

ORIGINAL ARTICLE

Open Access



Dynamic population mapping with AutoGluon

Yimeng Song^{1*} , Yong Xu², Bin Chen³, Qingqing He⁴, Ying Tu⁵, Fei Wang⁶ and Jixuan Cai⁷

Abstract

Timely and accurate population mapping plays an essential role in a wide range of critical applications. Benefiting from the emergence of multi-source geospatial datasets and the development of spatial statistics and machine learning, multi-scale population mapping with high temporal resolutions has been made possible. However, the over-complex models and the strict data requirement resulting from the constant quest for increased accuracy pose challenges to the repeatability of many population spatialization frameworks. Therefore, in this study, using limited publicly available datasets and an automatic ensemble learning model (AutoGluon), we presented an efficient framework to simplify the model training and prediction process. The proposed framework was applied to estimate county-level population density in China and received a good result with an r^2 of 0.974 and an RMSD of 427.61, which is better than the performances of current mainstream population mapping frameworks in terms of estimation accuracy. Furthermore, the derived monthly population maps and the revealed spatial pattern of population dynamics in China are consistent with earlier studies, suggesting the robustness of the proposed framework in cross-time mapping. To our best knowledge, this study is the first work to apply AutoGluon in population mapping, and the framework's efficient and automated modeling capabilities will contribute to larger-scale and finer spatial-temporal population spatialization studies.

Keywords: Population spatialization, AutoML, Geospatial big data, Social sensing, GeoAI

1 Introduction

Timely and accurate population distribution data play an essential role in a wide range of critical applications, including but not limited to epidemiological studies (Li et al., 2020), natural disaster management (Chen, Wu, Jin et al., 2022), climate change analysis (O'Neill et al., 2020), ecological vulnerability assessment (He et al., 2018), emergency response (Tian et al., 2020), environmental monitoring (He et al., 2021), and sustainable urban planning (Long et al., 2022). In most scenarios, the population data used in these applications are derived from officially published census data. Although census data have a scientific collection process and high accuracy, the insufficient information

in the temporal dimension caused by low update frequency (5–10 years) makes census data hard to meet the needs of many real-world applications, leading to the increase of analysis bias, imbalance of resource allocation, and inefficiency of government management (Song, 2019; Xu et al., 2021).

To address this limitation, scientists seek observation data reflecting human activity to provide richer spatial and temporal information for population mapping. Therefore, remote sensing data are extensively used due to their ability to record continuous observations across time and space (Chen et al., 2020). The most popular remote sensing data include nighttime light (NTL) and optical satellite data (Wang et al., 2018). NTL data capture the footprint of light on the earth's surface generated by human activities, and its brightness level directly indicates the intensity of human activities, providing support for inferring population distribution (Song et al., 2020).

*Correspondence: yimeng.song@yale.edu

¹ School of the Environment, Yale University, New Haven, CT 06511, USA
Full list of author information is available at the end of the article

NTL from the US Air Force Defense Meteorological Satellite Program Operational Linescan System (DMSP/OLS), released in the early 1990s, is widely used for population mapping (Yu et al., 2018). However, its limitations of low spatial and radiometric resolutions, blooming effect, and over-saturation problem would result in underestimation in high-density areas (e.g., urban centers), making it only applicable to large-scale estimations (e.g., city or larger scales) (Elvidge et al., 1999). Another frequently used NTL data are from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor carried by the weather satellite Suomi NPP. Given the advanced luminous technology equipped in VIIRS, NPP/VIIRS NTL images come with a higher spatial resolution and are demonstrated to have better performance in population spatialization compared with its predecessor, DMSP/OLS (Wang et al., 2018).

Optical remote sensing images record reflection information of the earth's surface in multiple wavebands. Based on the principle that different types of land cover have different population carrying capacity, optical remote sensing images with different resolutions are also widely employed for population spatialization (Wang et al., 2018). Some works also indicate that optical remote sensing images and derived land cover/use products could perform better than NTL data in population mapping (Zeng et al., 2011), and a higher spatial resolution of optical data would also contribute to a higher accuracy (Linard et al., 2011). The development of geographic information and remote sensing technology has increased the accessibility of numerous datasets. Therefore, beyond using a single type of dataset, most of the well-known population count data products are obtained based on the fusion of multiple sources of data, for example, WorldPop from the University of Southampton, UK (Stevens et al., 2015; Tatem, 2017), LandScan from Oak Ridge National Laboratory, US (Bright et al., 2016), Gridded Population of the World (GPW) from Socioeconomic Data and Application Center, NASA, US (Doxsey-Whitfield et al., 2015), and Global Human Settlement Layer (GHSL) from Joint Research Centre, European (Freire et al., 2016). Despite the satisfactory performance of these population data in terms of spatial accuracy, information like short-term population distribution dynamics caused by intra- and inter-city human mobility is highly essential in various applications but still cannot be provided by these static datasets with an update frequency of 1 year or longer (Cheng et al., 2022).

Since 2010, the gradual popularization of smartphones along with the rapid development of mobile Internet triggered by 4G communication technology have completely changed the way of observing human behaviors. Specifically, smart devices users actively or passively upload

their spatial-temporal information during their access to various mobile Internet services and applications, and the collection of such a vast amount of information makes it possible to continuously observe people's real-time spatial behavior at a large spatial scale (Song et al., 2019), which is also named as "social sensing" (Liu et al., 2015). Social sensing big data have therefore attracted much attention of the scientific communities and are gradually adopted with satellite data in population mapping to improve accuracy and enrich information in the temporal dimension. Some widely used datasets include geotagged social media data (Patel et al., 2017), mobile phone data (Deville et al., 2014), point of interests (POI) data (Bakillah et al., 2014), and smart card data (Ma et al., 2017).

