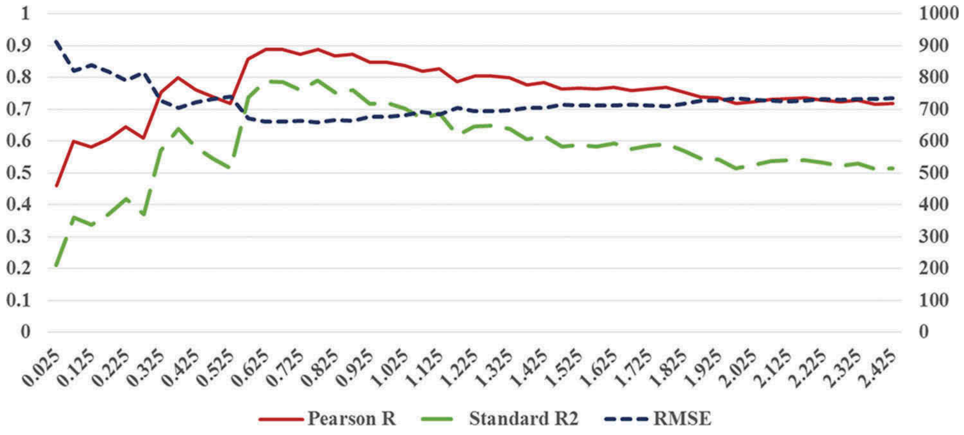


**Figure 9.** Changes in the accuracy indices (*y* axis) for the mapping population distribution according to different sampling radii (*x* axis, unit: kilometers).

with previous studies (Langford 2013, Bakillah *et al.* 2014). Landscan data yielded the lowest accuracy at the local scale. During the process of building the POI-based population model of Bakillah *et al.* (2014), we tuned the non-linear parameter q to .3 and set the PPs of census units as control points. The results showed that the POI-based model obtained the second best population mapping accuracy after our proposed method.

The accuracy of the population distribution first increased and then decreased with increasing sampling radius (Figure 9), and the accuracy peaked when the sampling radius was between 500 m and 750 m.

## 5. Discussion

As shown in the results (Table 5), our proposed method achieved the highest accuracy. The method of (Deville *et al.* 2014) yielded the lowest Pearson correlation coefficient among the methods assessed (.2931). This is because it does not include the spatial heterogeneity of the population distribution in each census unit. The second method assessed also exhibited lower mapping accuracy than the proposed method. This method assumed that people generally dwell in regional centers, when in reality some census units might have complicated spatial structures in which the population spatial distributions do not correspond to an idealized distribution.

Clearly, our proposed method had the highest population mapping accuracy because of the utilization of multisource geospatial big data. In future studies, sources of some official GIS data, such as more detailed building data and population data, will be considered to improve the model. State-of-the-art methods of semiautomatic land use classification and building extraction from high-spatial-resolution stereo image pairs can be introduced into our proposed model for dynamic population mapping at the building level. Finally, more case studies should be conducted in the future because population mapping results are likely to differ in different cities with varying urban spatial structures.

The decreasing trend in the fitting accuracy of the proposed method gradually declined when the sampling distance was larger than 750 m (Figure 9) because the data cleaning process in the sampling method enhanced the reliability of the proposed population-fitting model. These fitting accuracy variations with the sampling radius suggest that the majority of the population was mainly concentrated within 1 km of the PPs in the current census units.

