Science of the Total Environment 658 (2019) 936–946

Contents lists available at ScienceDirect

Science of the Total Environment

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

Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model

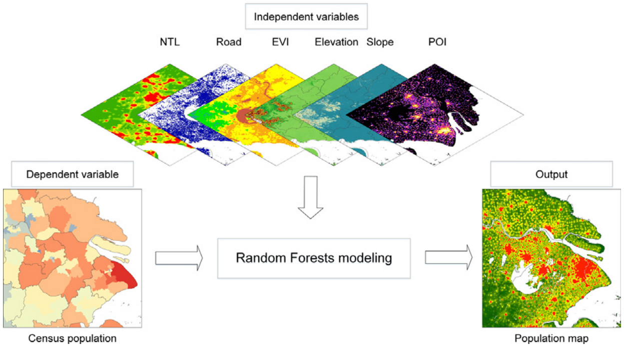
Tingting Ye [a], Naizhuo Zhao [b], Xuchao Yang [a,c,*], Zutao Ouyang [c], Xiaoping Liu [d], Qian Chen [a], Kejia Hu [a], Wenze Yue [e], Jiaguo Qi [c], Zhansheng Li [c,f,**], Peng Jia [g,h]

[a] Ocean College, Zhejiang University, Zhoushan, China
[b] Center for Geospatial Technology, Texas Tech University, Lubbock, TX, USA
[c] Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI, USA
[d] School of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, China
[e] Department of Land Management, Zhejiang University, Hangzhou, China
[f] China University of Geosciences, Wuhan, China
[g] Department of Earth Observation Science, Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands
[h] International Initiative on Spatial Lifecourse Epidemiology (ISLE), Enschede, the Netherlands

HIGHLIGHTS

- A population map for China at 100-m spatial resolution was produced by random forests.
- Remote sensing and POI data were jointly used to disaggregate census population.
- The new population map showed higher accuracy than the Worldop dataset.
- The use of POI reduced under-allocation in urban and over-allocation in rural areas.
- POIs have more strengths than brightness of nighttime lights for population estimation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 September 2018
Received in revised form 18 December 2018
Accepted 18 December 2018
Available online 19 December 2018

Editor: SCOTT SHERIDAN

Keywords:
Points of interest
Population
Random forests
Nighttime light
China

ABSTRACT

Remote sensing image products (e.g. brightness of nighttime lights and land cover/land use types) have been widely used to disaggregate census data to produce gridded population maps for large geographic areas. The advent of the geospatial big data revolution has created additional opportunities to map population distributions at fine resolutions with high accuracy. A considerable proportion of the geospatial data contains semantic information that indicates different categories of human activities occurring at exact geographic locations. Such information is often lacking in remote sensing data. In addition, the remarkable progress in machine learning provides toolkits for demographers to model complex nonlinear correlations between population and heterogeneous geographic covariates. In this study, a typical type of geospatial big data, points-of-interest (POIs), was combined with multi-source remote sensing data in a random forests model to disaggregate the 2010 county-level census population data to 100 × 100 m grids. Compared with the WorldPop population dataset, our population map showed higher accuracy. The root mean square error for population estimates in Beijing, Shanghai, Guangzhou, and Chongqing for this method and WorldPop were 27,829 and 34,193, respectively. The large under-allocation

* Correspondence to: X. Yang, Ocean College, Zhejiang University, Zhoushan Campus, Haike Building 357, 1 Zheda Road, Zhoushan 316021, China.
** Correspondence to: Z. Li, China University of Geosciences, Wuhan 430074, China.
E-mail addresses: yangxuchao@zju.edu.cn (X. Yang), lizhsh@cug.edu.cn (Z. Li).

https://doi.org/10.1016/j.scitotenv.2018.12.276

of the population in urban areas and over-allocation in rural areas in the WorldPop dataset was greatly reduced in this new population map. Apart from revealing the effectiveness of POIs in improving population mapping, this study promises the potential of geospatial big data for mapping other socioeconomic parameters in the future.

1. Introduction

Accurate maps of human population distribution are of critical importance in studies of disaster assessments, public health, and urban planning (Ahola et al., 2007; Aubrecht et al., 2013; Dobson et al., 2000; Hay et al., 2005; Jia et al., 2014). Official population figures derived from census data are typically reported at administrative unit levels (e.g., province/state and city/county). The practicality of such census data is limited, as human populations are not uniformly distributed within administrative units and the administrative boundaries are also changing. Consequently, census data fail to elaborately reveal spatial heterogeneity of population density (Bhaduri et al., 2007; Mao et al., 2017). Due to the lack of explicit and detailed georeferences, combining census population data with georeferenced environmental data is difficult, a problem that impedes interdisciplinary studies in coupled human–environment systems (Zandbergen and Ignizio, 2010). Thus, producing gridded population datasets that can be easily integrated with gridded environmental data is an urgent task.

During the past three decades, various approaches have been developed to spatially disaggregate census population data to grid cells, such as areal weighting (Tobler et al., 1997), pycnophylactic interpolation (Tobler, 1979), dasymetric mapping (Briggs et al., 2007; Su et al., 2010), and intelligent interpolation (Mennis and Hultgren, 2006). These methods have produced many well-known gridded population datasets covering large geographic areas, for example, the Gridded Population of the World (Tobler et al., 1997), the Global Rural Urban Mapping Project urban–rural population (Balk et al., 2006), the LandScan (Bhaduri et al., 2007; Dobson et al., 2000), the WorldPop (Tatem et al., 2013), and the Global Human Settlement Population Grid datasets (European Commission, 2015). During the production of these datasets, satellite image products, e.g., land cover/land use types and nighttime light (NTL) imagery, have been widely adopted as ancillary data to ensure the accuracy of disaggregated population data at relatively fine spatial resolutions (e.g. 1 × 1 km) (Jia and Gaughan, 2016; Li and Zhou, 2018; Sutton, 1997; Wang et al., 2018a; Zandbergen and Ignizio, 2010). However, such remotely-sensed ancillary data at medium spatial resolutions are not directly indicative of land utilization or human presence. They also have limited capabilities in extracting demographic and socioeconomic features related to human activities, specifically in complex urban environments (Liu et al., 2017; Liu et al., 2015; Wu et al., 2009).

The applications of geospatial big data have been greatly popularized in recent years, thereby providing new opportunities to produce accurate gridded population datasets at fine spatial resolutions (Yao et al., 2017b). Points-of-interest (POIs) are a typical kind of geospatial big data. Apart from exact location information (i.e., latitude and longitude), each single POI contains a short textual description to define the category that the POI belongs to (McKenzie et al., 2015; Yoshida et al., 2010). Different categories of POI (e.g., school, bus station, and factory) represent different human activities within and surrounding them, and subsequently have different levels of correlation with population density (Bakillah et al., 2014; Cai et al., 2017). Therefore, elaborate information regarding urban or social systems can be extracted from POIs. Recently, POIs have been utilized to define urban functional districts and land use types (Gao et al., 2017; Hu et al., 2016; Jiang et al., 2015; Liu et al., 2017; Wang et al., 2018b; Yao et al., 2017a; Zhang et al., 2017). Furthermore, each POI holds a point coordinate, thus converting POIs to raster layers with different grid sizes and then combining them with remote sensing data is convenient and flexible (Bakillah et al., 2014).

China has the largest population in the world and its census is conducted every 10 years. The finest-resolution geographic census data that the public can obtain are reported at the county level. Some algorithms for mapping China's population distribution at a 1-km spatial resolution have been developed based on multiple geographic and remotely sensed variables (e.g., elevation, slope, net primary productivity, land use/land cover type, distance to major roads, and brightness of NTL) as distributing weights (Liu et al., 2003; Yue et al., 2003; Yue et al., 2005). Gaughan et al. (2016) applied a random forests (RF) model to map China's population density at 100-m spatial resolution. Complex nonlinear relationships between population density and the multiple geographic and remote sensing variables have been modeled, producing the WorldPop Mainland China dataset with high accuracy at the finest spatial resolution to date (Bai et al., 2018; Gaughan et al., 2016).

The major objective of this study was to jointly use multi-source remote sensing data and POIs to produce a more accurate population map than the WorldPop dataset for China. A previous study suggested that a population dataset with 100-m resolution has specific superiority over other datasets at 1-km resolution (Azar et al., 2013). Moreover, the RF-based methodology has been successfully applied at a 100-m resolution for China (Gaughan et al., 2016). Therefore, we produced a new population density map in 2010 at the 100-m resolution for this country. To fulfill the study objective, we first introduced a method to convert the individual POI of different categories to a raster layer, thus POI data can be jointly used with remote sensing images in an RF model to disaggregate population by county to each gridded area. Next, the allocated population was evaluated using the census data reported at the township level (i.e., Jiedao/Xiangzhen in Chinese). The accuracy of our population map was compared with that of the WorldPop dataset. Then, we discussed the importance of different variables in the RF model for population reallocation. Finally, we analyzed the reasons why the results of our population map were more accurate than the WorldPop dataset.