Besides fusing multiple data sources, estimation methods (regression models) for exploring the relationship between population and various features also play a critical role in population mapping. According to the algorithm, these models can be divided into three categories, including (1) statistical model, (2) spatial statistics model, and (3) machine learning. Statistical models are a powerful and widely used class, such as linear regression model (Bagan & Yamagata, 2015), and log-linear regression model (Liu et al., 2018). To address the complex relationship between variables arising from spatial heterogeneity, many works employed local-model-based spatial statistics models, ranging from geographically weighted regression (GWR) (Wang et al., 2018; Xu et al., 2021), geographically and temporally weighted regression (GTWR) (Liu et al., 2021), and Bayesian spatio-temporal model (Wang et al., 2021). In the past few years, the boom of Artificial Intelligence (AI) has made machine learning models increasingly employed in population spatialization, which plays a vital role in improving estimation accuracy. These machine learning models include not only single models like random forest regression models (Cheng et al., 2022), but also ensemble models like XGBoost (Tu et al., 2022).

While many well-designed population spatialization frameworks have been proposed and reported satisfactory accuracy, some new issues have emerged. (1) The fusion of an excessive number of data sources increases the difficulty of data acquisition, especially for studies that use non-public data, making these methods difficult to be replicated by others. (2) Although the spatial statistical models have ideal fit and prediction performance, their transferability is limited due to their nature as local models, which means that such models may not be suitable for estimating non-modeled areas or increase uncertainties in results. (3) For the use of machine learning models, in addition to the cost of parameter tuning required, selecting a suitable machine learning model is

also a challenge that must be faced. To alleviate the above problems, this study proposes a population estimation framework based on publicly available spatial datasets and an automatic ensemble learning framework. The framework will be used to estimate the monthly county-level population densities to reveal population distribution dynamic in China in 2015.

2 Materials and methods

Four main procedures are involved in mapping dynamic population density in this study: (1) data collection and feature extraction; (2) automatic ensemble learning model training and prediction; (3) accuracy assessment and comparison; (4) monthly population density mapping.

2.1 Data collection and feature extraction

We incorporated multiple categories of geospatial datasets (Table 1) for dynamic population density mapping, including demographics data, social sensing data, medium-resolution (MR) multispectral images, NTL data, land use data, and topography data. All the datasets

except topography data were obtained for 2015, and details of each dataset and extracted features for model training are provided as follows.

2.1.1 Demographics

County-level demographics data from the 1% population sample survey in 2015 and corresponding administrative boundaries were collected, spatially combined, and adopted for both model training and validation. County-level in China is equivalent to level-3 of the Global Administrative Unit Layer (GAUL) as defined by the Food and Agriculture Organization of the United Nations (Chen, Wu, Song et al., 2022). A total of 2851 counties in China were included in this study, without the records of Hong Kong, Macau, and Taiwan. We further calculated each county's population density (/km²) based on total population and administrative area.

2.1.2 Tencent LBS data

Thanks to Tencent's large number of active users, the LBS data from Tencent have good performance in describing the digital footprints of human activities (Chen, Song,

Table 1 List of datasets and extracted features used in population density mapping

Category	Dataset	Format, Resolution, and Year	Feature(s)	Variable(s)
Demographics	Census (1% population sample survey)	Polygons, county-level, 2015	Density of population	Pop_den
Social sensing	Tencent location-based service (LBS) data	Grid, 30 arc-second, 2015	Mean of monthly LBS data Sum of monthly LBS data	LBS_mean LBS_sum
MR multispectral	Landsat-8 Operational Land Imager (OLI)	Grid, 30-m, 2015	Mean of NDVI, NDWI, and NDBI Sum of NDVI, NDWI, and NDBI	NDVI_mean, NDWI_mean, NDBI_mean NDVI_sum, NDWI_sum, NDBI_sum
Nighttime light data	NPP-VIIRS nighttime light (NTL) data	Grid, 15-arc-second, 2015	Mean of nighttime light Sum of nighttime light	NTL_mean NTL_sum
Land-use	Land-use Status Remote Sensing Monitoring Database of China	Grid, 30-m, 2015	Coverage of urban area, rural area, water, forest, grassland, and cropland Sum of urban area, rural area, water, forest, grassland, and cropland	Urban_mean, Rural_mean, Water_mean, Forest_mean, Grassland_mean, Cropland_mean Urban_sum, Rural_sum, Water_sum, Forest_sum, Grassland_sum, Cropland_sum
Topography	SRTM V3 digital elevation data (DEM) and slope	Grid, 1 arc-second, 2000	Mean of DEM and slope Sum of DEM and slope	DEM_mean, slope_mean DEM_sum, slope_sum
Population data (for accuracy assessment and comparison)	WorldPop	Grid, 100-m, 2015	Density of WorldPop population	
	LandScan	Grid, 30 arc-second, 2015	Density of LandScan population	
	Gridded Population of the World (GPW)	Grid, 30 arc-second, 2015	Density of GPW population	
	Global Human Settlement Layer (GHSL)	Grid, 1-km, 2015	Density of GHSL population	

Kwan, et al., 2018), and have been successfully applied in many fields like population distribution mapping (Xu et al., 2021), land use classification (Gong et al., 2020), environmental exposure assessment (Chen, Song, Jiang, et al., 2018; Song et al., 2021), and human mobility mining (Zhu et al., 2018). We collected Tencent LBS data (<https://heat.qq.com>) generated in 2015 using the method introduced in Song et al. (2018). The raw data is tabulated as the number of location service requests with a spatial resolution of 30 arc-second and a 5-minute update frequency. We then transformed tabular data into hourly aggregated raster data and generated the average hourly raster for each month in 2015. The monthly mean and sum of LBS data were finally derived over each county as features.