Errors in land use classification can cause some imperfections in population estimation, so we only used artificial surfaces in LULC data sets to estimate preliminary population disaggregation data, eventually resulting in estimation errors for two reasons: (1) artificial surfaces are not the only locations where populations are located; and (2) the spatial heterogeneity is more complex on artificial surfaces, which extends beyond our hypothesis that population distributions obey a normal spatial distribution. Therefore, much finer LULC data are required to build a more complex distribution model and disaggregate census data. Moreover, massive geospatial big data could be input as auxiliary spatial variables and could be easily obtained from the website while building the RFA-based population model. A previous study showed that geospatial big data can reflect the characteristics of the ground surface, such as urban land use types and human activity preferences (Liu *et al.* 2012). However, the computational cost significantly increases with increasing spatial variable inputs, and noise in geospatial big data create instability in the fitting model. A future study should calculate the RFA-based contribution weights of these vast spatial variables. Thus, we can introduce more multisource geospatial big data sets into the proposed population distribution model, explore the driving forces of China's urban population distribution at different scales, and select effective geospatial variables from huge geospatial big data sets via RFA, which is a state-of-the-art feature selection approach that has been recently applied in remote sensing (Ghosh and Joshi 2014).

An iterative gravity model of residential buildings and population was proposed here for the first time. This model efficiently estimated a reasonable population density in every building and study area. The correction factor in the proposed gravity model can be used to estimate reasonable average nonresidential living areas in census units (community level) and reflect the vacancy rate in each unit. However, population accuracy evaluation is difficult at the building level because of a lack of accurate statistical data regarding population density at certain levels. Thus, the accuracy evaluation was chiefly conducted through a comparison of existing models at the minimum scale (community level) with sufficient statistical data. As seen in Table 4, our method produced the highest accuracy (Pearson R = .8615, RMSE = 663.3250, p < .0001) among all the models. Moreover, our building-level population distribution result fit the official per capita housing area reasonably and accurately according to a comparison with official statistical housing data. In future studies, finer statistical population data will be required to calibrate the proposed gravity model. Moreover, the building types are actually related to various preferences that are associated with human activity destinations. Thus, the attractive factors of different building types should be considered in the proposed gravity model to obtain a more accurate population distribution at the building level.

## 6. Conclusion

To our knowledge, no previous studies could allocate population distributions at the building level by using multisource geospatial big data because of a lack of effective models. In this study, we developed a proper framework that was powered by multisource geospatial big data to tackle this problem and thus successfully perform the fine-scale mapping of urban population distributions. By fusing multisource information from official survey data to geospatial big data, this study built a multiscale population model to downscale census data and obtained a high-precision population map at a fine spatial resolution of 25 m. A gravity model between residential buildings and populations was proposed to generate reasonable population distributions at the building level for the first time. According to a comparison with several popular population mapping methods, the proposed method achieved the highest accuracy (Pearson R = .8615, RMSE = 663.3250, p <.0001) at the community level. Moreover, the estimated housing area per capita obtained via proposed model was in good agreement with that of the study area. In future studies, finer population census data and more practical factors will be introduced into our method to validate and improve the gravity model. Additionally, we will combine fine-scale population maps and practical social issues, which can help policymakers optimize resource allocation and determine a more scientific development path in the future.

## Acknowledgment

We really appreciate Prof. May Yuan, Assoc Prof. Shawn Laffan and the two anonymous reviewers for their useful comments and suggestions. This study was supported by the National Natural Science Foundation of China (Grant No. 41671398), the Key National Natural Science Foundation of China (Grant No. 41531176) and the National Natural Science Foundation of China (Grant No. 41601420).

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the National Natural Science Foundation of China [41671398, 41531176 and 41601420].

## ORCID

Yao Yao http://orcid.org/0000-0002-2830-0377
Xia Li http://orcid.org/0000-0003-3050-8529
Jinbao Zhang http://orcid.org/0000-0001-8510-149X
Zhaotang Liang http://orcid.org/0000-0001-9261-5261
Ke Mai http://orcid.org/0000-0002-3532-9872
Yatao Zhang http://orcid.org/0000-0001-5701-2836

## References

Aunan, K. and Wang, S., 2014. Internal migration and urbanization in China: impacts on population exposure to household air pollution (2000–2010). *Science Of The Total Environment*, 481, 186–195. doi:10.1016/j.scitotenv.2014.02.073

Azar, D., *et al.*, 2010. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. *International Journal Of Remote Sensing*, 31 (21), 5635–5655. doi:10.1080/01431161.2010.496799

Bakillah, M., *et al.*, 2014. Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal Of Geographical Information Science*, 28 (9), 1940–1963. doi:10.1080/13658816.2014.909045

Bhaduri, B., *et al.*, 2007a. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69 (1–2), 103–117. doi:10.1007/s10708-007-9105-9

Bhaduri, B., *et al.*, 2007b. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69 (1–2), 103–117. doi:10.1007/s10708-007-9105-9

Biau, G.E.R., 2012. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1), 1063–1095.

Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324

Chan, K.W. and Zhang, L., 1999. The hukou system and rural-urban migration in China: processes and changes. *The China Quarterly*, 160, 818–855. doi:10.1017/S0305741000001351

Chang, X., *et al.*, 2014. Estimating the distribution of economy activity: a case study in Jiangsu Province (China) using large scale social network data. *IEEE*, 1126–1134.

Chen, J., Chen, S., and Landry, P.F., 2013. Migration, environmental hazards, and health outcomes in China. *Social Science Medicine*, 80, 85–95. doi:10.1016/j.socscimed.2012.12.002

Chen, Y., *et al.*, 2016. Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied Geography*, 75, 200–212. doi:10.1016/j.apgeog.2016.08.011

Ciesin, I., 2004. *WRI, 2000. Gridded Population of the World (GPW), version 2*. Palisades, NY: Center for International Earth Science Information Network (CIESIN) Columbia University, International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI).

Ciesin, I., 2005. *CIAT (2005) Global Rural-Urban Mapping Project (GRUMP), alpha version*. Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI).

Co., T., 2015. *Tencent data report of wechat users in the first quarter of 2015, Tencent company (In Chinese)*.

Deichmann, U., 1996. *A review of spatial population database design and modeling*. National Center for Geographic Information and Analysis.

Deville, P., *et al*., 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111 (45), 15888–15893. doi:10.1073/pnas.1408439111

Eicher, C.L. and Brewer, C.A., 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. *Cartography and Geographic Information Science*, 28 (2), 125–138. doi:10.1559/152304001782173727

Fakhraei, S., Soltanian-Zadeh, H., and Fotouhi, F., 2014. Bias and stability of single variable classifiers for feature ranking and selection. *Expert Systems With Applications*, 41 (15), 6945–6958. doi:10.1016/j.eswa.2014.05.007

Fan, C.C., 2008. China on the move: migration, the State, and the Household. *The China Quarterly*, 196, 924–956.

Gaughan, A.E., *et al*., 2013. High resolution population distribution maps for Southeast Asia in 2010 and 2015. *Plos One*, 8 (2), e55882. doi:10.1371/journal.pone.0055882

Ghosh, A. and Joshi, P.K., 2014. A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. *International Journal of Applied Earth Observation and Geoinformation*, 26, 298–311. doi:10.1016/j.jag.2013.08.011

Grafarend, E.W., 2006. *Linear and nonlinear models: fixed effects, random effects, and mixed models*. Walter de Gruyter.

Holt, J.B., Lo, C.P., and Hodler, T.W., 2004. Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31 (2), 103–121. doi:10.1559/1523040041649407

Hu, Q.W., Wang, M., and Li, Q.Q., 2014. Urban hotspot and commercial area exploration with check-in data. *Acta Geodaetica Et Cartographica Sinica*, 43 (3), 314–321.

Hu, T., *et al*., 2016. Mapping urban land use by using landsat images and open social data. *Remote Sensing*, 8 (2), 151. doi:10.3390/rs8020151

Jacobs-Crisioni, C. and Koomen, E., 2012. Linking urban structure and activity dynamics using cell phone usage data. *In: 15th AGILE international conference on Geographic Information Science,* Avignon.

Jones, H.R., 1990. *Population geography*.Guilford Press.