2. Data and preprocessing

Table 1 lists nine types of data that were used to fit the RF model and evaluate the accuracy of the new RF-based population map. The retrieval and preprocessing of these datasets in the current study are described below.

Table 1
Datasets used in this study.

Dataset	Format	Source
POIs	Point features	Baidu Map Services, China
DMSP-OLS nighttime lights imagery	Grid	The National Oceanic and Atmospheric Administration's National Geophysical Data Center, USA
SPOT NDVI	Grid	Vlaamse Instelling Voor Technologish Onderzoek, Belgium
DEM	Grid	Land Processes Distributed Active Archive Center, USA
Road network	Line features	Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences, China
Census population data (2010)	Table	National Bureau of Statistics of China
Boundary maps	Polygon features	Administration of Surveying Mapping and Geoinformation, China
WorldPop Mainland China dataset	Grid	WorldPop Mainland China dataset for 2010

2.1. Census data

China's population data for 2010 were obtained from the Sixth National Population Census of Mainland China. Hong Kong, Macao, and Taiwan were excluded from this study because of their distinct political and economic status from other Chinese provinces. The census data are reported at the county level (equivalent to the level 3 of the Global Administrative Unit Layer defined by the Food and Agriculture Organization) with 2837 units. We also collected census data at the township level (equivalent to the level 4 of the Global Administrative Unit Layer) with 3155 units from four metropolises (i.e., Beijing, Shanghai, Chongqing, and Guangzhou) with high population densities and 4631 units from four developing provinces (i.e., Jilin, Hubei, Yunnan, and Xinjiang). Following Gaughan et al. (2016), the census data at county level were used to fit the RF model, whereas those at township level were employed to evaluate the accuracy of the output population map.

2.2. Remote sensing datasets

The Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) radiance-calibrated NTL products for 2010 was downloaded from the National Oceanic and Atmospheric Administration's National Centers for Environmental Information (available from https://ngdc.noaa.gov/eog/dmsp/download_radcal.html, last accessed: August 3, 2018). Compared with another widely used type of DMSP-OLS image product (i.e., stable light image composites), radiance-calibrated NTL product comprises different fixed-gain images to avoid the severe saturation problem from emerging in urban areas while maintaining dim lights in suburban/rural areas (Hsu et al., 2015). This image product has a spatial resolution of 1 × 1 km.

The normalized difference vegetation index (NDVI) image products for 2010 were obtained from the Vlaamse Instelling Voor Technologish Onderzoek (available from http://www.vito-eodata.be/, last accessed: August 3, 2018) and collected by the Satellite Pour I'Observation de la Terre (SPOT) Vegetation sensor. This set of SPOT NDVI image products has a spatial resolution of 1 km × 1 km and a temporal granularity of 10 days. The maximum value composite method (Lu et al., 2008), which can be generalized by Eq. (1), was used to generate an annual maximum NDVI image to properly separate human settlements from other land use/land cover types and diminish the influences of cloud contamination:

$$\mathrm{NDVI}_{\max} = \mathrm{MAX}\,(\mathrm{NDVI}_1, \mathrm{NDVI}_2, \ldots, \mathrm{NDVI}_{36}) \tag{1}$$

where NDVI_1, NDVI_2, …, NDVI_{36} are NDVI values of the same pixel in the 36 10-day SPOT image products for 2010.

Following Gaughan et al. (2016), the NTL and the $\mathrm{NDVI}_{\max}$ images were resampled to 100 m using the nearest neighbor approach in ArcGIS 10.4.1 to avoid changing any pixel value during the resampling process.

The Advanced Spaceborne Thermal Emission and Reflection Radiometer digital elevation model (DEM) dataset at 30-m spatial resolution was obtained from the Land Processes Distributed Active Archive Center (available from https://gdex.cr.usgs.gov/gdex/, last accessed: August 3, 2018). These 30-m DEM data were resampled to 100-m using the bilinear interpolation method. Elevation and slope datasets were created using the resampled DEM data.

2.3. POIs

POIs were retrieved from the Baidu Map (http://map.baidu.com), which is the largest desktop and mobile map service provider in China (Yao et al., 2017a, 2017b). We obtained 4,471,696 POI records for 2010 using Baidu Map's application programming interface. Baidu Map classified these POIs into 20 categories on the basis of their Chinese semantic phrase (Yao et al., 2017a, 2017b). Table 2 presents the 20 categories and the number of POI records for each category.

Table 2
Categories of POIs and %IncMSEs used to calculate combining weights with a 5000-m bandwidth of KDE.

Category	Counts	For producing layers of POI density		For producing layers of DtN-POI	
		%IncMSE	Weight	%IncMSE	Weight
Airport	670	2.94	0.01	3.24	0.01
Auto service	149,593	13.27	0.05	16.48	0.06
Bank	301,892	22.6	0.09	15.88	0.06
Commercial building	34,733	12.04	0.05	20.98	0.08
Company	572,129	10.89	0.04	9.26	0.04
Education facility	285,438	20.07	0.08	13.68	0.05
Factory	104,927	14.63	0.06	14.99	0.06
Gas station	86,844	29.68	0.11	20.61	0.08
Government agency	468,794	16.11	0.06	10.48	0.04
Hospital and clinic	175,572	16.63	0.06	29.43	0.11
Hotel	148,156	13.85	0.05	14.45	0.06
Motor passenger station	10,815	12.77	0.05	12.93	0.05
Park	13,041	12.05	0.05	16.98	0.07
Railway station	1864	1.9	0.01	4.75	0.02
Residential community	167,598	16.28	0.06	19.18	0.07
Restaurant and entertainment	781,214	15.31	0.06	15.34	0.06
Retail	1,132,295	13.93	0.05	9.64	0.04
Service zone of Highway	21,873	2.21	0.01	2.61	0.01
Toll station	14,248	13.56	0.05	9.4	0.04
Others	534,357	−0.71	0	−0.15	0

2.4. Boundary and road network data

The boundary map at county and Township levels were obtained from the Administration of Surveying Mapping and Geoinformation, China. Acquired from the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences, the Chinese road network data were used to generate the Euclidean distance to the closest road (DtC-road). Such data include China's national highways, city roads, provincial, county, and township-level roads.

2.5. WorldPop mainland China dataset

The WorldPop Mainland China dataset is a relatively new gridded population dataset and has the finest spatial resolution (i.e., 100-m) for the Chinese territory (Bai et al., 2018; Gaughan et al., 2016). The accuracy of our newly produced population map was compared with that of the WorldPop dataset to highlight the effects of involved POIs in an RF model to improve population mapping. The WorldPop dataset for 2010 was obtained from the WorldPop project website (http://www.worldpop.org.uk/, last accessed: August 3, 2018). To ensure the correctness of area information, all mentioned raster data were re-projected to the Albers Conical Equal Area projection.

3. Method

The workflow for processing POI data, fitting the RF model to produce the dasymetric population map, and assessing accuracy is exhibited in the flowchart in Fig. 1.

3.1. Processing POI data

All POI records in this study were produced to two raster layers of POI density and distance to the nearest POI (DtN-POI) to be jointly used with remote sensing images in the RF model. The kernel density estimation (KDE) (Peng et al., 2016), with a bandwidth initially set as 5000 m, was adopted to convert discrete individual POIs to continued and smooth density surfaces for each of the 20 categories. The density surfaces were output as raster layers at 100 × 100 m spatial resolution. To reduce the computing burden of the final RF model, the 20 raster