2.1.3 Landsat-8 OLI imagery

Launched in February 2013 as part of a long-term Landsat program led by the U.S. Geological Survey and NASA, Landsat-8 is designed to collect medium-resolution multispectral image data and provide seasonal coverage of land surface (Roy et al., 2014). The sensor of Operational Land Imager (OLI) carried by Landsat-8 has nine reflective wavelength bands, six of which (i.e., blue, green, red, NIR, SWIR-1, SWIR-2) are designed for land applications with a spatial resolution of 30-m and a 16-day repeat cycle (Zhang et al., 2018). Here we collected Landsat-8 imagery from January 1 to December 31, 2015, for model training. A pixel-based quality check was first performed to eliminate observations contaminated with clouds and shadows from the entire Landsat-8 archive, using cloud masking and quality assessment information from Landsat-8 metadata. We then calculated the normalized difference vegetation index ($NDVI = (NIR - Red) / (NIR + Red)$) for each retained pixel. The whole-year maximum NDVI values were further selected and used as the quality index to generate the 2015 cloud-free greenest Landsat-8 composite. Based on this composite image, we calculated the normal difference built-up index ($NDBI = (SWIS - NIR) / (SWIR + NIR)$) and the normal difference water index ($NDWI = (NIR - SWIR) / (NIR + SWIR)$). The mean and sum of NDVI, NDBI, and NDWI were then derived for each county.

2.1.4 NPP-VIIRS nighttime light data

NPP-VIIRS NTL data with a spatial resolution of 15 arc-second are highly performing in characterizing various human activities, such as determining urban expansion patterns (Song et al., 2020), and monitoring human mobilities (Cai et al., 2017). In this study, we collected the annual cloud-free composites of 2015 without interference from stray light and lunar illumination. Each county's annual mean and sum radiance value was then derived as features.

2.1.5 Land-use data

We collected land-use data in 2015 from the Land-use Status Remote Sensing Monitoring Database of China provided by the Chinese Academy of Sciences Resource and Environmental Science Data Center (www.resdc.cn). The data with a spatial resolution of 30-m were produced based on Landsat TM/ETM remote sensing images, having six first-class and 25 second-class land-use types. The county-level coverage rate and total area of land-use types of "urban area", "rural area", "water", "forest", "cropland", and "grassland" were derived as features.

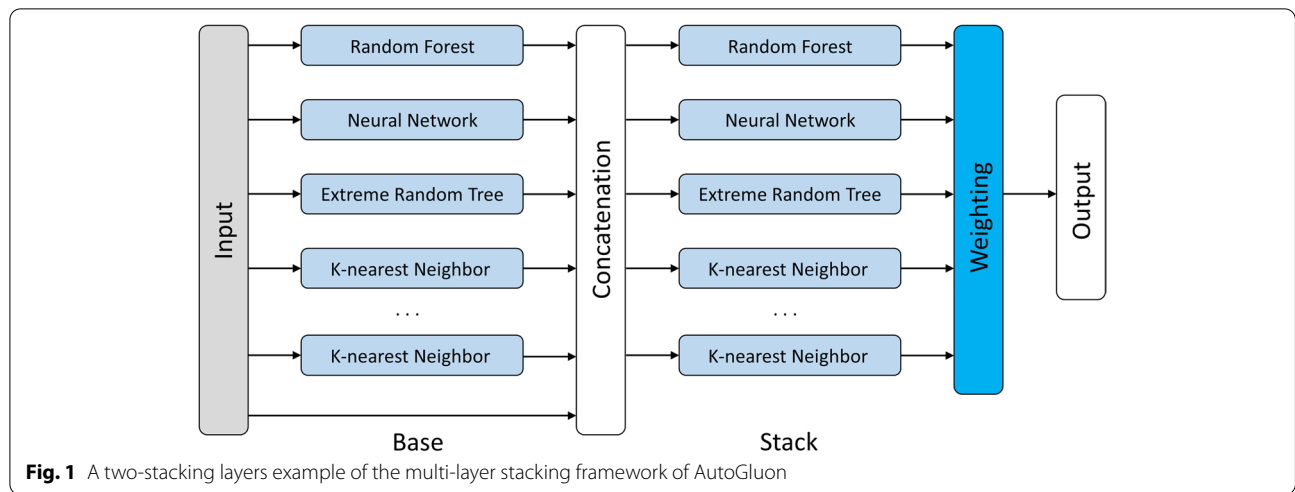
2.1.6 DEM and slope

The Shuttle Radar Topography Mission (SRTM) V3 digital elevation data (DEM) (Farr et al., 2007) were collected for China, which was provided by NASA in 2000 with a spatial resolution of 1 arc-second. Since the earth's surface elevation change is a slow process, the temporal inconsistency of this data with other data (collected in 2015) will have little effect on the estimation results. The mean and sum of DEM and slope were then derived for each county.

2.2 Mapping population density with automatic ensemble learning

We utilized the automatic ensemble learning framework embedded in AutoGluon (Version: 0.5.2) (<https://github.com/awsmlabs/autogluon>) to train the regression models for population density mapping. AutoGluon adopts a novel automated machine learning (AutoML) framework that employs a multi-layer stacking strategy (Erickson et al., 2020). As shown in Fig. 1, the multi-layer stacking framework of AutoGluon is constructed by one base layer and a minimum of one stacking layer. The base layer has several base models, and their prediction outputs are concatenated and fed into the stacker models in next layer, which then serve as base models for additional higher stacking layers. To avoid expensive costs of algorithm selection and hyperparameter optimization, AutoGluon simply reuses the same models as stackers in each layer with the same hyperparameter values (Erickson et al., 2020). This technique can be seen as another form of deep learning, using layer-wise training, where the units connected between layers could be arbitrary machine learning models. AutoGluon also enables stacker models in any higher layer to take as input both the predictions from the previous layer and original input features during training. The final stacking layer combines the predictions of the stacker models in a weighted manner by applying ensemble selection (Caruana et al., 2004).