Kang, C., *et al*., 2012. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19 (4), 3–21. doi:10.1080/10630732.2012.715479

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. doi: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. doi:10.1111/gean.2013.45.issue-3

Langford, M., Maguire, D.J., and Unwin, D.J., 1991. The areal interpolation problem: estimating population using remote sensing in a GIS framework. *Handling Geographical Information: Methodology and Potential Applications*, 55–77.

Liu, X., *et al*., 2014. Simulating urban growth by integrating landscape expansion index (LEI) and cellular automata. *International Journal Of Geographical Information Science*, 28 (1), 148–163. doi:10.1080/13658816.2013.831097

Liu, Y., *et al*., 2010. Urban villages under China's rapid urbanization: unregulated assets and transitional neighbourhoods. *Habitat International*, 34 (2), 135–144. doi:10.1016/j.habitatint.2009.08.003

Liu, Y., *et al*., 2015. Social sensing: a new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105 (3), 512–530. doi:10.1080/00045608.2015.1018773

Liu, Y., *et al*., 2012. Urban land uses and traffic "source-sink areas": evidence from GPS-enabled taxi data in Shanghai. *Landscape And Urban Planning*, 106 (1), 73–87. doi:10.1016/j.landurbplan.2012.02.012

Loibl, W. and Peters-Anders, J., 2012. Mobile phone data as source to discover spatial activity and motion patterns. *G1_Forum*, 524–533.

Lu, D., Weng, Q., and Li, G., 2006. Residential population estimation using a remote sensing derived impervious surface approach. *International Journal Of Remote Sensing*, 27 (16), 3553–3570. doi:10.1080/01431160600617202

Lwin, K. and Murayama, Y., 2009. A GIS approach to estimation of building population for micro-spatial analysis. *Transactions in GIS*, 13 (4), 401–414. doi:10.1111/tgis.2009.13.issue-4

Mennis, J., 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55 (1), 31–42.

Palczewska, A., *et al*., 2014. Interpreting random forest classification models using a feature contribution method. *In Integration of Reusable Systemsspringer*, 193–218.

Poston Jr, D.L. and Yaukey, D., 2013. *The population of modern China*. Springer Science & Business Media.

Ratti, C., *et al*., 2006. Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33 (5), 727–748. doi:10.1068/b32047

Sevtsuk, A. and Ratti, C., 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17 (1), 41–60. doi:10.1080/10630731003597322

Stevens, F.R., *et al*., 2015. *Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one*, 10 (2), e0107042.

Sun, J.B., *et al*., 2011. Exploring space-time structure of human mobility in urban space. *Physica A: Statistical Mechanics and Its Applications*, 390 (5), 929–942. doi:10.1016/j.physa.2010.10.033

Tong, W., 2012. The power of social media in china: the government, websites and netizens on weibo. phdtong2012power.

Ural, S., Hussain, E., and Shan, J., 2011. Building population mapping with aerial imagery and GIS data. *International Journal of Applied Earth Observation and Geoinformation*, 13 (6), 841–852. doi:10.1016/j.jag.2011.06.004

Wu, C. and Murray, A.T., 2005. A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems*, 29 (5), 558–579. doi:10.1016/j.compenvurbsys.2005.01.006

Yao, Y., *et al*., 2016. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal Of Geographical Information Science,* 1–24. doi:10.1080/13658816.2016.1244608.

Yue, Y., *et al*., 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. *IEEE*, 1–6.

Zha, Y., Gao, J., and Ni, S., 2003. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal Of Remote Sensing*, 24 (3), 583–594. doi:10.1080/01431160304987

Zhou, Y. and Ma, L.J., 2005. China's urban population statistics: a critical evaluation. *Eurasian Geography and Economics*, 46 (4), 272–289. doi:10.2747/1538-7216.46.4.272

Zhu, Y., 2007. China's floating population and their settlement intention in the cities: beyond the Hukou reform. *Habitat International*, 31 (1), 65–76. doi:10.1016/j.habitatint.2006.04.002