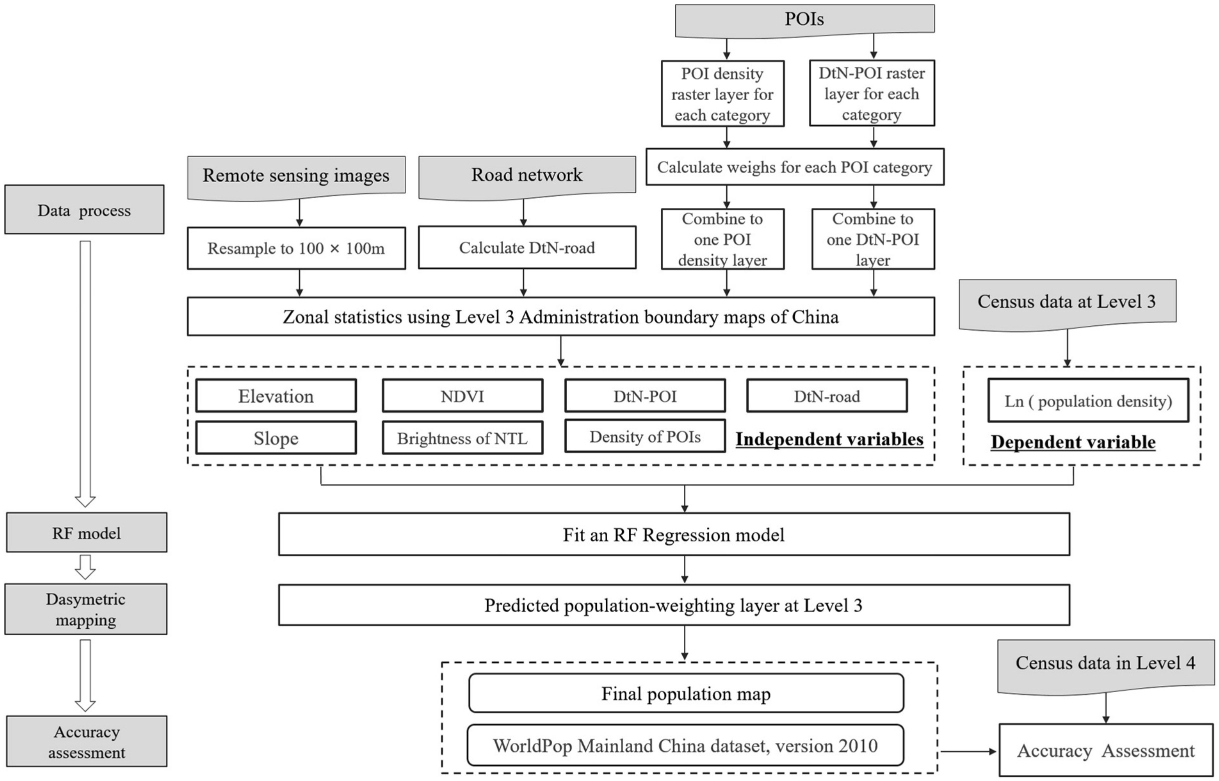

POIs
POI density raster layer for each category
DtN-POI raster layer for each category
Calculate weighs for each POI category
Combine to one POI density layer
Combine to one DtN-POI layer
Remote sensing images
Road network
Data process
Resample to 100 × 100m
Calculate DtN-road
Zonal statistics using Level 3 Administration boundary maps of China
Census data at Level 3
Elevation
NDVI
DtN-POI
DtN-road
Slope
Brightness of NTL
Density of POIs
Independent variables
Ln (population density)
Dependent variable
RF model
Fit an RF Regression model
Dasymetric mapping
Predicted population-weighting layer at Level 3
Census data in Level 4
Final population map
WorldPop Mainland China dataset, version 2010
Accuracy assessment
Accuracy Assessment

Fig. 1. Flowchart for producing and assessing accuracy of the dasymetric population map.

layers for the 20 different POI categories were further combined into one layer. Specifically, each density layer was aggregated by county. Thus, any county held a census population and 20 summed POI densities. These densities were used as the 20 predictor variables of a regression RF model to estimate population at the county level. After developing the RF model, each predictor variable had an output value, namely, %IncMSE, indicating the increase in the mean squared error (MSE) of prediction (i.e., population in this study) after permuting this variable. The higher the value of %IncMSE, the more important the variable in the out-of-bag cross-validation process (Breiman, 2001). Thus, we utilized %IncMSEs to calculate the weights of the 20 individual POI density layers to combine them into one POI density layer, as follows:

$$D_c = D_1W_1 + D_2W_2 + \cdots + D_{20}W_{20} \tag{2}$$

$$W_i = \frac{\%IncMSE_i}{\sum_{k=1}^{20} \%IncMSE_k} \quad (\%IncMSE_k = 0, \text{if} \%IncMSE_k < 0) \tag{3}$$

where D_c denotes the digital number (DN) value of a pixel in the combined POI density layer, D_i ($i = 1, 2, \ldots, 20$) represents the DN value of a pixel in an individual POI density layer for the ith category, and W_i is the weight for the ith POI category.

The combined POI density and additional another five predictor variables (i.e., elevation, slope, brightness of NTL, NDVI, and distance to the closest road) were included in another RF regression model (see Section 3.2 for details of this RF model). Moreover, %IncMSE was obtained for the predictor variable of POI density. We repeated these steps at 100-m increments in the bandwidth of KDE and found that the out-of-bag error was minimized when the bandwidth was 5000 m. Further increasing the bandwidth led to an increase in out-of-bag errors. Therefore, we selected the POI density layer produced by the 5000-m bandwidth of KDE. Table 2 exhibits the specific %IncMSEs of POI densities for the 20 individual categories.

A fishnet with empty attributes at the 100 × 100 m cell size covering the entire mainland China was created in ArcGIS 10.4.1. Each cell was valued by the Euclidean distance from the center of the cell to the nearest POI of a category. We produced a total of 20 raster layers as DtN-POIs for the 20 POI categories. Using the same method of combining the individual POI density layers, the 20 DtN-POI layers were integrated as one raster dataset by their %IncMSEs in an RF regression model for estimating population at the county level. Table 2 displays the specific %IncMSEs used to combine the DtN-POI layers.

3.2. Fitting the RF model and dasymetric population mapping

RF is a classic machine learning approach developed from decision trees (Breiman, 2001). Different from traditional linear regression models, RF is a non-parametric method that can model complex nonlinear relationships between predictions and heterogeneous predictor variables (Hastie et al., 2009). We skipped the formulated details on RF in this research and instead referred to Breiman (2001) and Liaw and Wiener (2002) for such details.

The seven 100 × 100 m raster layers of elevation, slope, brightness of NTL, NDVI, DtC-road, POI density, and DtN POI were aggregated by county and then linked with the natural logarithm of the census population to fit the RF model. Next, the same raster layers were positioned to the fitted RF model to calculate the distribution weight for each 1 ha (i.e., 0.01 km^2) gridded area (see Fig. 2a). Finally, the weights were used to disaggregate the census population at the county level (Fig. 2b) to pixels. A dasymetric population density map (Fig. 3) for mainland China was produced using Eq. 4, as follows:

$$POP_{grid} = \frac{POP_{county} \times W_{grid}}{W_{county}} \tag{4}$$

where W_{grid} is the population-distribution weight for a 1-hectare gridded area, W_{county} denotes the summed population-distribution weight of a county that contains the gridded area, POP_{county} represents the county's census population, and POP_{grid} is the predicted population for the gridded area.

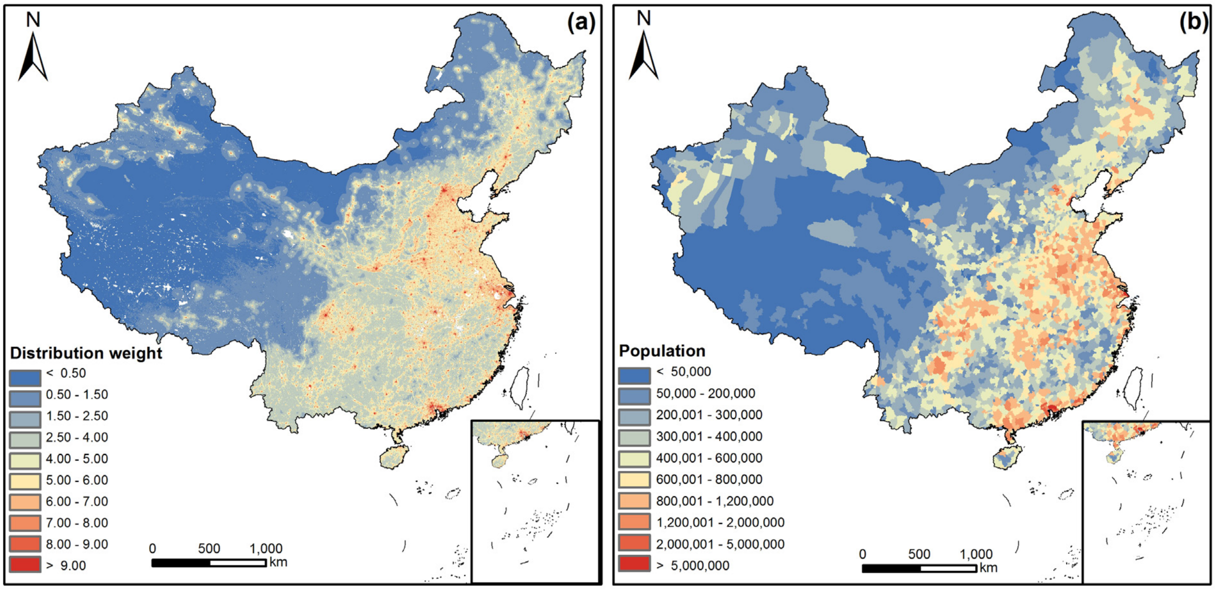


Fig. 2. (a) Distribution weights at 100-m spatial resolution and (b) China census population at the county level.