AutoGluon has the capacity to automate the process of data pre-processing, base models search, hyperparameter tuning, and model ensembling during training. In this



study, we employed automatic multi-layer stacking and 5-fold cross-validation as the parameters of the Tabular Prediction of AutoGluon model training. The customized base models used in the Tabular Prediction (for regression) include neural network algorithms (e.g., NeuralNetFastAI, NeuralNetTorch), random forest algorithms (e.g., RandomForestMSE), extreme random tree algorithms (e.g., ExtraTreesMSE), K-nearest neighbor algorithm (e.g., KNeighborsDist, KNeighborsUnif), and boosting tree algorithms (e.g., LightGBM, CatBoost, XGBoost, LightGBMXT, LightGBMLarge). In addition, ensemble learning algorithms is used for combining predictions in the final stacking layer.

All the 26 variables listed in Table 1 were used for model training which was performed using CPU. Note that since the national 1% sample survey population data we used represent the population with a standard time of November 1, 2015, at 00:00 (National Bureau of Statistics of China, 2016), we only used the Tencent LBS data collected from November for model training, and the data from other months were only used for dynamic population density mapping presented in Section 3.4.

2.3 Accuracy assessment and comparison

We split the 2851 samples into 80% for training and 20% for validation. Root mean square error (RMSE, Eq.1) was used as the indicator for accuracy validation. To detect the contribution of different inclusive features, we calculated feature importance scores for the final weighted ensemble model via permutation importance. To be specific, the importance score of a feature represents the decrease in the model’s prediction accuracy (namely RMSD in this study) when the values of

the feature have been randomly shuffled across rows. The higher a feature score, the more important it is for the prediction accuracy of a model. A negative importance score means the features may be harmful to the model’s prediction, and a model without the feature having a negative score is likely to achieve better predictive performance (Erickson et al., 2020). We further included three more indicators for model comparison: relative root mean square error (%RMSE, Eq.2), mean absolute error (MAE, Eq.3), and coefficient of determination r^2 .

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - \hat{p}_i)^2}{n}} \tag{1}$$

$$\%RMSE = \frac{RMSE}{\frac{1}{n} \sum_{i=1}^n \hat{p}_i} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^n |p_i - \hat{p}_i|}{n} \tag{3}$$

where p_i is the estimated population density of the i^{th} county, \hat{p}_i is the observed population density of the i^{th} county, and n is the number of counties included in this study.

Four mainstream population distribution datasets (Table 1) were further collected and used to compare the estimation results, including WorldPop, Lands-can, GPW, and GHSL. Using census population density as the benchmark, we compared the accuracy of our estimated population density by RMSE with these four grided population datasets at the county level and calculated the county-average %RMSE at the national scale.

3 Results and discussion

3.1 Comparison of different models

A two-layer stacking structure was finally generated by AutoGluon for county-level population density estimation in China, which used 11 individual machine learning models and one weighted ensemble model. Table 2 shows the comparison of performance between different models.

In the training phase, the NeuralNetFastAI model is the highest scoring among all the 11 individual models with an RMSD of 752.12, followed by ExtraTreesMSE (RMSD = 841.74), CatBoost (847.61), LightGBMXT (850.97), and LightGBMLarge (862.86). The weighted ensemble model that combines prediction from all the individual models achieved the best prediction performance with an RMSD of 728.37. Also, the training time varied significantly from 0.01 to 32.52 seconds for different models (Table 2), with a mean time of 8.19 seconds and a standard deviation of 11.06 seconds. Models of K-Neighbors take the least training time, while Neural Network models are the most difficult ones to train. The total time used for weighted ensemble model training is 79.63 seconds.

Four criteria were further used to assess the testing performances of models. Among all the individual models, the NeuralNetFastAI has the best performance in RMSD (497.58), %RMSD (55.86), and r^2 (0.967), and the ExtraTreesMSE received the lowest MAE (151.45). On the other hand, the weighted ensemble model has the best performance in all criteria, and the received RMSD, %RMSD, MAS, and r^2 are 443.01, 49.74, 158.65, and 0.974, respectively. At the same time, we found that the increase in the number of layers did not contribute

significantly to the improvement in prediction accuracy in this study. For example, when we increased the layer number of stacking from 2 to 3, the model performance just increased 0.005 in r^2 (and 4.45 in RMSD).

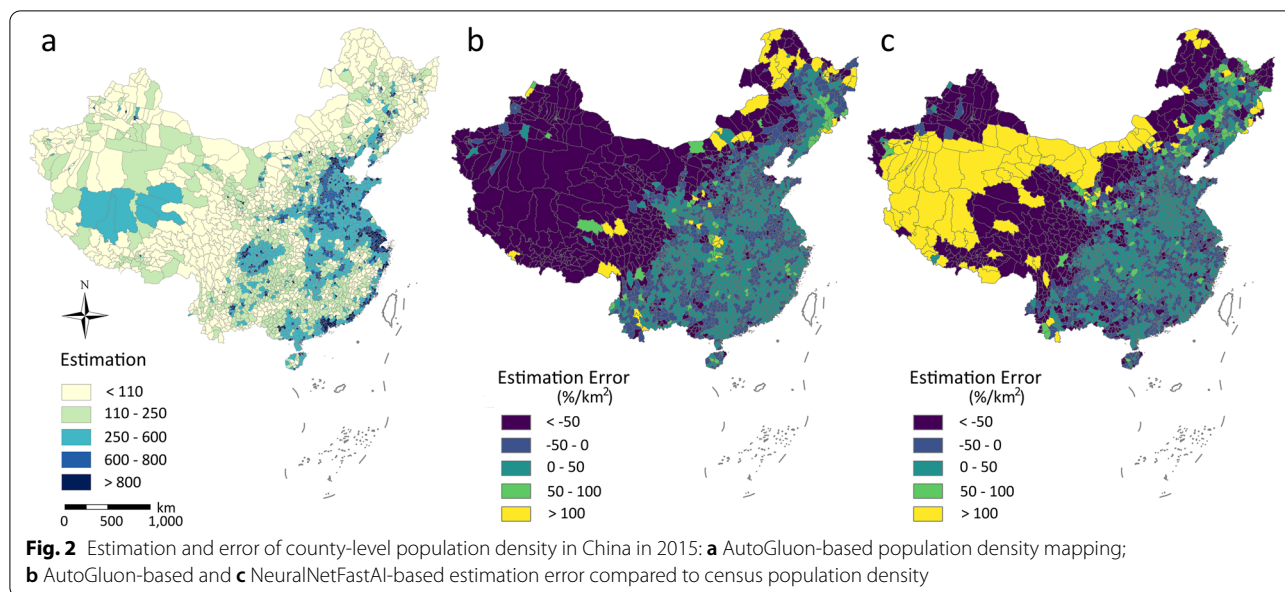
3.2 Population density mapping for China

