

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Mapping monthly population distribution and variation at 1-km resolution across China

Zhifeng Cheng, Jianghao Wang & Yong Ge

To cite this article: Zhifeng Cheng, Jianghao Wang & Yong Ge (2022) Mapping monthly population distribution and variation at 1-km resolution across China, International Journal of Geographical Information Science, 36:6, 1166-1184, DOI: 10.1080/13658816.2020.1854767

To link to this article: https://doi.org/10.1080/13658816.2020.1854767

Published online: 07 Dec 2020.

Submit your article to this journal

Article views: 1094

View related articles

View Crossmark data

Citing articles: 1 View citing articles

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

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2022, VOL. 36, NO. 6, 1166–1184
https://doi.org/10.1080/13658816.2020.1854767

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Mapping monthly population distribution and variation at 1-km resolution across China

Zhifeng Cheng [a,b], Jianghao Wang [a,b] and Yong Ge [a,b]

[a]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [b]University of Chinese Academy of Sciences, Beijing, China

ABSTRACT

Fine-grained inner-annual population data are instrumental in climate change response, resource allocation, and epidemic control. However, such data are currently scarce due to the lack of human-related indicators with both high temporal resolution and long-term coverage that can be used in the process of population spatialization. Here, we estimate monthly 1-km gridded population distribution across China in 2015 using time-series mobile phone positioning data. We construct a hybrid downscaling model to map the gridded population by incorporating random forest and area-to-point kriging. The estimated monthly population products appear to capture inner-annual population variations, especially during special periods, such as the festival, holiday, and short-term labor flow period, which are characterized by large-scale population movements. Additionally, compared with census data, the hybrid model-based results obtained exhibit higher consistency than popular global population products across all spatial extents. Our monthly 1-km data products for the population distribution across China in 2015 provide a credible dataset that can be employed in studies aimed at accurate population-dependent decisions.

ARTICLE HISTORY
Received 21 May 2020
Accepted 19 November 2020

KEYWORDS
Population distribution; random forest; area-to-point kriging; China

1. Introduction

"Where people are" is a fundamental question in regional science (Hägerstrand 1970), the answer to which influences decision-making, including resource allocation, epidemic control, and disaster response.

Census data commonly serve as the primary source of population distribution information. However, they only provide information on the total population counts within predefined enumeration units. Recently, there has been a high demand for downscaling census data to obtain fine-grained information on population distribution, ranging from grid level (Stevens *et al*. 2015, Patel *et al*. 2017, Pezzulo *et al*. 2017, Wang *et al*. 2018, Chen *et al*. 2019, Leyk *et al*. 2019) to building level (Lwin and Murayama 2009, Ural *et al*. 2011, Greger 2015, Yao *et al*. 2017), and from reconstruction (Fang and Jawitz 2018) to prediction (Chen *et al*. 2018). It has also been observed that earlier studies relied heavily on environmental data that are closely related to human activities (Stevens *et al*. 2015, Wardrop *et al*.

CONTACT Jianghao Wang wangjh@lreis.ac.cn
© 2020 Informa UK Limited, trading as Taylor & Francis Group

2018), including terrain, land use, and night light remote sensing imagery (Lloyd *et al*. 2017). However, these static data lack sufficient information on short-term (e.g., seasonal or monthly) fluctuations in population distribution. With the availability of emerging data from mobile phones (Deville *et al*. 2014, Ma 2018, Lai *et al*. 2019) and social media platforms (Longley and Adnan 2016, Dong *et al*. 2019), an increasing number of studies have applied such dynamic big data to overcome the limitations associated with static models. Such dynamic data reveal people's movement at the individual level, improve population estimates (Patel *et al*. 2017), and also enable dynamic population mapping.

Population distribution modeling at different temporal scales can reflect diverse human activity patterns. Many scholars have developed decadal (Wang *et al*. 2018) and annual (e.g., WorldPop: https://www.worldpop.org/; Landscan: https://landscan.ornl.gov/) population products, which support studies on urban expansion and climate change. Others have used daily or hourly population distribution data (Douglass *et al*. 2015, Chen *et al*. 2018) to predict inner-city human mobility, which is essential for urban planning and disaster management. However, these products hardly apply to studies related to inner-annual population variation on a national extent because their temporal resolution is either too low or too fine, and to the best of our knowledge, no data products for monthly or seasonal population distribution in China is currently available.

To fill this gap, in this study, we propose a hybrid model that combines the random forest algorithm and area-to-point kriging (Kyriakidis 2004, Liu *et al*. 2008, Chen *et al*. 2019). Thereafter, we evaluate the potential of this hybrid model to map the monthly population across China at a spatial resolution of 1-km by incorporating time-series mobile phone positioning data. Three considerations motivate this research. First, China frequently experiences mass migration movement, with a volume that reached 240 million in 2015; however, few evidence-based decision-making tools exist for public policymakers in this regard. Second, compared with conventional linear regression methods (Bagan and Yamagata 2015, Greger 2015, Lwin *et al*. 2016, Yao *et al*. 2017), machine learning algorithms, such as random forest, are more expert in accounting for the complex non-linear relationships between variables (Stevens *et al*. 2015, Hengl *et al*. 2017, 2018, Bi *et al*. 2019). But the ignorance of the samples' locations is their inherent weakness. Accordingly, we further apply kriging methods to characterize the remaining spatial dependence information in the estimation residuals. Finally, the mobile-phone penetration rate in China, which is 98 and 68% for mobile phones and smartphones, respectively, is higher than the world average (90 and 59%, respectively) (Poushter *et al*. 2018). Positioning data based on such a high mobile-phone penetration rate can provide information on people's whereabouts and describe how they move, which can improve inner-annual population variation estimations. The population distribution products derived in this study can benefit other human-related studies.

2. Data

2.1. Data collection

We used the year 2015 as the baseline year for collecting all the data that were used in this study (Table 1).

Table 1. Input dataset used for the modeling of gridded population distribution.

Name	Data format	Spatial resolution	Temporal resolution	Data source
Response variable				
Census data/1% population sample survery data	Polygon	Subdistrict/town/ township	5 years	National Bureau of Statistics of China
Covariates				
Elevation	Grid, Continuous	1-km	Static	ASTER DEM
Average slope	Grid, Continuous	1-km	Static	ASTER DEM
Land-use type	Grid, Categorical	100-m	5 years since 1980s	Landsat TM/ETM+, OIL
Proportion of cropland	Grid, Continuous	1-km	5 years since 1980s	Landsat TM/ETM+, OIL
Proportion of forest				
Proportion of grassland				
Proportion of waters				
Proportion of urban land				
Proportion of rural land				
Proportion of other buildings land				
Proportion of unutilized land				
Night light remote sensing imagery	Grid, Continuous	15″ (~500 m)	1 month since 2012	VIIRS-DNB
Mobile phone positioning data	Grid, Continuous	/	10 minutes	Tencent location-based service

2.1.1. Census data

Fine-grained and accurate census data play crucial roles in population spatialization. Given that the most recent census data (2010) do not effectively describe the current population characteristics of fast-paced urbanization areas, the 2015 1% population sample survey data were adopted for this study. Official population data reports corresponding to three fine-grained administrative units in China, including subdistrict (inner-city zoning), town, and township (rural zoning), were used. For the year 2015, there were a total of 43,327 town-level units in China. We obtained resident population data for 32,513 towns and townships from the 'China County Statistical Yearbook, 2016' (https://data.cnki.net/Yearbook/Single/N2017050134, data for 2015), and calculated resident population for 8,577 subdistricts by multiplying the population of these districts based on the 2010 census data by the population growth rate of the upper-level district or county. To fill up missing data corresponding to 2,237 blocks, a widely used open-source population mapping product, WorldPop, was used.