3.3. Accuracy assessment

The WorldPop dataset was reported as an accurate gridded population dataset with the finest spatial resolution (i.e., 100 m) for China in the current literature (Bai et al., 2018). Beijing, Shanghai, Chongqing, and Guangzhou are the most populated cities in China. Gaughan et al. (2016) used the four cities' census data to evaluate the accuracy of the WorldPop dataset for China. To highlight that our new dasymetric population map (henceforth referred to as PoiPop to distinguish it from the WorldPop map) holds a competitive accuracy at 100-m spatial

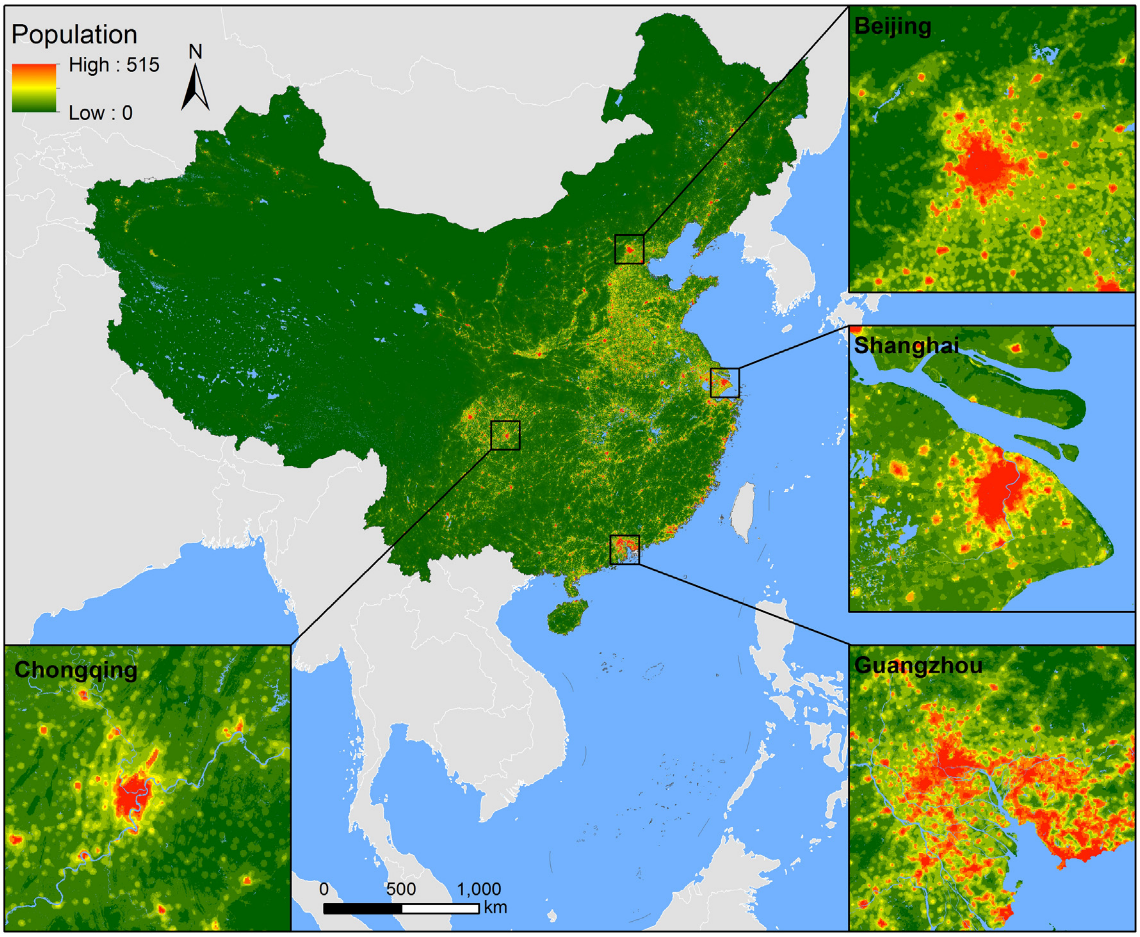


Fig. 3. Predicted people per grid cell (100-m) for 2010 across mainland China.

resolution, we utilized the census data at the township level from the same four cities with high population density (henceforth referred to as Group1) for validation. Four other provinces with relatively low population density (i.e., Jilin, Hubei, Yunnan, and Xinjiang, henceforth referred to as Group2) were also used to elaborately compare accuracy levels between the PoiPop and WorldPop datasets. Two measures, namely, root mean square error (RMSE) and mean absolute deviation (MAE), were selected to quantify and compare the errors of the two gridded population datasets.

4. Results

4.1. Accuracy assessment

The RMSE of PoiPop for Group1 and Group2 was 27,829 and 20,805, respectively, whereas that of WorldPop was correspondingly 34,193 and 24,203. The MAE of PoiPop for Group1 and Group2 was 16,030 and 11,720, respectively, whereas that of WorldPop was correspondingly 20,233 and 13,918. Therefore, PoiPop had a better overall accuracy than WorldPop, as demonstrated by the small RMSE and MAE values.

Fig. 4 illustrates the relationship between the predicted population density and the census population density in which each data point corresponds to a township. The WorldPop had an acceptable accuracy ($R^2 = 0.73$) in regions with medium population densities in Group1. However, such accuracy severely reduced in the highly or lowly populated townships ($R^2 = 0.35$ or $R^2 = 0.15$) (Fig. 4b). The errors were mainly derived from underestimations in townships with large populations (demonstrated by the fact that most red points fell under the diagonal line), whereas overestimations were common in those with small populations (demonstrated by the fact that most blue points fell above the diagonal line) (Fig. 4). Compared with WorldPop, PoiPop held higher accuracy ($R^2 = 0.80$) in the medially populated regions in Group1, and its accuracy in either highly or lowly populated areas ($R^2 = 0.57$ or $R^2 = 0.47$) was increased. Relative to Group1, overestimations in low population regions were not evident in Group2 for PoiPop