We estimated the population density for 2851 Chinese counties using the generated weighted ensemble model. The high goodness of fit ($r^2=0.974$) and low RMSD (427.61) for all counties indicate that the inclusive features and the weighted ensemble model can well estimate county-level population density. As shown in Fig. 2a, the derived map could accurately characterize the numerical and spatial distribution patterns of county population density in China. According to the map, we can identify that, except for the counties within or around provincial capitals, counties with high population density are mainly concentrated in the eastern and southeastern coastal regions. In particular, counties located in provinces of Hebei, Henan, Shandong, Zhejiang, and Jiangsu form the main high-density population concentration areas in China. In addition, counties located in the Sichuan Basin (mainly Sichuan and Chongqing) have higher population densities in southwest China.

The difference between census population density and estimation was then calculated via (census- estimation) / census * 100%. As shown in Fig. 2b, due to the good prediction capability of the weighted ensemble model, the vast majority of the estimation errors are within 50%/km². Some underestimations appear in counties with higher population density, while counties with overestimation are mainly concentrated in lower densely populated regions in western China, such as regions (or

Table 2 Performance of different machine learning models in terms of root mean squared error (RMSE), relative root mean square error (%RMSE), mean absolute (MAE), and R-squared (r^2)

No	Models	Training performance		Testing performance			
		Time (s)	RMSD	RMSD	%RMSD	MAE	r^2
1	Weighted Ensemble Model (Level 2)	79.63	728.37	443.01	49.74	158.65	0.974
2	NeuralNetFastAI	5.7	752.12	497.58	55.86	176.08	0.967
3	ExtraTreesMSE	0.36	841.74	587.01	65.91	151.45	0.954
4	CatBoost	25.13	847.61	634.39	71.22	167.26	0.947
5	LightGBMXT	2.81	850.97	637.64	71.59	188.71	0.946
6	LightGBMLarge	32.52	862.86	704.28	79.07	190.33	0.934
7	RandomForestMSE	0.88	865.15	665.95	74.77	166.97	0.941
8	LightGBM	2.37	879.14	601.34	67.51	173.86	0.952
9	NeuralNetTorch	13.28	930.18	664.34	74.59	201.09	0.942
10	XGBoost	7.08	930.74	622.42	69.88	165.55	0.949
11	KNeighborsDist	0.01	2142.95	2124.85	238.56	549.34	0.403
12	KNeighborsUnif	0.01	2310.01	2150.67	241.46	554.72	0.388



provinces) of Tibet, Xinjiang, and Qinghai. We also performed the same estimation using the NeuralNetFastAI model which achieved the best performance second to weighted ensemble model in both training and testing processes. Compared with the NeuralNetFastAI-based result (Fig. 2c), the AutoGluon framework not only has better fitting accuracy, but also alleviates the severe over-estimation in less populated areas in the western region. Statistically, 1483 counties showed underestimation, accounting for 51.91% of all the selected counties.

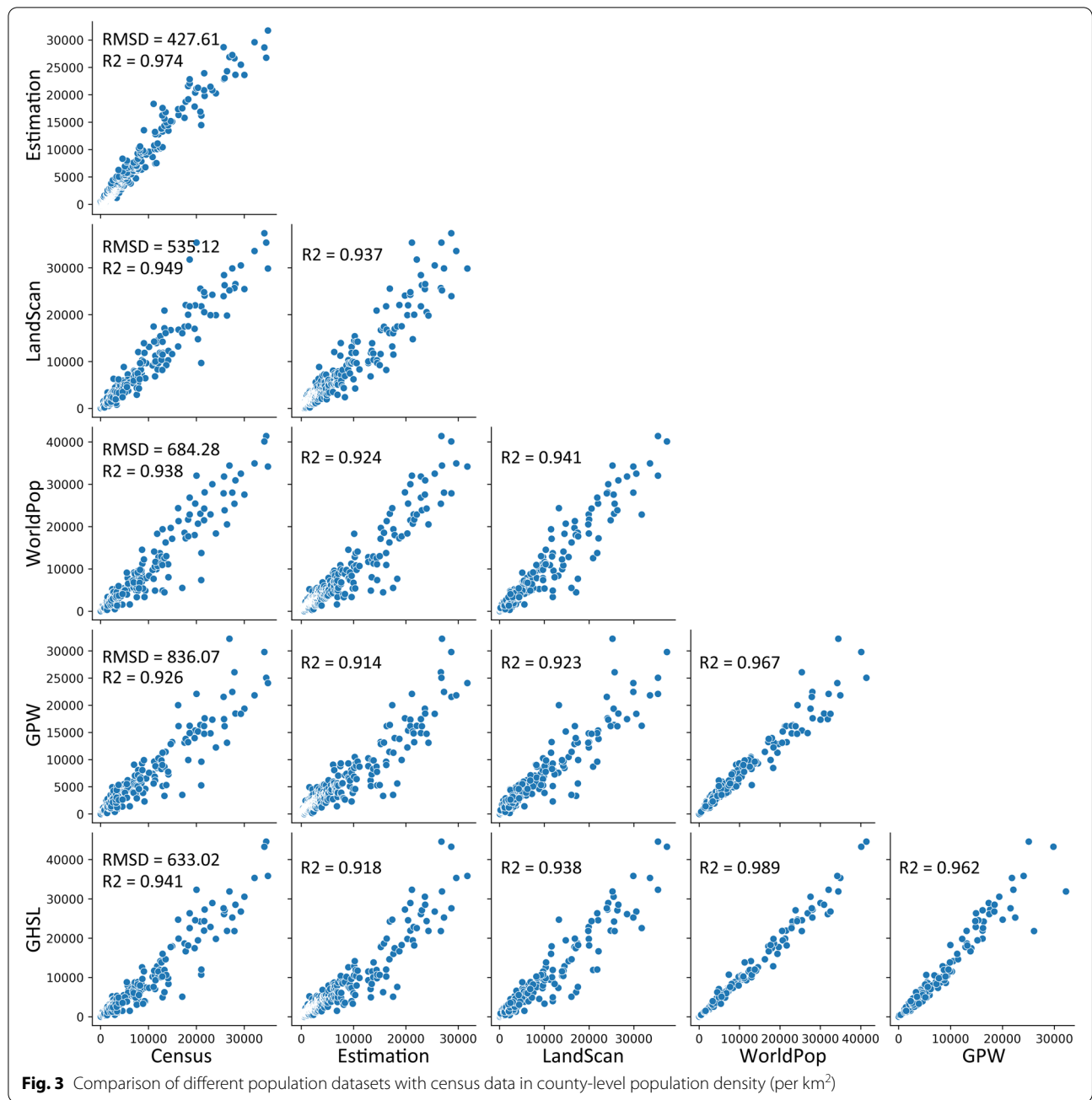
We further compared the estimated county-level population density with the other four population datasets, namely LandScan, WorldPop, GPW, and GHSL. The accuracies of these datasets were evaluated by RMSD based on census population data, and the relationship between any two of them were measured by r^2 . As the scatterplot matrix shown in Fig. 3, the population density estimated by the proposed model in this study obtained the best performance in both r^2 and RMSD, followed by LandScan ($r^2=0.949$, RMSD=535.12), GHSL ($r^2=0.941$, RMSD=633.02), WroldPop ($r^2=0.938$, RMSD=684.28), and GPW ($r^2=0.926$, RMSD=836.07). The results also show that all tested population datasets show good reliability in characterizing the spatial pattern of population density in China, with LandScan and GHSL data having better accuracy.