2.1.2. Environmental ancillary data

Given that the natural environment shapes population distribution, features including elevation, slope, land use, and night light remote sensing imagery were coupled in the estimation model as environmental covariates.

2.1.2.1. Elevation and slope. We produced a digital elevation model (DEM) by mosaicking ASTER DEM downloaded from the United States Geological Survey (USGS, https://earthexplorer.usgs.gov/). Thereafter, we resampled it from the original resolution of 1 arc-second (WGS-84 geographic reference system, approximately 30 m at the equator) to a resolution of 1-km on the Albers Equal Area projection and calculated the corresponding average surface slope based on it.

2.1.2.2. Land Use. We collected land use data from the Landuse Status Remote Sensing Monitoring Database of China (http://www.resdc.cn/data.aspx?DATAID=184). This database contains seven major categories (including, cropland, forest, grassland, waters, urban/rural/industrial, unutilized, and ocean) and 25 secondary categories. We aggregated it from the original 100-m resolution to a resolution of 1-km to generate the proportion of each land-use type. This was mainly done based on the major categories. However, considering that population density varies based on building type, location, and density, each built-up area was subdivided into three sub-categories, namely, urban land, rural land, and other buildings land (based on secondary categories under the major category, urban/rural/industrial).

2.1.2.3. Night light remote sensing. We obtained VIIRS-DNB nighttime light data from the National Oceanic and Atmospheric Administration (NOAA, https://ngdc.noaa.gov/eog/viirs/). To best approach reality, we chose the version with cloud-free averaged, outliers (e.g., lightning, aurora, and fires) removed, and background retained. The raw data are stored in 15-arc-second grids (approximately 500 m at the equator) in the WGS-84 geographic reference system.

2.1.3. Mobile phone positioning data

Tencent location-based service (https://heat.qq.com/) provides data on the number of mobile phone positioning signals in grids with a spatial resolution of 0.01 decimal degree (approximately 900 m at the equator) every 10 minutes. As the largest social media platform in China, Tencent offers location-needed applications such as online ride-hailing, near-by amenities searching, and real-time location sharing, by periodically sending mobile phone positioning data to the Tencent data center. We collected data from 1 January to 31 December 2015. Subsequently, we summarized the raw data by month to obtain monthly averaged data.

2.2. Data integration

A completely unified format of the covariates ensures the smooth execution of the estimation model. Thus, we resampled all the other raster data, including the proportion of each land-use type, night light, and rasterized mobile phone positioning, based on the DEM data, ensuring that they all had the same projection (Albers Equal Area), resolution (1-km), extent, and number of rows and columns. To resample land use and night light data, which have continuously changing characteristics, we used the bilinear method; and to resample rasterized mobile phone positioning data with skewness distribution characteristics, we used the nearest neighbor method.

3. Methodology

This section presents the hybrid downscaling model based on the random forest algorithm and area-to-point residual kriging method (Figure 1). We used the population density data corresponding to each town as the input response variable, and the input explanatory variables included a series of environmental variables as well as rasterized monthly averaged mobile phone positioning data.

First, we adjusted town-level census data to monthly census data to fix its insufficient representation of the dynamic population distribution. The mobile phone positioning data were leveraged as an indicator in this step because they reveal population distribution directly. Second, to characterize the non-linear relationship between the population distribution and the covariates, we constructed a random forest model for each month as a trend analysis. To construct this model, we used the average population density of each town and the corresponding mean values of the covariates. Thereafter, we used the model to estimate the population on each pixel assuming that the parameters are constant. Third, areal residuals of each town were obtained by subtracting aggregated random forest results from the adjusted census data, and area-to-point kriging was, then, subsequently used to realize a reasonable redistribution of the random forest estimation residuals from towns to covering pixels at the same spatial resolution (1-km). Thus, the random forest model was used to analyze the numerical correlation information from the variables, and the area-to-point kriging was employed to extract the remaining spatial autocorrelation information from the residuals. We considered that these two resultant layers combined contain all the information that can be extracted from our data. The final model could be expressed as:

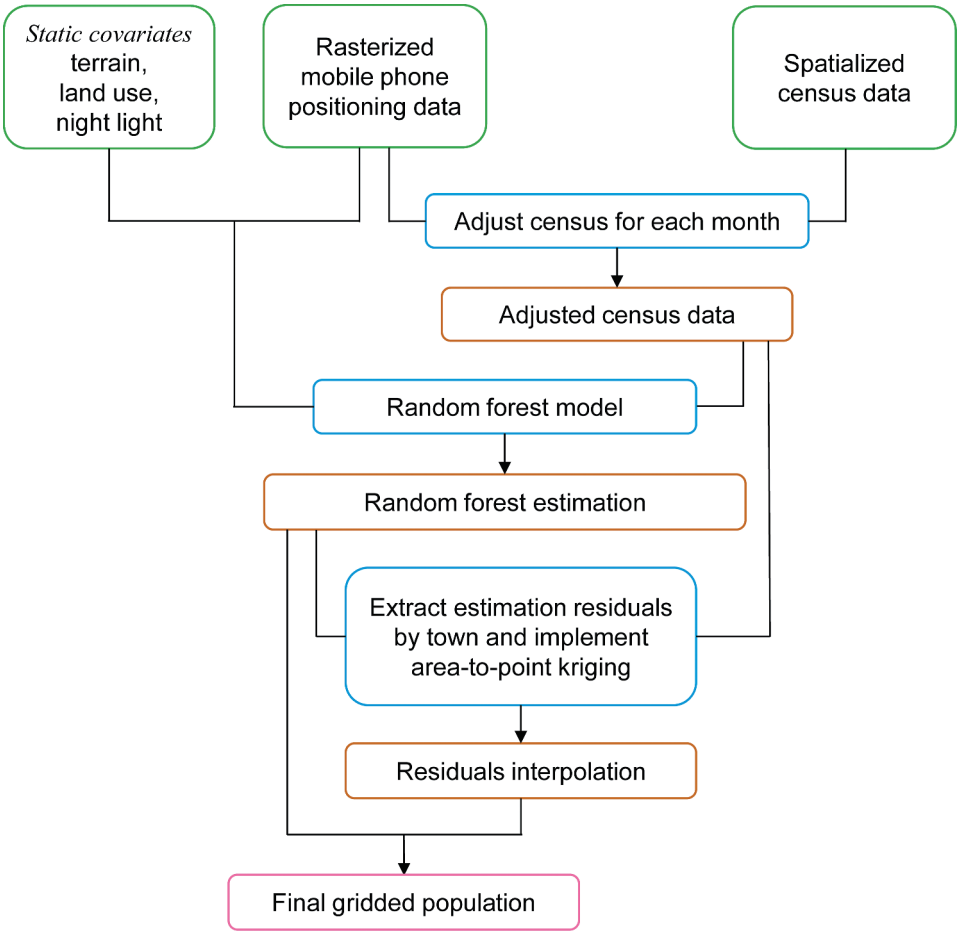


Figure 1. Dynamic population mapping workflow.