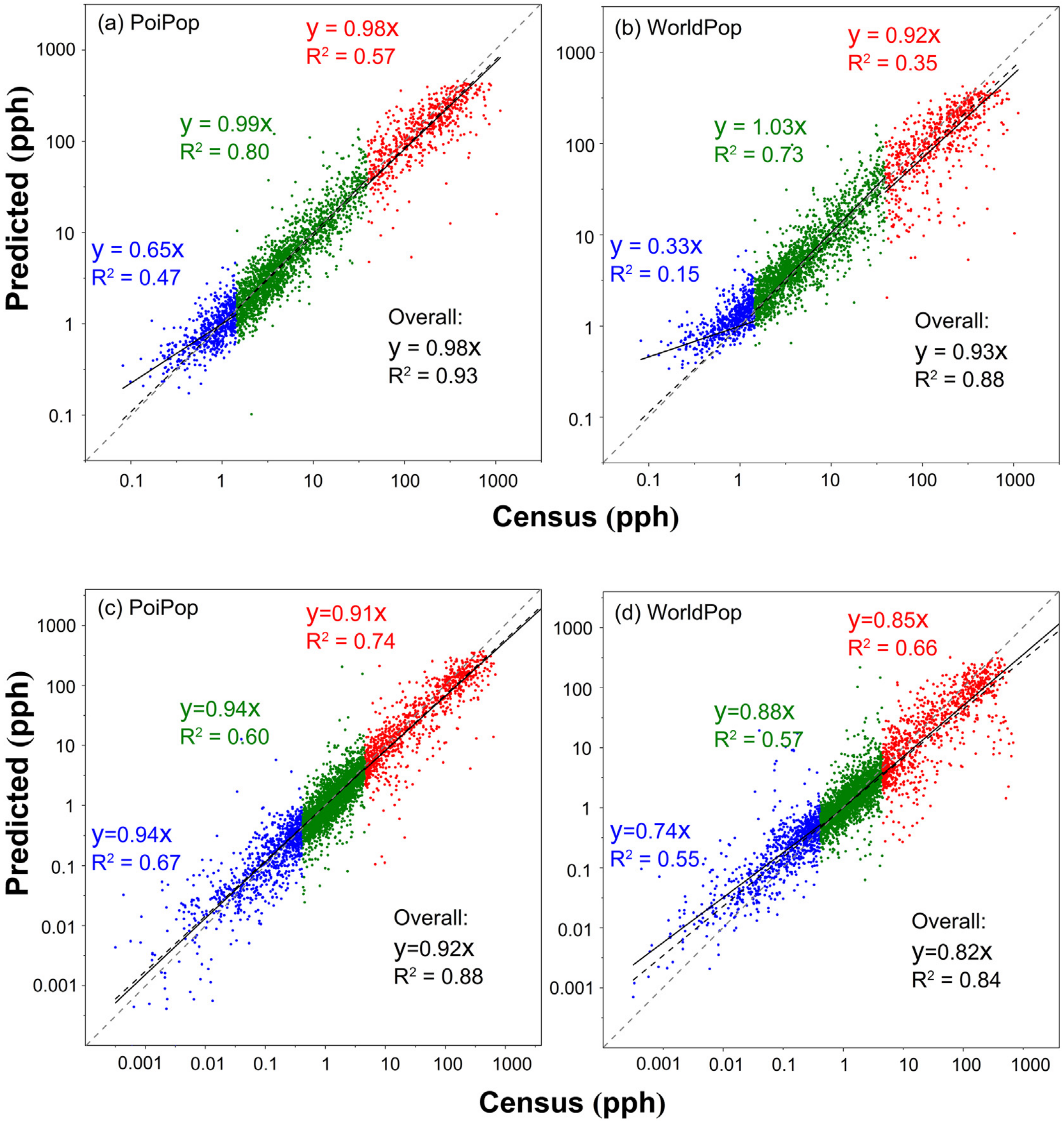


Fig. 4. Scatterplots of the predicted and the census population density at the township level. A $\log_{10}$-$\log_{10}$ transformation was conducted for the population density. The red points indicate the townships with the largest 20% population densities, the blue points correspond to those with the smallest 20% population densities, and the remaining townships are represented by the green points. The black dash line is the global fitting line. (a) and (b) represent the four cities (i.e., Beijing, Shanghai, Chongqing, and Guangzhou) in Group1. (c) and (d) represent the four provinces (i.e., Jilin, Hubei, Yunnan, and Xinjiang) in Group2. pph: population per hectare.

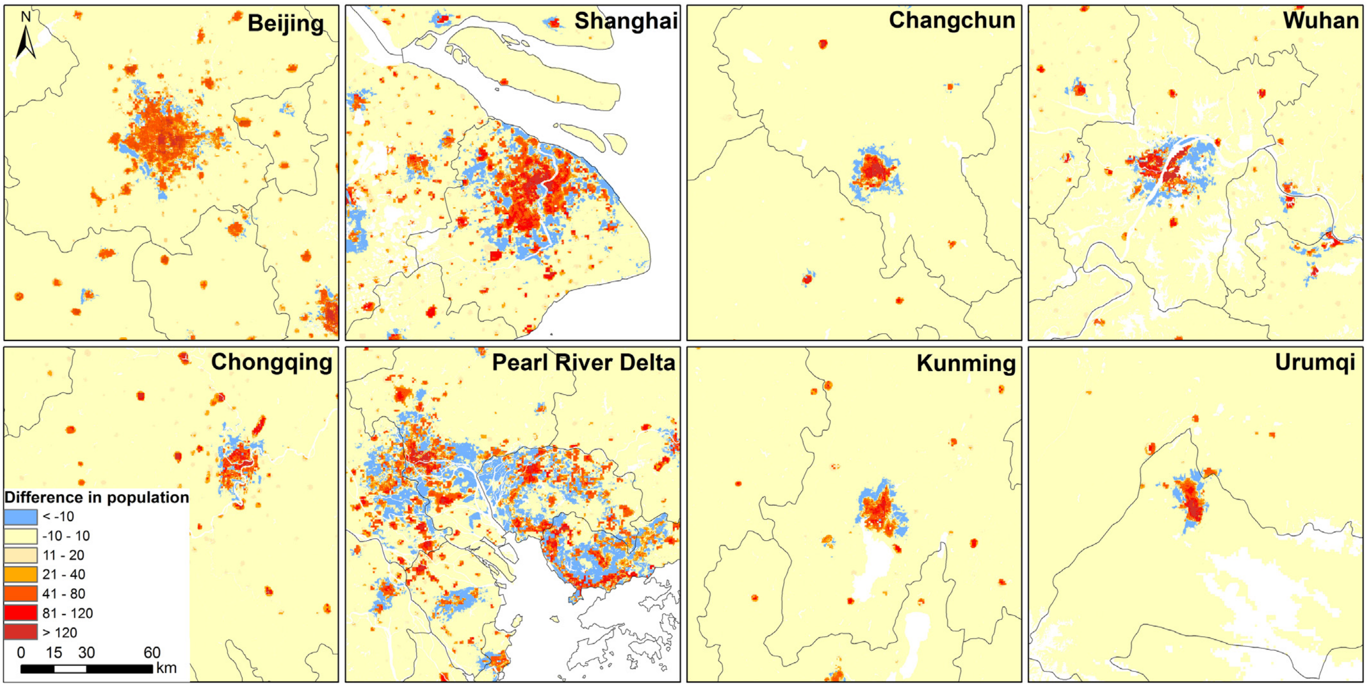


Fig. 5. Differences between PoiPop and WorldPop in eight cities by subtracting the WorldPop dataset from the PoiPop dataset.

and WorldPop (Fig. 4c and d, respectively. The trend lines indicated that the PoiPop product was closer to the one-to-one line than the WorldPop dataset in Group1 and Group2.

Fig. 5 exhibits the differences between the PoiPop and the WorldPop datasets for the four cities in Group1 and four capital cities (i.e., Changchun, Wuhan, Kunming, and Urumqi) of the provinces in Group2. Compared with the WorldPop population map, our newly produced PoiPop map showed larger populations in urban areas and smaller populations in suburban areas, in line with the results in Fig. 4. The PoiPop map benefited from the incorporation of POI-related variables and agreed well with the actual population distribution.

4.2. Responses of population density to the variables

The partial dependency plots (Fig. 6) demonstrate the correlations between population density and geographic factors. Most human settlements distribute in the eastern low-elevation areas of China. The population density steadily decreases as elevation increases to 1300 m. Areas with elevation greater than 3000 m are mostly located in the Qinghai-Tibet Plateau where the population density also decreases with the rise in elevation until the elevation reaches 4000 m. Humans live in flat areas, and population density decreases with increased slope until approximately 20%.

NDVI reflects vegetation coverage on the earth's surface. Human beings also live in areas with plenty of water resources and plants (Nieves et al., 2017). Hence, as the NDVI value of a region rises, the population density also steadily increases. However, when NDVI reaches 0.8, the population density sharply decreases with any further increase in NDVI. Regions with an NDVI value larger than 0.8 can be defined as dense forest (Zhuo et al., 2009).

In this study, we selected radiance-calibrated NTL instead of the widely used stable NTL image product. This selection ensures sufficient variation in the brightness of NTL in urban areas by avoiding appearances of saturated pixels (Hsu et al., 2015). As the partial dependency plot demonstrates, the largest DN value in the 2010 radiance-calibrated NTL image product was greater than 1000. However, most (93%) of the lit pixels' DN values were lower than 100. Thus, population density steadily increased with rising brightness of NTL. However, when the brightness of NTL reached 100, population density remained nearly unchanged with further increase in NTL brightness. Similarly, population density steadily increased with the increments in the POI density until it reached 3.5. Then, population density remained unchanged.

Population density logarithmically decreased with increasing distance to the nearest road or POI (see Fig. 6). Moreover, 97% (91%) of Chinese people live in areas with a distance to the nearest road (POI) shorter than 25 km (12 km). Thus, population density was nearly unvaried with the change after such a distance.

5. Discussion

5.1. POI versus brightness of NTL

The partial dependency plots (Fig. 6) show that the relationships between population density and geographic variables were nonlinear and/or piecewise. Thus, compared with the geographically weighted regression models, which were mainly dependent on linear regressions in previous population studies (Chen et al., 2007; Xu and Ouyang, 2018), the machine learning method of RF can more accurately capture the complex correlations between population density and geographic variables (Liu et al., 2018b). PoiPop and WorldPop were produced by the same method (i.e., RF), census data (i.e., the Sixth National Population Census of Mainland China), and geographic variables of elevation, slope, and brightness of NTL. The major difference between PoiPop and WorldPop was the use of additional POI-related variables (i.e., density of POI and DtN-POI). Aside from slope, the two POI-related variables had the largest contributions to modeling population density (Fig. 7). Specifically, the %IncMSEs of the two variables were larger than that of brightness of NTL, which was a good measure of population density according to a number of previous studies (Lo, 2001; Sutton et al., 2001; Sutton et al., 1997; Zhuo et al., 2009).