3.3 Importance of inclusive features

Table 3 lists inclusive features' importance scores measured by RMSE. "Importance" and "Standard deviation" mean the estimated importance score and standard deviation of a feature. " p -value" measures confidence level of

importance score, and a p -value of 0.01 indicates a 1% probability that the feature negatively affects the prediction of the model. "p99_high" and "p99_low" refer to the upper and lower end of 99% confidence interval for true feature importance score, respectively.

According to Table 3, features of the "mean" value in different categories play a more critical role in the county-level population density prediction. Nine of the top ten features with the highest importance score are mean values. Of all the inclusive features, mean Tencent LBS (LBS_mean), urban area coverage (Urban_mean), mean nighttime light brightness (NTL_mean), and mean normal difference built-up index (NDBI_mean) are the four most important features for prediction with both high importance score and low p _value. This result makes perfect sense because of the following reasons. (1) The Tencent LBS data generated by active mobile phone users and record their real-time locations, obviously having an excellent ability to characterize the population distribution. (2) Similarly, the brightness of nighttime light represents the intensity of human socio-economic activity. Except for industrial facilities such as docks and power plants, areas with high lighting levels usually have high densities of human activity. (3) Urban areas are often accompanied by relatively high population density, making it easy to explain that higher urban coverage results in higher population densities. (4) The NDBI could be considered as a representation of the intensity of human modification of the land surface, and a higher NDBI will surely refer to higher population density. More importantly, the significant contribution of variables' mean value to model performance is a valuable finding



of this study. Specifically, the more important contribution of the “mean” value makes the proposed framework reasonable and interpretable, as the observation (ground true for training) and prediction target in the model were set as population density rather than total population of a county. Simply speaking, the “mean” of features well predict the “mean” of the population (i.e., population density). In addition, the less importance of “sum” values provides vital evidence that the proposed framework has potential for cross-scale prediction, for example, the

applicability to town and fine-grid-level (e.g., 1-km) population density estimation.

Besides, negative importance scores were also found in six features, namely NDVI_sum, Grassland_mean, Rural_sum, Forest_sum, DEM_sum, and Slop_sum, which means they are likely to play a harmful role in the prediction results. However, the *p*-values of all these features’ importance scores are larger than 0.5, indicating that their harmful effect should be very weak. We then retained the model by removing these features and found

Table 3 Importance score of inclusive features in terms of root mean squared error (RMSE)

Variables	Importance	Standard deviation	p-value	p99_high	p99_low
LBS_mean	914.42	149.19	0.000082	1221.61	607.23
Urban_mean	710.37	69.47	0.000011	853.42	567.32
NTL_mean	555.92	49.14	0.000007	657.12	454.73
NDBI_mean	73.61	47.72	0.013035	171.86	-24.64
Cropland_mean	67.32	12.62	0.000141	93.29	41.35
Rural_mean	64.87	38.37	0.009717	143.88	-14.14
NDVI_mean	62.79	30.94	0.005256	126.49	-0.91
NDWI_mean	59.93	23.35	0.002285	108.02	11.84
Grassland_sum	49.84	33.82	0.015036	119.48	-19.80
Water_mean	35.18	16.96	0.004876	70.12	0.25
NTL_sum	32.70	35.30	0.053542	105.38	-39.98
Urban_sum	23.29	36.11	0.111303	97.64	-51.05
NDWI_sum	21.01	18.80	0.033410	59.71	-17.69
Forest_mean	11.08	26.63	0.202458	65.91	-43.75
Slope_mean	9.70	4.30	0.003615	18.55	0.86
NDBI_sum	7.85	12.01	0.108925	32.58	-16.88
DEM_mean	6.33	5.36	0.028717	17.36	-4.70
MP_sum	3.57	11.82	0.268335	27.90	-20.76
Cropland_sum	2.50	2.21	0.032038	7.04	-2.04
Water_sum	1.01	4.23	0.311156	9.73	-7.71
NDVI_sum	-0.05	10.25	0.504266	21.05	-21.16
Grassland_mean	-1.58	6.39	0.694978	11.58	-14.74
Rural_sum	-6.72	4.41	0.986466	2.36	-15.79
Forest_sum	-11.23	11.73	0.950494	12.92	-35.37
DEM_sum	-11.30	12.12	0.947263	13.66	-36.26
Slope_sum	-16.87	10.71	0.987811	5.17	-38.91

that the prediction performance remained almost the same ($r^2 = 0.973$, $RMSE = 448.13$).