$$Pop_{i,t} = f_{RF}(X_{i,t}) + kriging_{i,t} + \varepsilon \tag{1}$$

where i represents the serial number of each pixel, t represents the month, Pop represents the population estimation, X represents the covariate vector, $f_{RF}(X_i)$ represents the random forest regression component at i, $kriging_i$ represents the residual component obtained from area-to-point kriging at i, and ε represents white noise with no information.

3.1. Adjusting census data

The monthly population change rate for each town was calculated as the ratio of the percentage of mobile phone positionings for that month to the percentage of the 12-month mean as shown below:

$$Pop_{j,t} = \frac{\frac{MP_{j,t}}{\sum MP_t}}{\frac{MP_{j,yearmean}}{\sum MP_{yearmean}}} \times Pop_{j,resident} \tag{2}$$

where *j* represents the town serial number, *t* represents the month, *yearmean* represents the mean value for 12 months, *resident* implies the resident population data obtained from the census data, *Pop* represents the population, and *MP* represents the number of mobile phone positionings.

3.2. The random forest model

Random forest (Breiman 2001) is an ensemble data-driven statistical method, which has been primarily adopted for classification, regression, and prediction. It grows a forest of individual trees, reserves the best randomly selected variable combination for each node of each tree, and executes the final prediction by averaging the predictions of individual trees.

In this study, random forest models were built to preliminarily downscale the census data. Given that determining the real population under a specific grid covering within a certain period is still challenging, we used adjusted town-level census data as input instead. The mean values of the covariates corresponding to each town were used as the explanatory variables, and the average population density of each town was used as the response variable; this average population density was converted from a skewed distribution to a normal distribution by performing log-transformation to ensure that it obeys Gaussian distribution like the environmental covariates. Thus, it was possible to estimate population distribution at pixel-level using the prediction model established at a coarser spatial resolution, which is a process that is only valid in a linear multivariate Gaussian regression context (Ribeiro Sales *et al*. 2013).

The modeling process was performed using the 'randomForest v6.4–14' package in an R 3.6.1 environment. To ensure the optimal performance of the final model, we iteratively adjusted the covariate selection as well as two fine-tuning parameters, namely, *ntree* and *mtry*. Specifically, *ntree* controls the number of trees in the forest, which was fixed at 500 given that the model had converged sufficiently by this time. Further, *mtry* represents the number of covariates that were randomly selected for each node; it was tuned using the *tuneRF* function by minimizing the out-of-bag (OOB) prediction error. We first tuned its initial value by including all the covariates into the function. Thereafter, it was used to construct an initial random forest model, and the sequence of the measure of covariate importance was extracted. Covariates with importance below zero (implying that no decline in OOB error was observed after randomly permuting their values) were eliminated. Next, we tuned a new *mtry* and constructed a new random forest model using the remaining covariates. This procedure was performed iteratively until no unessential covariates remained. Finally, we used the established optimal model to estimate the log-population on each pixel, and, then performed back-transformation.

3.3. Area-to-point kriging

The conventional kriging interpolation method was later adopted to supplement the limitations of the established random forest model in dealing with geographic autocorrelation between samples. Owing to the lack of true values corresponding to each pixel, town-level areal residuals were used instead. Thus, area-to-point kriging (Kyriakidis 2004), a variant of ordinary kriging, was used to realize a reasonable redistribution of the areal

residual to the corresponding pixels. This method ensures that the summary of the redistributed residuals of the pixels from the same town covering is equal to the town's residual. For each specific pixel, the residual was calculated as a weighted linear combination of the residuals of neighboring towns with an unbiased weight constraint.

$$\begin{cases} \widehat{e_p} = \sum_{k=1}^{K} \lambda(v_k) e(v_k) \\ \sum_{k=1}^{K} \lambda(v_k) = 1 \end{cases} \tag{3}$$

where e_p represents the estimated residual of a pixel, K represents the number of neighboring towns under consideration, $e(v_k)$ represents the known town-level areal residual, and $\lambda(v_k)$ represents the weight assigned to each town, which was solved as:

$$\begin{cases} \sum_{l=1}^{K} \lambda(v_l) \bar{C}(v_k, v_l) - \chi_p = \bar{C}(p, v_k), k = 1, \ldots, K \\ \sum_{l=1}^{K} \lambda(v_l) = 1 \end{cases} \tag{4}$$

where $\bar{C}(v_k, v_l)$ denotes the covariance between any two towns $v_k$andv_l under consideration, $\bar{C}(p, v_k)$ represents the area-to-point covariance between the target pixel and the town, v_k. It is worth noting that the covariances were exported from the coarse spatial resolution to the fine resolution via deconvolution, which is a technique that generates variograms from the discrete points of input areal data by minimizing the differences between the theoretically regularized variogram and the variograms corresponding to the input areal data (Goovaerts 2008, Chen *et al*. 2019).

This procedure was performed using the 'gstat v2.0–0' package in an R environment. The specific steps were as follows:

- Random forest results were aggregated to the town-level.
- The aggregation results were subtracted from the adjusted census data to obtain the areal residual of each town.
- A variogram at point support was fitted from areal residuals via deconvolution.
- An empty grid (the same as the target layer) was created as the kriging output template, and the areal residuals as well as the variogram model were used to execute area-to-point kriging.

4. Results

4.1. Evaluating the quality of the hybrid model

4.1.1. Random forest model evaluation

The *randomForest* function in the 'randomForest' package in the R environment generates two specific evaluation indexes automatically during modeling, namely, the *% var explained* and the *mean of squared residuals*, which represent the overall interpretation of the model and the mean value of the squared residuals, respectively. These furnish measurable indicators that reflect the quality of the model. It also provides a sequence of

the importance measures of the explanatory variables (*%IncMSE*), which can facilitate the filtering out of inessential covariates and clarify the inequitable contributions of the eventually selected covariates.