The basic logic of using brightness of NTL to map or distribute population was that a region with bright lights at night typically has a large population (Sutton et al., 2001). However, blooming is inherent to NTL, demonstrating that urban peripheries are brightened by urban lights (Imhoff et al., 1997). Liu et al. (2016) found that in China less than 5% of lit areas were developed. Therefore, the lit area from the

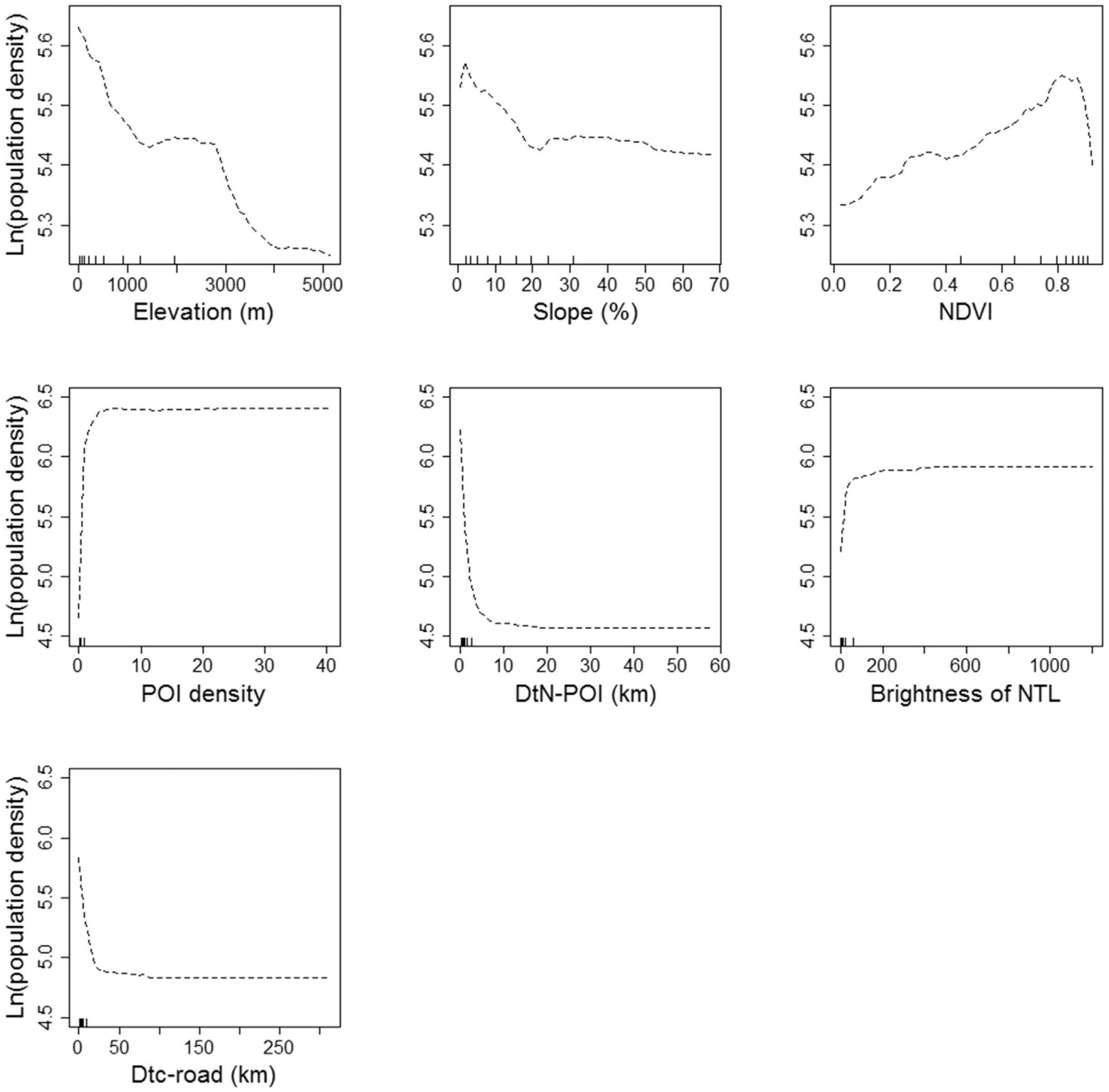


Fig. 6. Partial dependency plots for the variables in the RF model predicting population density. The black ticks at the base of the plots are deciles of the input variable. Tick marks at the x-axis indicate the deciles (10% quantiles) of the observed distribution of continuous predictor variables.

NTL image was much larger than the developed area extracted from the Landsat image (Fig. 8). The West Lake and Qiantang River inside or close to the urban area of Hangzhou were well-lit and held considerably large DN values in the NTL image. Thus, when brightness of NTL was used as the only or chief variable indicating the status of human systems, a considerable amount of the population was allocated to undeveloped regions, thereby leading to over-distributions in rural and suburban areas (i.e., the lowly populated townships) and under-distributions in urban areas (i.e., the highly populated townships) (see Fig. 4). In addition, NTL within a small land area inside a city can brighten a huge surrounding area (Esch et al., 2014). In an NTL image, a pixel's brightness of NTL is usually a combined outcome of the brightness of its adjacent pixels. Thus, variation in the brightness of NTL can be small, as demonstrated by the relatively flat curve of DN values across the urban area (Fig. 8). These reasons led to the failure of brightness of NTL collected by satellite images to accurately reflect population densities at a relatively small geographic scale.

In modern societies, human existence inevitably generates demand for different kinds of services, driving the appearance of different service entities (e.g., gas stations, convenience stores, schools, and hospitals). Moreover, the larger the population, the greater the demand for such service entities. Hence, a region with more POIs or closer distance to POIs has a larger population than its counterparts. For example, the primary factor determining the construction of an additional school or hospital is its surrounding population. The number of gas stations or convenience stores in an area is determined by its purchasing power, which is mainly affected by population. A region without POIs or is far from POIs should be allocated with a low population although the region is lit in the NTL imagery. Moreover, POI-related variables can properly capture fabric and function inside of cities (Gao et al., 2017; Jiang et al., 2015), indicated by the dramatic changes of the POI density across urban areas (Fig. 8). Therefore, adding POI-related variables can greatly enhance the variation of the distributed population and reduce the underestimation of populations in urban areas, because a considerable proportion of population was not allocated to suburban and rural areas where pixels were lit but corresponded to small POI densities.

Furthermore, populations inside or surrounding different categories of POIs are dramatically different. Certain categories of POI, such as

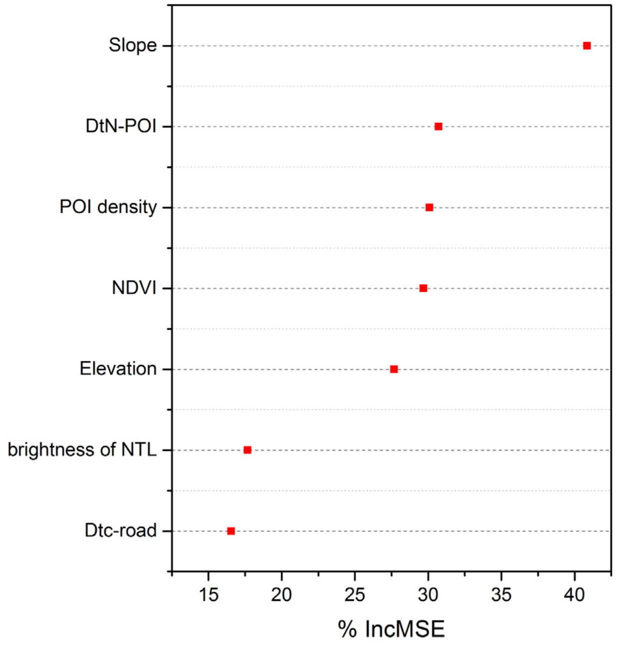


Fig. 7. Percentage-increased MSE indicates the variable importance for RF regression.

hospitals and schools, are correlated with high population density (Bakillah et al., 2014). Baidu's POI data contain semantic information on categories of POIs that cannot be extracted from a satellite image and consequently is a distinctive feature of POI data superior to remote sensing imagery. In this study, POIs belonging to disparate categories were allocated different weights, thereby allowing combined POI density or DtN-POI to more accurately represent population density than NTL brightness.

5.2. Uncertainty of the POI data

At present, the most widely used POIs (e.g., OpenStreetMap's POIs) are a typical type of volunteered geographic information (VGI) data (Bakillah et al., 2014). Uncertainty over the data quality of such POI datasets has adversely affected their uses in practical studies (Fonte et al., 2017). These uncertainties are derived not only from positional and thematic accuracies but also from spatial biases in the information. Previous studies (e.g., Ma et al., 2015; Neis and Zielstra, 2014) indicated that most POIs are collected in urban areas, whereas many POIs in rural areas are unreported. Moreover, in VGI-based POI datasets, popular service entities (e.g., restaurants and tourist attractions) attract more attention. By contrast, POIs are considerably insufficient in unobtrusive urban areas (e.g., residential areas) (Antoniou and Schlieder, 2014). Thus, such spatial biases can generate severe errors in distributing census populations.

To avoid such problems, we adopted the commercial POI data acquired from Baidu. Collected by trained persons and after undergoing strict inspections and corrections, positional and thematic accuracies of these commercial POI data are found reliable. These POI data are also used in Baidu's navigation software and consequently, spatial biases have been greatly controlled although these spatial biases cannot be thoroughly eradicated.