3.4 Intra-annual population density dynamics in China

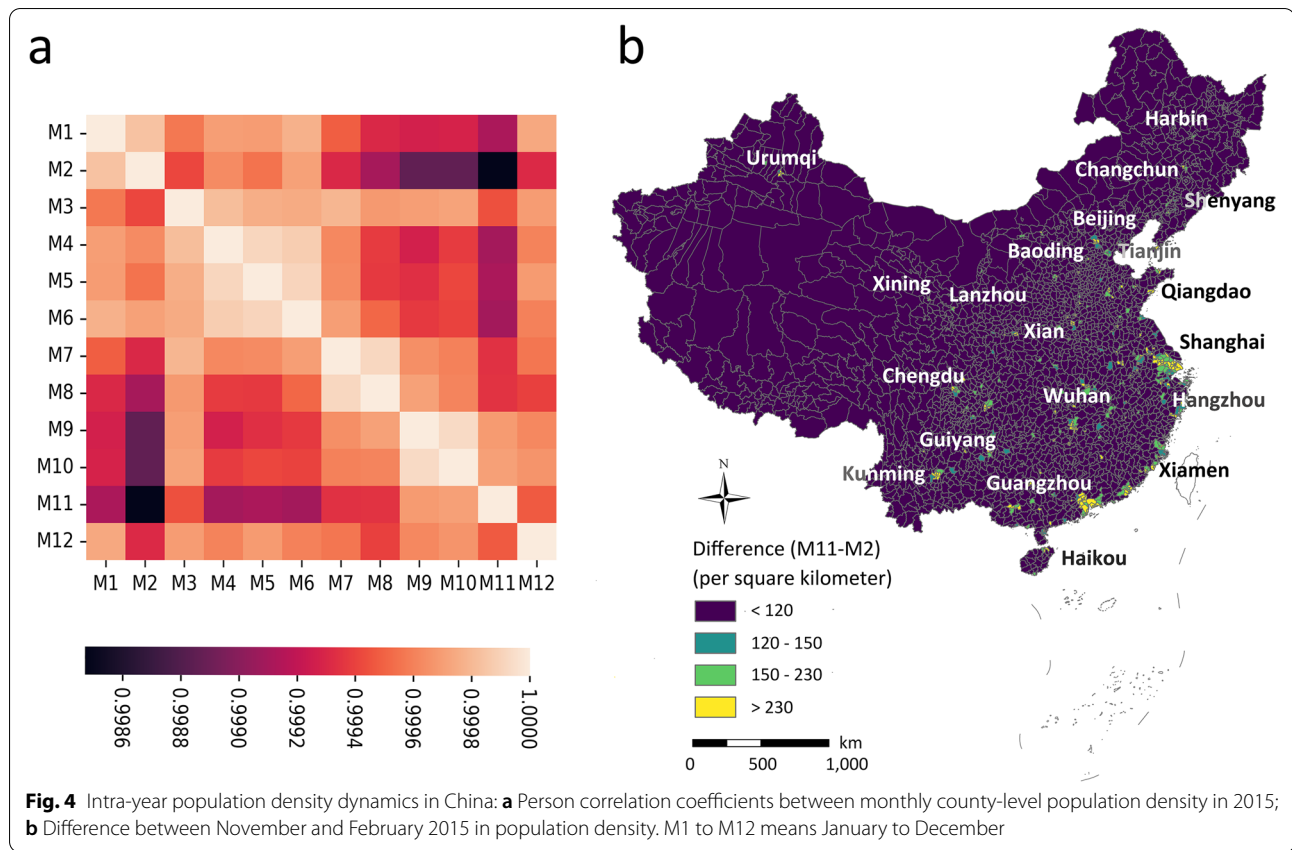
We used the weighted ensemble model to estimate county-level population density for months other than

November 2015, and only changed the mean and sum of Tencent LBS data to the corresponding month and let the other variables be the same. Person correlation coefficients between monthly county-level population density (Fig. 4a) reveal intra-annual population distribution dynamics in China. Using November as a reference month, we can be found that September and October have the highest similarity with November, while January and February have the most significant differences from November. This result is consistent with earlier studies (Li et al., 2016; Pan & Lai, 2019) and reasonably respond to the population mobility characteristic in China. Specifically, the largest population movement in China occurs during the Spring Festival, when large numbers of labor force population and students return to their hometowns before the Spring Festival and go back to the cities where they work within a short period after the Spring Festival. The spring festival of 2015 is February 19, which explains the relatively large differences between February and the other months, especially from September to November.

We further calculated the difference in population density between November and February ($population_density_{Nov} - population_density_{Feb}$) to uncover the spatial pattern of population distribution dynamics during the spring festival in 2015. As shown in Fig. 4b, in general, the most dramatic population density changes occur in areas that are more economically developed. These cities or towns are more attractive to the labor force because of more job opportunities and high salaries. Specifically, cities with significant declines in population density include three major economic zones of the Pearl River Delta city cluster (e.g., Guangdong, Shenzhen, Dongguan, Foshan, and Zhuhai), the Yangtze River Delta city cluster (e.g., Shanghai, Wuxi, Suzhou, Ningbo, Changzhou, Hangzhou, Taizhou, and Wenzhou), and the Beijing-Tianjin-Hebei city cluster (e.g., Beijing, Tianjin, Zhangjiakou, Tangshan, and Baoding), provincial capital cities (e.g., Shenyang, Jilin, Harbin, Jinan, Fuzhou, Haikou, Nanning, Kunming, Guiyang, Chengdu, Wuhan, Changsha, Xian, Zhengzhou, Xining, Lanzhou, and Urumqi), eastern coastal cities (e.g., Dalian, Yantai, Qingdao, and Xiamen), and Chongqing. This result is generally consistent with the results of earlier studies using different study designs (Zhou et al., 2020; Zhu et al., 2021), illustrating the good performance of our estimated monthly population densities in characterizing population spatial distribution dynamics in China.