We summarize the parameters of 12 random forest models and observe that the *%var explained* of all 12 models are above 94%, demonstrating that random forest models could sufficiently explain the numeric correlation between variables. The 12-month sequence of the importance measures of the covariates (Figure 2(a)) demonstrates the ascendancy of mobile phone positioning data in the random forest models, with a mean *%IncMSE* value of 3.39, which is approximately 5-fold larger than that of the second variable. Night light and the proportion of urban land use also exhibit comparatively higher importance with mean *%IncMSE* values of 0.62 and 0.58, respectively, while those of the remaining covariates are less significant (i.e., < 0.5).

4.1.2. Semivariograms of random forest estimation residuals

The existence of spatial autocorrelation is the premise of kriging interpolation. Here, we map variograms for each month with an appropriate model to detect the spatial dependence of random forest estimation residuals. The criteria of the log-transformed squared average residual of each town yield Matern models with a range of 100-km with the best fit (Figure 2(b)). The nugget parameters of the 12 models are found to be similar; however, the sill parameters vary across months, resulting in nugget-to-sill ratios ranging from 33.24 to 40.04%.

4.2. Validating monthly population data via census and mainstream products

We evaluate the consistency of monthly population data by averaging 12-month population distribution and comparing them with town-level census data owing to the absence of real pixel-level population data (Figure 3(a)). We also compare our results with mainstream population products. Over the past decade, several scholars have developed and shared

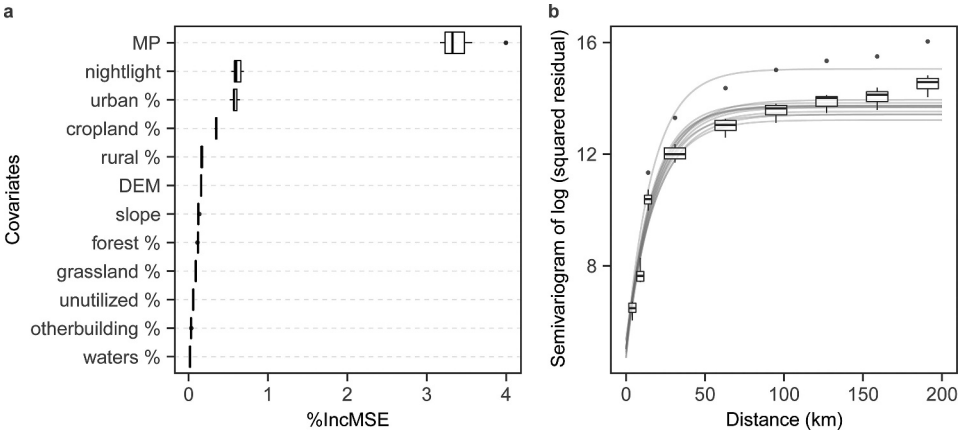


Figure 2. Evaluation of model quality. (a) Boxplot of the 12-month %IncMSE value of each covariate. (b) Semivariograms of the log-transformed squared average residual of each town. The grey lines indicate variogram models fitted for 12 months, and superimposed, are the boxplots corresponding to semivariances at specific distances for 12 months.

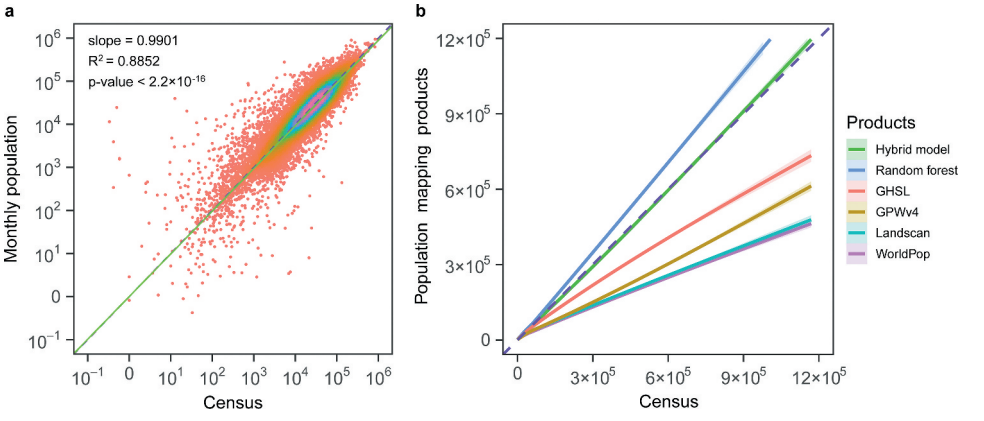


Figure 3. Evaluation of estimation consistency. (a) Town-level comparison of census data and the mean values of estimated 12-month population. (b) Town-level comparison of census data and the mean values of estimated 12-month population, the mean values of 12-month random forest-based results, and mainstream population products.

various global gridded population products, including WorldPop, Landscan, GHSL (Global Human Settlement Layer), and GPWv4 (Gridded Population of the World). We collect the data corresponding to China in 2015 from these population products and compare them with our census data at multiple spatial extents, using the results of random forests and hybrid models as references. The comparison indicates a higher consistency of our hybrid model-based products (Figure 3(b)). Further, the hybrid model also neatly addresses the problem associated with the random forest model results that their values are generally larger than real census data. Along with the aggregation of the evaluation extent, all the mainstream products exhibit sharp increases in R-square values (from < 0.8 to > 0.95) and a slightly sharper slope of the fitted lines. As shown in Table 2, the results obtained using the hybrid model are still the best among those corresponding to these products.

4.3. Exploring inner-annual population variation patterns using monthly population data

4.3.1. Yearly averaged population distribution

To reflect the general population distribution in China, we averaged 12-month population products. As shown in Figure 4, China's population is found to be concentrated in the eastern coastal and central plains, while the western and northern areas are found to be relatively sparsely populated. This distribution exactly coincides with the 'HuHuanyong line,' a famous empirical line that divides China into the populous region and less populous region.

In the populous region, central China (30°N–40°N), which predominantly consists of plains and basins with flat terrain, shows an evener population density, while southern China (< 30°N), most of which is densely covered by hills, shows an uneven population distribution.

Table 2. Consistency assessment of hybrid model results, random forest-based results, and mainstream products at multiple levels.