Aside from the geospatial big data of POIs, location-based social media data such as those from Twitter (Patel et al., 2017) and WeChat (Yao et al., 2017b) have also been used to improve population mapping. These social media data have considerable biases among subpopulations of different age groups. For example, young people more likely post social media data than seniors (Jiang et al., 2018). Therefore, when the location-based social media data are used to disaggregated census populations, rural areas may be under-allocated because many young countrymen migrate to urban areas for work or to receive higher education, leaving elderly people. In addition, the volume of location-based social media data is greatly influenced by the prevalence of computers and smartphones. In the economically developed eastern areas of

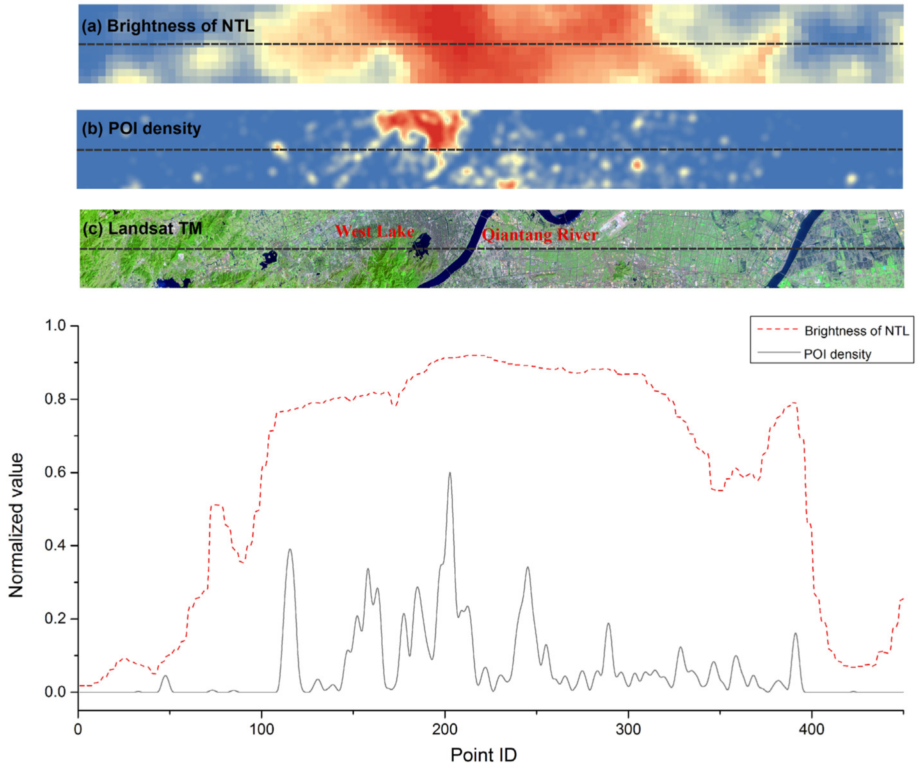


Fig. 8. Latitudinal transects of NTL and the POI density raster layer for the city of Hangzhou, China.

China, nearly every family or adult has a computer or smartphone, yet in the economically repressed western areas, computers and smartphones are still not prevalent. Therefore, large spatial biases likely exist within the Chinese social media data and generate considerable adverse effects on the accuracy of population distribution.

To sum up, commercial POI data (e.g. Baidu's POIs used in this study) have relatively fewer uncertainties and smaller spatial biases than VGI-based POIs and social media big data. Thus, commercial POI data are suitable in mapping population distribution at fine spatial resolutions and across large geographic scales.

6. Conclusions

In the study, we used multiple kinds of remote sensing image products and Baidu's POIs within an RF model to spatially disaggregate census data to produce a gridded population map for China at 100-m spatial resolution. Our population map showed higher accuracy than the WorldPop dataset. The considerably evident underestimations in urban and overestimations in rural and suburban areas that once existed in the WorldPop dataset were markedly reduced in the new PoiPop date due to the integrated of POIs. In the RF model, brightness of NTL, which was extensively believed to be a good proxy of population distribution, showed less importance than the two POI-related variables (i.e. POI density and DtN-POI). Compared with brightness of NTL, POI density had more adequate variation across urban areas and did not have overly high values in undeveloped areas (i.e., no blooming effects).

Apart from distributing census population, brightness of NTL has also been extensively used as an indicator of socioeconomic factors (e.g., GDP, electric power consumption, and CO_2) (Chen and Nordhaus, 2011; Doll et al., 2006; He et al., 2014; Liu et al., 2018a; Ou et al., 2015). Although NTL can generally delimit the extent of human activities and indicate intensities of these activities, categories of such activities cannot be discerned. For example, two regions' total brightness of NTL may be the same, despite one region being a residential area and the other being a business district. In such cases, the business district should have more GDP than the residential area although they have the same total brightness of NTL. POI data contain semantic information, thus showing locations at which wealth is produced and indicating kinds of industry from which wealth is derived. This characteristic shows great potential for POI data to be used in spatially disaggregating other socioeconomic parameters in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 41671035), the Fundamental Research Funds for the Central Universities, and the Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University (Grant No. 2017LSDMIS08). The authors thank Editor Scott C. Sheridan and two anonymous reviewers for their constructive comments.

References

Ahola, T., Virrantaus, K., Krisp, J.M., Hunter, G.J., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision-making. Int. J. Geogr. Inf. Sci. 21 (8), 935–953.

Antoniou, V., Schlieder, C., 2014. Participation patterns, VGI and gamification. Proceedings of AGILE 2014. Presented at the AGILE, pp. 3–6.

Aubrecht, C., Özceylan, D., Steinnocher, K., Freire, S., 2013. Multi-level geospatial modeling of human exposure patterns and vulnerability indicators. Nat. Hazards 68, 147–163.

Azar, D., Engstrom, R., Graesser, J., Comenetz, J., 2013. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. Remote Sens. Environ. 130 (0), 219–232.

Bai, Z., Wang, J., Wang, M., Gao, M., Sun, J., 2018. Accuracy assessment of multi-source gridded population distribution datasets in China. Sustain. For. 10 (5), 1363.

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., Zipf, A., 2014. Fine-resolution population mapping using OpenStreetMap points-of-interest. Int. J. Geogr. Inf. Sci. 28 (9), 1940–1963.

Balk, D.L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S.I., Nelson, A., 2006. In: Hay, S.I., Graham, A., Rogers, D.J. (Eds.), Determining Global Population Distribution: Methods, Applications and Data. Adv. Parasit. 62. Academic Press, pp. 119–156.

Bhaduri, B., Bright, E., Coleman, P., Urban, M.L., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69 (1), 103–117.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Briggs, D.J., Gulliver, J., Fecht, D., Vienneau, D.M., 2007. Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote Sens. Environ. 108 (4), 451–466.

Cai, J., Huang, B., Song, Y., 2017. Using multi-source geospatial big data to identify the structure of polycentric cities. Remote Sens. Environ. 202, 210–221.

Chen, X., Nordhaus, W.D., 2011. Using luminosity data as a proxy for economic statistics. Proc. Natl. Acad. Sci. 108 (21), 8589–8594.

Chen, M., Xu, C., Wang, R., 2007. Key natural impacting factors of China's human population distribution. Popul. Environ. 28 (3), 187–200.

Dobson, J.E., Bright, E.A., Colemen, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogramm. Eng. Remote. Sens. 66 (7), 849–857.

Doll, C.N.H., Muller, J.P., Morley, J.G., 2006. Mapping regional economic activity from night-time light satellite imagery. Ecol. Econ. 57 (1), 75–92.

Esch, T., Marconcini, M., Marmanis, D., Zeidler, J., Elsayed, S., Metz, A., et al., 2014. Dimensioning urbanization – an advanced procedure for characterizing human settlement properties and patterns using spatial network analysis. Appl. Geogr. 55, 212–228.

European Commission, C. U. f. I.-C, 2015. GHS Population Grid, Derived From GPW4, Multitemporal [1975, 1990, 2000, 2015]. European Commission, Joint Research Centre, Brussels, Belgium See. http://data.europa.eu/89h/jrc-ghslghs_pop_gpw4_globe_r2015a.

Fonte, C.C., Antoniou, V., Bastin, L., Bayas, L., See, L., Vatseva, R., 2017. Assessing VGI data quality. In: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V. (Eds.), Mapping and the Citizen Sensor, pp. 137–163.

Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467.

Gaughan, A.E., Stevens, F.R., Huang, Z., Nieves, J.J., Sorichetta, A., Lai, S., et al., 2016. Spatiotemporal patterns of population in mainland China, 1990 to 2010. Sci. Data 3, 160005.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. second edition. 2009. Springer Science+Business Media, LLC, New York, NY.