Comparison level	Hybrid model	Random forest	WorldPop	Landscan	GHSL	GPWv4
Town-level	slope = 0.99*** R^2 = 0.89***	slope = 1.18*** R^2 = 0.77***	slope = 0.51*** R^2 = 0.77***	slope = 0.56*** R^2 = 0.74***	slope = 0.82*** R^2 = 0.77***	slope = 0.58*** R^2 = 0.75***
County-level	slope = 1.02*** R^2 = 0.98***	slope = 1.23*** R^2 = 0.88***	slope = 0.58*** R^2 = 0.97***	slope = 0.62*** R^2 = 0.96***	slope = 0.86*** R^2 = 0.96***	slope = 0.63*** R^2 = 0.96***
City-level	slope = 1.02*** R^2 = 0.99***	slope = 1.18*** R^2 = 0.96***	slope = 0.59*** R^2 = 0.99***	slope = 0.63*** R^2 = 0.98***	slope = 0.88*** R^2 = 0.99***	slope = 0.64*** R^2 = 0.99***

Note: with the p-values of all t-tests of coefficients $< 2 \times 10^{-16}$ and the p-values of all adjusted $R^2 < 2.2 \times 10^{-16}$

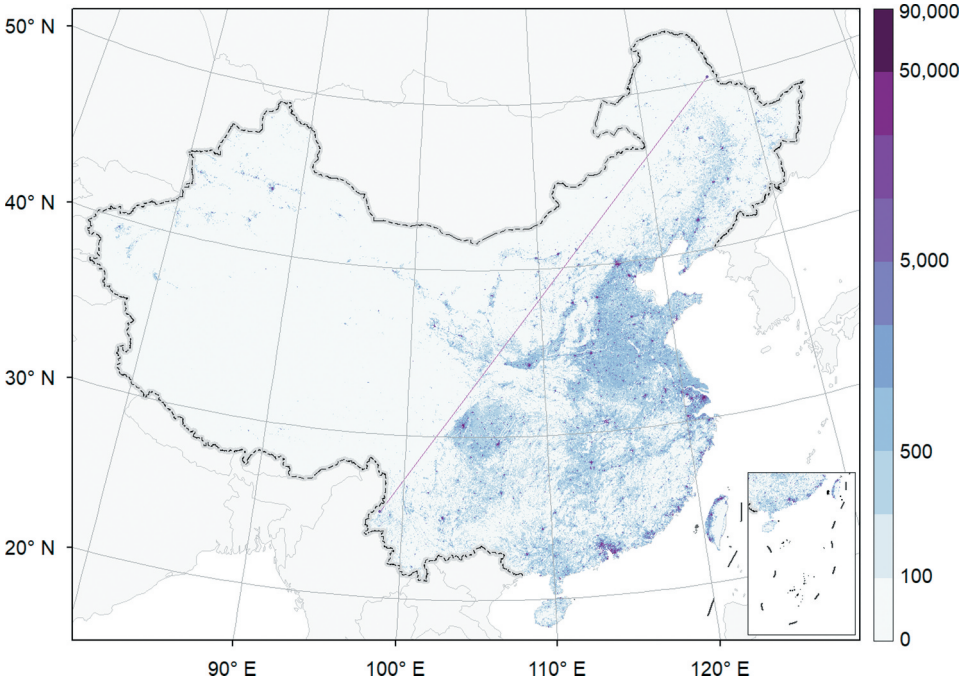


Figure 4. Yearly averaged population distribution in China calculated by averaging population products corresponding to 12 months. The northeast-southwest violet line represents the 'HuHuanyong Line,' which is a meaningful empirical line that divides the population distribution in China; proposed by Huanyong Hu in 1935. (* IJGIS remains strictly neutral with respect to jurisdictional claims on disputed territories in published maps in the figure.).

4.3.2. Population variation on primary land use in cities of different levels

Each city has its own characteristics based on its historical development and location, which affect population mobility to varying degrees. For example, people prefer to work in a city with a more robust economy; thus, emigration and immigration are affected. Similarly, high-tech talents are more likely to flock to big cities to enjoy the density of tech companies. Additionally, such incentives for migration and relocation further reinforce the characteristics of cities. Thus, we expect that monthly population data can be used to support the exploration of inner-annual population variation across different city classes and population types.

The mainstream media in China categorizes Chinese cities into six levels based on their economic development levels, lifestyle diversities, and population vitalities as follows: first-tier cities or world-class cities (i.e., Beijing, Shanghai, Guangzhou, Shenzhen); new first-tier cities (15 in total, upgraded from the second-tier city level); second-tier cities (30 in total); third-tier cities (70 in total); fourth-tier cities (90 in total); and fifth-tier cities (the remaining 128 cities). This classification scheme closely reveals the attractiveness of different cities to migrant workers and high-tech talents. We further recategorized these six city levels under three classes: global cities (first-tier), major cities (new first- and second-tier cities), and small cities (third- to fifth-tier cities).

To determine the number of people living in urban and rural areas, we multiplied the monthly population and the proportion of urban land use, rural and cropland land use, respectively. The mean value of the estimated 12-month population was selected as a standard, based on which variations were computed. We later joined them according to the three city classes mentioned above as shown in Figure 5.

Three periods require special attention. First, the 'Chunyun' period during which the traffic corresponding to the Spring Festival, which is reported to be the largest human migration on Earth, is observed. Each year, during this period, over 200 million people return to their hometowns from the cities where they live and work, and one to two weeks later, they return to the cities. Thus, there is a sharp decline in the urban population. Another notable pattern is that the urban population of the global and major cities rebound to their original sizes rapidly after the 'Chunyun' period ends, whereas the urban population in small cities continues to be low until April, and only recovers gradually.

The second period during which dramatic population movement is observed is the summer vacation (from July through August). Compared with the Spring Festival, fewer people return to their hometowns during this period. Rather, more people travel to global cities for traveling, sightseeing, or internship, resulting in a reverse variation trend in the urban population of global cities and small cities.

The third noteworthy period is the short-term labor flow in November. Basically, in November, there is no holiday or special day in the working calendar in China. However, this month signifies an important phase in agricultural production, the end of the autumn harvest. After this period, many farmers travel to nearby cities to work as short-term workers to subsidize their households, leading to a sudden increase in the urban population and a consequent decline in the population in rural areas. Thus, the urban population in all the cities at different levels reaches maxima in November, with more than 10% above the mean population, and in small cities, the urban population could even increase by 16.36%. Simultaneously, the sharpest drops in the year are also observed in the rural population of small cities.

4.3.3. Population variation distributions during specific periods

Given the importance of the three periods mentioned above, analyzing population variation patterns over the whole country promises to be significant. We considered the population distribution in February as the population during the Spring Festival, the mean value for July and August as that corresponding to the summer vacation, and the

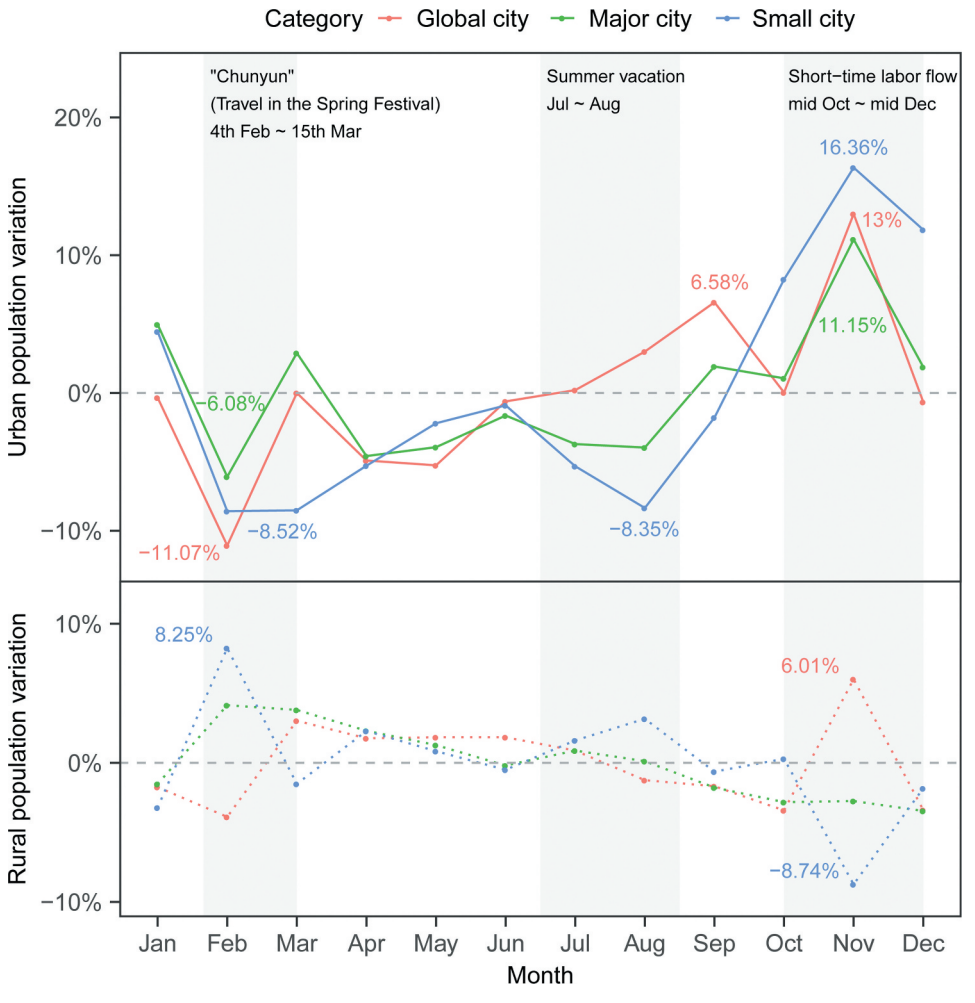


Figure 5. Monthly population variation in different land-use types and city categories. The upper and lower panels correspond to urban and rural population variation over 12 months. Both of them were equally reclassified into three categories, i.e., the population in global cities, major cities, and small cities. Three of the more volatile periods with grey background correspond to the 'Chunyun' (Spring Festival holiday), summer vacation, and short-term labor flow period, respectively.

population in November as that corresponding to the short-term labor flow. Thereafter, we calculated population changes as deviations from these and the 12-month mean.

During the Spring Festival (Figure 6(a)), there is a noticeable drop in the urban population. Remarkable regional declines appear in the Beijing-Tianjin-Hebei city cluster (A), the Yangtze River Delta city cluster (B), and the Pearl River Delta city cluster (C), all of which exactly correspond to the locations of the global cities. Conversely, there are widespread increases in the population in rural areas, indicating migration from urban to rural. A similar pattern is observed during the summer vacation (Figure 6(b)). However, Shanghai, which is one of the largest financial and tourist cities in the world, exhibits significant dissimilarities. It shows the largest population decline in the country during the Spring Festival, owing to the outflow of migrant workers, and in summer, it shows the largest population increase, mainly due to the

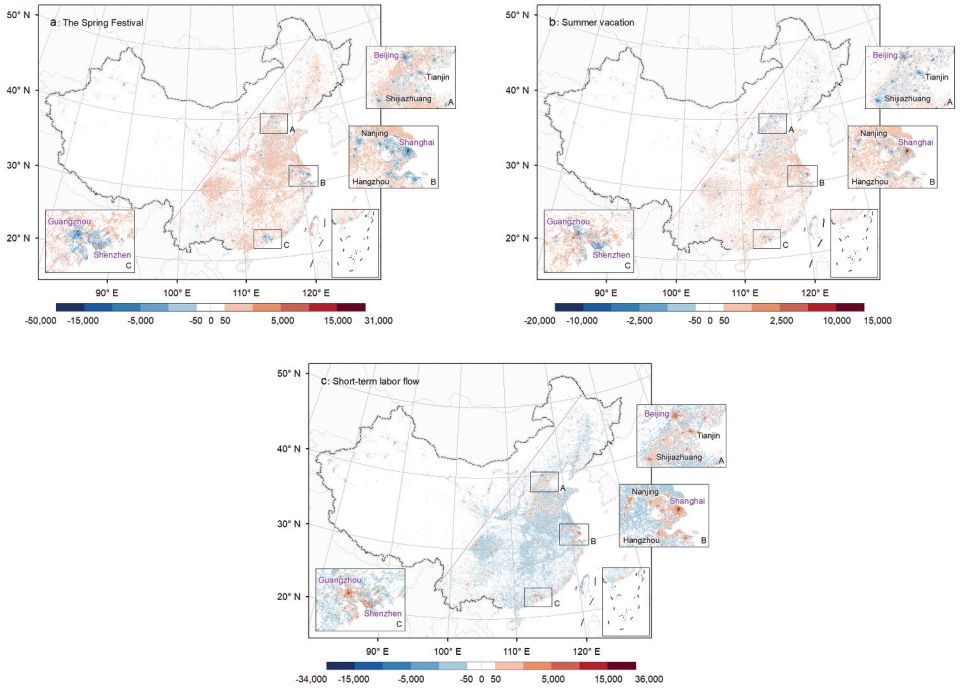


Figure 6. Population variation distributions within specific periods. (a) The Spring Festival, (b) The summer vacation, and (c) The short-term labor flow period. The three zoom-in maps correspond to (A) the Beijing-Tianjin-Hebei city cluster, (B) the Yangtze River Delta city cluster, and (C) the Pearl River Delta city cluster, which are the exact locations of the global cities in China, where significant variations are always observed. The global cities are labeled purple and other major cities are labeled black. (* IJGIS remains strictly neutral with respect to jurisdictional claims on disputed territories in published maps in the figure.).

influx of millions of tourists. Simultaneously, a similar trend is observed in Beijing, Nanjing, and Guangzhou. In November, the population variation exhibits patterns of population gain and loss in urban and rural areas the other way around compared with the Spring Festival period (Figure 6(c)).