Hay, S.I., Noor, A.M., Nelson, A., Tatem, A.J., 2005. The accuracy of human population maps for public health application. Tropical Med. Int. Health 10 (10), 1073–1086.

He, C., Ma, Q., Liu, Z., Zhang, Q., 2014. Modeling the spatiotemporal dynamics of electric power consumption in Mainland China using saturation-corrected DMSP/OLS nighttime stable light data. Int. J. Digital Earth 7 (12), 993–1014.

Hsu, F.-C., Baugh, K., Ghosh, T., Zhizhin, M., Elvidge, C., 2015. DMSP-OLS radiance calibrated nighttime lights time series with intercalibration. Remote Sens. 7 (2), 1855.

Hu, T., Yang, J., Li, X., Gong, P., 2016. Mapping urban land use by using Landsat images and open social data. Remote Sens. 8 (2), 151.

Imhoff, M.L., Lawrence, W.T., Stutzer, D.C., Elvidge, C.D., 1997. A technique for using composite DMSP/OLS "city lights" satellite data to map urban area. Remote Sens. Environ. 61 (3), 361–370.

Jia, P., Gaughan, A.E., 2016. Dasymetric modeling: a hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. Appl. Geogr. 66, 100–108.

Jia, P., Qiu, Y., Gaughan, A.E., 2014. A fine-scale spatial population distribution on the high-resolution gridded population surface and application in Alachua County, Florida. Appl. Geogr. 50 (Supplement C), 99–107.

Jiang, S., Alves, A., Rodrigues, F., Ferreira Jr., J., Pereira, F.C., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput. Environ. Urban. Syst. 53, 36–46.

Jiang, Y., Li, Z., Ye, X., 2018. Understanding demographic and socioeconomic biases of geotagged twitter users at the county level. Cartogr. Geogr. Inf. Sci. 1–15.

Li, X., Zhou, W., 2018. Dasymetric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data. Sci. Total Environ. 643, 1248–1256.

Liaw, A., Wiener, M., 2002. Classification and Regression by RandomForest. 3. R News, pp. 18–22.

Liu, J., Yue, T., Wang, Y., Qiu, D., Liu, M., Deng, X., et al., 2003. Digital simulation of population density in China. Acta Geograph. Sin. 58 (1), 17–24.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al., 2015. Social sensing: a new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105 (3), 512–530.

Liu, Y., Delahunty, T., Zhao, N., Cao, G., 2016. These lit areas are undeveloped: delimiting China's urban extents from thresholded nighttime light imagery. Int. J. Appl. Earth Obs. Geoinf. 50, 39–50.

Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., et al., 2017. Classifying urban land use by integrating remote sensing and social media data. Int. J. Geogr. Inf. Sci. 31 (8), 1675–1696.

Liu, X., Ou, J., Wang, S., Li, X., Yan, Y., Jiao, L., et al., 2018a. Estimating spatiotemporal variations of city-level energy-related CO_2 emissions: an improved disaggregating model based on vegetation adjusted nighttime light data. J. Clean. Prod. 177, 101–114.

Liu, Y., Cao, G., Zhao, N., Mulligan, K., Ye, X., 2018b. Improve ground-level $PM_{2.5}$ concentration mapping using a random forests-based geostatistical approach. Environ. Pollut. 235, 272–282.

Lo, C.P., 2001. Modeling the population of China using DMSP operational linescan system nighttime data. Photogramm. Eng. Remote. Sens. 67 (9), 1037–1047.

Lu, D., Tian, H., Zhou, G., Ge, H., 2008. Regional mapping of human settlements in southeastern China with multisensor remotely sensed data. Remote Sens. Environ. 112 (9), 3668–3679.

Ma, D., Sandberg, M., Jiang, B., 2015. Characterizing the heterogeneity of the OpenStreetMap data and community. ISPRS Int. J. Geo-Info. 4 (2), 535.

Mao, H., Ahn, Y.-Y., Bhaduri, B., Thakur, G., 2017. Improving land use inference by factorizing mobile phone call activity matrix. J. Land Use Sci. 12 (2–3), 138–153.

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., Hu, Y., 2015. POI pulse: a multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. Cartographica 50 (2), 71–85.

Mennis, J., Hultgren, T., 2006. Intelligent dasymetric mapping and its application to areal interpolation. Cartogr. Geogr. Inf. Sci. 33 (3), 179–194.

Neis, P., Zielstra, D., 2014. Recent developments and future trends in volunteered geographic information research: the case of OpenStreetMap. Future Internet 6 (1), 76.

Nieves, J.J., Stevens, F.R., Gaughan, A.E., Linard, C., Sorichetta, A., Hornby, G., et al., 2017. Examining the correlates and drivers of human population distributions across low- and middle-income countries. J. R. Soc. Interface 14 (137).

Ou, J., Liu, X., Li, X., Li, M., Li, W., 2015. Evaluation of NPP-VIIRS nighttime light data for mapping global fossil fuel combustion CO_2 emissions: a comparison with DMSP-OLS nighttime light data. PLoS One 10 (9), e0138310.

Patel, N.N., Stevens, F.R., Huang, Z., Gaughan, A.E., Elyazar, I., Tatem, A.J., 2017. Improving large area population mapping using Geotweet densities. Trans. GIS 21 (2), 317–331.

Peng, J., Zhao, S., Liu, Y., Tian, L., 2016. Identifying the urban-rural fringe using wavelet transform and kernel density estimation: a case study in Beijing City, China. Environ. Model Softw. 83, 286–302.

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. Sci. Total Environ. 408 (20), 4807–4816.

Sutton, P., 1997. Modeling population density with night-time satellite imagery and GIS. Comput. Environ. Urban. Syst. 21 (3–4), 227–244.

Sutton, P., Roberts, D., Elvidge, C., Meij, H., 1997. A comparison of nighttime satellite imagery and population density for the continental United States. Photogramm. Eng. Remote. Sens. 63 (11), 1303–1313.

Sutton, P., Roberts, D., Elvidge, C., Baugh, K., 2001. Census from heaven: an estimate of the global human population using night-time satellite imagery. Int. J. Remote Sens. 22 (16), 3061–3076.

Tatem, A.J., Gaughan, A.E., Stevens, F.R., Patel, N.N., Jia, P., Pandey, A., et al., 2013. Quantifying the effects of using detailed spatial demographic data on health metrics: a systematic analysis for the AfriPop, AsiaPop, and AmeriPop projects. Lancet 381, S142.

Tobler, W.R., 1979. Smooth pycnophylactic interpolation for geographical regions. J. Am. Stat. Assoc. 74 (367), 519–530.

Tobler, W., Deichmann, U., Gottsegen, J., Maloy, K., 1997. World population in a grid of spherical quadrilaterals. Int. J. Popul. Geogr. 3 (3), 203–225.

Wang, L., Wang, S., Zhou, Y., Liu, W., Hou, Y., Zhu, J., et al., 2018a. Mapping population density in China between 1990 and 2010 using remote sensing. Remote Sens. Environ. 210, 269–281.

Wang, Y., Gu, Y., Dou, M., Qiao, M., 2018b. Using spatial semantics and interactions to identify urban functional regions. ISPRS Int. J. Geo-Info. 7 (4), 130.

Wu, S.-S., Qiu, X., Usery, E.L., Wang, L., 2009. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. Ann. Assoc. Am. Geogr. 99 (1), 76–98.

Xu, Z., Ouyang, A., 2018. The factors influencing China's population distribution and spatial heterogeneity: a prefectural-level analysis using geographically weighted regression. Appl. Spat. Anal. Policy 11 (3), 465–480.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., et al., 2017a. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 31 (4), 825–848.

Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., et al., 2017b. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. Int. J. Geogr. Inf. Sci. 31 (6), 1220–1244.

Yoshida, D., Song, X., Raghavan, V., 2010. Development of track log and point of interest management system using free and open source software. Appl. Geomatics 2 (3), 123–135.

Yue, T.X., Wang, Y.A., Chen, S.P., Liu, J.Y., Qiu, D.S., Deng, X.Z., et al., 2003. Numerical simulation of population distribution in China. Popul. Environ. 25 (2), 141–163.

Yue, T.X., Wang, Y.A., Liu, J.Y., Chen, S.P., Qiu, D.S., Deng, X.Z., et al., 2005. Surface modelling of human population distribution in China. Ecol. Model. 181 (4), 461–478.

Zandbergen, P.A., Ignizio, D.A., 2010. Comparison of dasymetric mapping techniques for small-area population estimates. Cartogr. Geogr. Inf. Sci. 37 (3), 199–214.

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., Wang, H., 2017. The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: a case study in Beijing, China. Remote Sens. 9 (9), 865.

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. Int. J. Remote Sens. 30 (4), 1003–1018.