4.3.4. Monthly population variation in the Yangtze River Delta city cluster

The Yangtze River Delta city cluster is one of the most populous, economically developed, and tech-gathered regions in China. It always demonstrates dramatic population changes; for example, Shanghai has always experienced population changes on the largest scale in the country (Figure 6). In this study, we map the monthly population changes in the core area of the Yangtze River Delta city cluster by determining the difference between the population of each month and the 12-month mean to develop a more intuitive visualization (Figure 7). For Shanghai, a distinct time division could be observed. The urban population is higher during the second half of the year than it is during the first half, with maximum and minimum values appearing in November and February, respectively, both of which deviate from the mean value by more than 20%, representing the general variation trend of the national urban population (Figure 5). In another view of the entire region, seasonal differences in rural

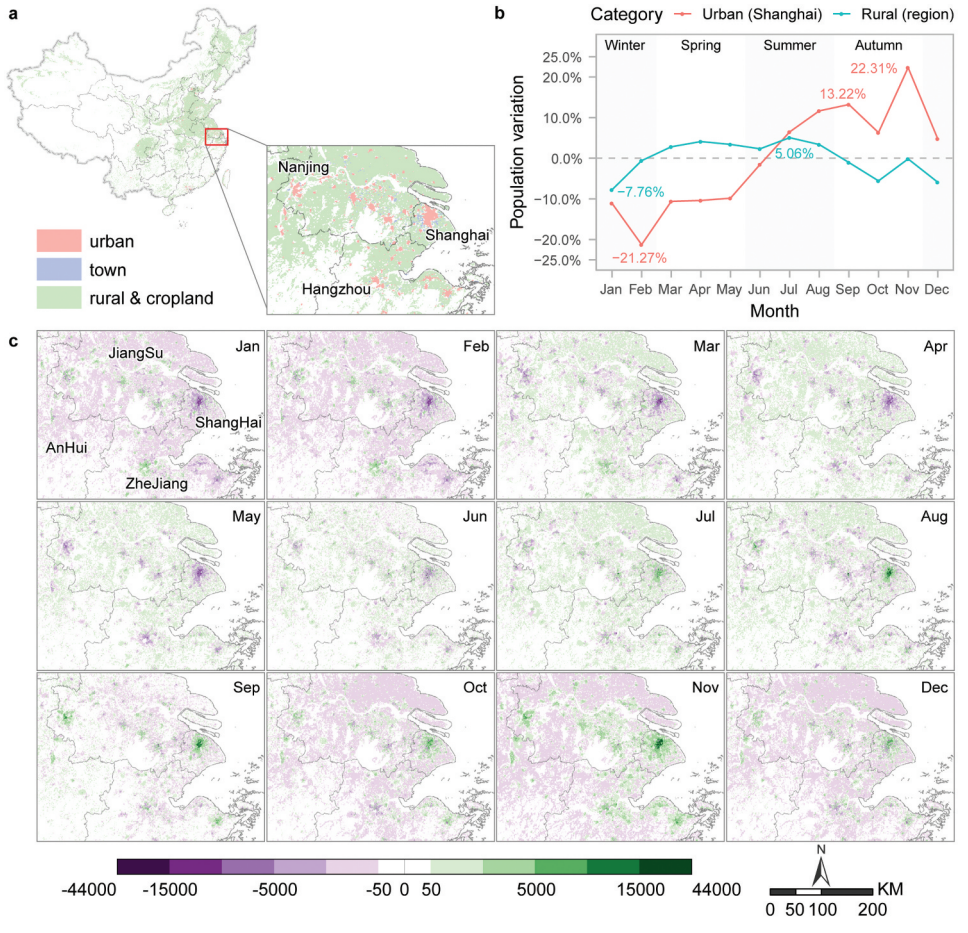


Figure 7. Monthly population variation in the core area of the Yangtze River Delta city cluster. (a) The distribution of three main land-use types in the core area of the Yangtze River Delta city cluster, (b) Variation of the urban population in Shanghai and the rural population in the whole region, (c) Population variation distributions in the core area of the Yangtze River Delta city cluster for 12 months.

population distribution are also clear. From spring to early autumn (the season for farming), the number of people residing in rural areas is higher than the yearly average, while in winter, it is lower.

5. Discussion

In this study, we propose a hybrid model, which is established by combining random forest and area-to-point kriging. This proposed model is, then, applied to downscale census data into population distribution at 1-km resolution across China for each month in 2015. The contributions and limitations of this work can be outlined as follows:

The monthly population distribution data provide a new perspective by which population dynamics can be observed. Based on the case analysis presented above, obvious

changes in the population distribution in China over months and seasons, across regions and cities, and between urban and rural areas, are documented. It is expected that these results will facilitate diverse governmental discussions, such as transportation dispatching during festivals and employment allocation optimization in response to short-term labor flow. Our results can also benefit a number of human-related studies, such as flood exposure estimation (Smith *et al*. 2019), determining whether people have easy access to open spaces and public transportation (goals 11.7 and 11.2 in the 2030 Sustainable Development Goals (SDG) of the United Nations), and capturing population flow responses to air pollution if such data for consecutive years are further available.

Our proposed hybrid downscaling model is highly expandable in three ways. First, in this study, we calibrate the population distribution for each month; however, we could still reproduce similar population distribution outputs at any finer spatial and temporal resolution as long as they are consistent with the mobile phone positioning data. Second, mobile phone positioning data could be replaced by emerging social media data, such as Geotweet data, which are easier to obtain from networks and are less constrained by user privacy considerations. Further studies can be conducted by subdividing these newer data based on detailed information, such as gender, age group, and education level, to produce more meaningful layers, and to add heterogeneity to the spatial population distribution. Third, our product and the underlying mobile phone positioning data reveal the general inner-annual population variation pattern in China, which coincides with the patterns observed in the past years, and probably in the coming years (unless the nation suffers a major disaster such as a coronavirus outbreak). We may recall or predict population variation for periods between censuses based on the mobile phone positioning data adjacent to time and the properly adjusted census data (for example, based on the natural growth rate or migration data) using our proposed model.

Additionally, in this study, we provide a feasible plan for applying machine learning algorithms in geographic downscaling research; however, the problem of scale is still unresolved. Due to the absence of the true population under certain pixels, random forest models were constructed based on average population densities of towns and the corresponding mean values of covariates. These town-level models were later used to estimate the population at pixel-level; analogously, variograms of geostatistical models were established based on the residuals of towns, and later used to execute pixel-level redistribution, which is not good enough even though they were deconvoluted. The latter problem seems to be an intractable question in Geostatistics as numerous researchers have already attempted to solve it but have failed to reach an agreed solution. The method of anchored distribution (MAD) (Rubin *et al*. 2010), which has been widely used in Hydrology (Chen *et al*. 2012, Heße *et al*. 2015), may provide new resolution. This method anchors non-local data in specific characteristic space using the Bayesian theory to realize multi-scale data transformation. Our future research will incorporate this into the existing model to seek a more accurate estimates.

6. Conclusions

Capturing population distribution within a year is essential for human-related studies. In this study, taking advantage of time-series mobile phone positioning data, a hybrid model was developed to realize monthly population distribution mapping in China. First, the regression between population density and covariates using a random forest model was performed.

Next, the area-to-point kriging method was adopted to supplement the assumption that regression residuals contain remaining spatial autocorrelation information. The fine-grained population distribution was ultimately estimated by combining the random forest and area-to-point kriging outputs. The results obtained demonstrate that this method has higher consistency compared with census data than mainstream population products. Additionally, the monthly population data obtained well capture inner-annual population variations in China and provide a new perspective for the study of population dynamics.

Acknowledgments

We are grateful to Yuehong Chen and Binzhe Wang for comments and feedback in the development of this paper.

Data and codes availability statement

The monthly population distribution product, coupling with source data and codes that support the findings of this study are available with the identifiers at: https://doi.org/10.6084/m9.figshare.12319334

Disclosure statement

The authors declare no competing interests.

Funding

This research was funded by the National Key Research and Development Program of China [2017YFB0503500], the National Natural Science Foundation of China [41971409, 41531174], and the Youth Innovation Promotion Association of the Chinese Academy of Sciences.

Notes on contributors

Zhifeng Cheng is a master student of the University of Chinese Academy of Sciences and the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (CAS). His research is now focused on mapping social-economic dynamics with big data.

Jianghao Wang is an associate professor in the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS. His research interests concentrate on spatiotemporal data mining.

Yong Ge is currently a Professor of the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS. Her research is focused on spatiotemporal statistics.

ORCID

Zhifeng Cheng http://orcid.org/0000-0001-9261-9537
Jianghao Wang http://orcid.org/0000-0001-5333-3827

References

Bagan, H. and Yamagata, Y., 2015. Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience and Remote Sensing*, 52 (6), 765–780.

Bi, J., *et al*., 2019. Impacts of snow and cloud covers on satellite-derived PM2.5 levels. *Remote Sensing of Environment*, 221, 665–674. doi:10.1016/j.rse.2018.12.002

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

Chen, J., *et al*. 2018. Fine-grained prediction of urban population using mobile phone location data. *International Journal of Geographical Information Science*, 32 (9), 1770–1786. doi:10.1080/13658816.2018.1460753

Chen, X., *et al*., 2012. Three-dimensional bayesian geostatistical aquifer characterization at the hanford 300 area using tracer test data. *Water Resources Research*, 48, W06501. doi:10.1029/2011WR010675

Chen, Y., *et al*., 2019. Downscaling census data for gridded population mapping with geographically weighted area-to-point regression kriging. *IEEE Access*, 7, 149132–149141. doi:10.1109/ACCESS.2019.2945000

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

Dong, L., Ratti, C., and Zheng, S., 2019. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences of the United States of America*, 116 (31), 15447–15452. doi:10.1073/pnas.1903064116

Douglass, R.W., *et al*. 2015. High resolution population estimates from telecommunications data. *EPJ Data Science*, 4 (1), 4. doi:10.1140/epjds/s13688-015-0040-6

Fang, Y. and Jawitz, J.W., 2018. High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Scientific Data*, 5, 180067. doi:10.1038/sdata.2018.67

Goovaerts, P., 2008. Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geology*, 40 (1), 101–128.

Greger, K., 2015. Spatio-temporal building population estimation for highly urbanized areas using GIS. *Transactions in GIS*, 19 (1), 129–150. doi:10.1111/tgis.12086

Hägerstrand, T., 1970. what about people in regional science? *Papers in Regional Science*, 24 (1), 7–24. doi:10.1111/j.1435-5597.1970.tb01464.x

Hengl, T., *et al*. 2017. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE*, 12 (2), e0169748. doi:10.1371/journal.pone.0169748

Hengl, T., *et al*., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518. doi:10.7717/peerj.5518

Heße, F., *et al*., 2015. Characterizing the impact of roughness and connectivity features of aquifer conductivity using Bayesian inversion. *Journal of Hydrology*, 531, 73–87. doi:10.1016/j.jhydrol.2015.09.067

Kyriakidis, P.C., 2004. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36 (3), 259–289. doi:10.1111/j.1538-4632.2004.tb01135.x

Lai, S., *et al*. 2019. Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications*, 5, 34.

Leyk, S., *et al*., 2019. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, 11, 1385–1409. doi:10.5194/essd-11-1385-2019

Liu, X.H., Kyriakidis, P.C., and Goodchild, M.F., 2008. Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, 22 (4), 431–447. doi:10.1080/13658810701492225

Lloyd, C.T., Sorichetta, A., and Tatem, A.J., 2017. High resolution global gridded data for use in population studies. *Scientific Data*, 4, 170001. doi:10.1038/sdata.2017.1

Longley, P.A. and Adnan, M., 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30 (2), 369–389. doi:10.1080/13658816.2015.1089441

Lwin, K.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/j.1467-9671.2009.01171.x

Lwin, K.K., Sugiura, K., and Zettsu, K., 2016. Space–time multiple regression model for grid-based population estimation in urban areas. *International Journal of Geographical Information Science*, 30 (8), 1579–1593. doi:10.1080/13658816.2016.1143099

Ma, T., 2018. Multi-level relationships between satellite-derived nighttime lighting signals and social media-derived human population dynamics. *Remote Sensing*, 10 (7), 1128. doi:10.3390/rs10071128

Patel, N.N., *et al*. 2017. Improving large area population mapping using geotweet densities. *Transactions in GIS*, 21 (2), 317–331. doi:10.1111/tgis.12214

Pezzulo, C., *et al*., 2017. Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. *Scientific Data*, 4, 170089. doi:10.1038/sdata.2017.89

Poushter, J., Bishop, C., and Chwe, H., 2018. *Social media use continues to rise in developing countries but plateaus across developed ones*. Pew Research Center. https://www.pewresearch.org/global/2018/06/19/2-smartphone-ownership-on-the-rise-in-emerging-economies/

Ribeiro Sales, M.H., Souza, C.M., and Kyriakidis, P.C., 2013. Fusion of MODIS images using kriging with external drift. *IEEE Transactions on Geoscience and Remote Sensing*, 51 (4), 2250–2259. doi:10.1109/TGRS.2012.2208467

Rubin, Y., *et al*., 2010. A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. *Water Resources Research*, 46, W10523. doi:10.1029/2009WR008799

Smith, A., *et al*. 2019. New estimates of flood exposure in developing countries using high-resolution population data. *Nature Communications*, 10 (1), 1814. doi:10.1038/s41467-019-09282-y

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. doi:10.1371/journal.pone.0107042

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

Wang, L., *et al*., 2018. Mapping population density in China between 1990 and 2010 using remote sensing. *Remote Sensing of Environment*, 210, 269–281. doi:10.1016/j.rse.2018.03.007

Wardrop, N.A., *et al*. 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences of the United States of America*, 115 (14), 3529–3537. doi:10.1073/pnas.1715305115

Yao, Y., *et al*., 2017. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, 31 (6), 1220–1244.