3.5 Advantages and limitations

Using China as an example, this study presents a new framework for population spatialization using multi-source geospatial data. The framework alleviates some of the problems arising in current population estimation



methods and offers the potential to extend to estimations with larger spatial scales for the following reasons. First, using publicly available, easily accessible, and limited data sources. Only five categories of data are used in this framework, including Landsat-8, nighttime light data, land-use map, DEM map, and Tencent LBS data, which are all globally available. The collection and pre-processing of some data can also be performed quickly and efficiently on cloud computing platforms, such as Google Earth Engine, NASA Earth Exchange, Amazon AWS, and PIE-Engine. While the Tencent data values only apply to the China study, other digital footprint data alternatives can be used to map countries and regions outside China, such as geo-tagged Twitter data. Second, employing cost-effective automatic ensemble learning models. AutoGluon not only achieves plausible accuracies in population mapping but also dramatically reduces the cost of model selection and parameter tuning, making the estimation framework proposed in this study easily reproducible and refinable by other scholars. Third, transferability of prediction across time and regions. The satisfactory results of monthly mapping and the revealed characteristics of the intra-annual population dynamics in China suggest that the proposed model has the ability to conduct across-time with only the change of social

sensing data (population digital footprints of population) and without the influence of changes in the quality of remote sensing data (e.g., cloudiness, phenology), which could directly contribute to high-temporal-resolution population mapping (e.g., seasonal, monthly, daily). Moreover, the proposed framework's stable performance and overall high accuracy affirm its transferability across regions. Therefore, for regions where it is challenging to train models due to the lack of observation data, well-trained models from other regions would have the potential to result in good accuracy levels.

Nevertheless, some limitations in this framework study should be pointed out. (1) Since we do not have quality-assured observations at smaller scales, such as the community scale, the performance of the proposed framework at small scales is unknown. (2) The estimated bias in China shows a spatial pattern with overestimation in the west (low population density areas) and underestimation in the east (high population density areas), which is an issue could be addressed. We will attempt to solve this problem in our future work by adding variables or introducing spatial information as appropriate.

4 Conclusions

Using AutoGluon and multi-source geospatial data, this study proposed an efficient framework for population spatialization. Based on this framework, we estimated the county-level population density in China, using a limited number of publicly available datasets ranging from Tencent LBS data, Landsat-8 OLI imagery, nighttime light data, land-use maps, and DEM map. The result showed that the proposed framework could well estimate the population density for a total of 2851 counties in China with a high goodness of fit (r^2) of 0.974 and a low RMSD of 427.61. The comparisons with WorldPop, LandScan, GPW, and GHSL data also illustrate that the framework outperforms the current mainstream population mapping frameworks in terms of estimation accuracy. Of all the features involved in the modeling, mean Tencent LBS, urban area coverage, mean nighttime light brightness, and mean normal difference built-up index are the four features that contribute most to the improvement of estimation capacity. Furthermore, the derived monthly county-level population density and the revealed spatial pattern of population dynamics in China are consistent with earlier studies, corroborating the robustness of the proposed framework. This study is the first to apply AutoGluon to population estimation and mapping, and its efficient and automated modeling capabilities will undoubtedly contribute to larger scale and finer spatial-temporal population spatialization studies.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (Grant NO. 42001385). The authors would like to thank two anonymous reviewers and editors for providing valuable suggestions and comments, which have greatly improved this manuscript.

Authors' contributions

Yimeng Song: Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft. Yong Xu: Writing – review & editing. Bin Chen: Writing – review & editing. Qingqing He: Writing – review & editing. Ying Tu: Writing – review & editing. Fei Wang: Writing – review & editing. Jixuan Cai: Writing – review & editing. The author(s) read and approved the final manuscript.

Funding

This work was supported by a grant from the National Natural Science Foundation of China (Grant NO. 42001385).

Availability of data and materials

The datasets used in this research are publicly available.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of the Environment, Yale University, New Haven, CT 06511, USA. ²School of Geography and Remote Sensing, Guangzhou University, Guangzhou, China. ³Future Urbanity & Sustainable Environment (FUSE) Lab, Division of Landscape Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong, China. ⁴School of Resource and Environmental Engineering, Wuhan University of Technology, Wuhan 430070, China. ⁵Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Institute for Global Change Studies, Tsinghua University, Beijing 100084, China. ⁶Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China. ⁷Tencent Inc., Shenzhen 518057, China.

Received: 1 November 2022 Revised: 10 November 2022 Accepted: 13 November 2022

Published online: 25 November 2022

References

- Bagan, H., & Yamagata, Y. (2015). Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience & Remote Sensing*, 52(6), 765–780. <https://doi.org/10.1080/15481603.2015.1072400>.
- Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 28(9), 1940–1963. <https://doi.org/10.1080/13658816.2014.909045>.
- Bright, E., Rose, A., & Urban, M. (2016). *LandScan Global 2015 Version 2015 [raster digital data]*. Oak Ridge National Laboratory. <https://doi.org/10.48690/1524210>.
- Cai, J., Huang, B., & Song, Y. (2017). Using multi-source geospatial big data to identify the structure of polycentric cities. *Remote Sensing of Environment*, 202, 210–221. <https://doi.org/10.1016/j.rse.2017.06.039>.
- Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). *Ensemble selection from libraries of models proceedings of the twenty-first international conference on machine learning*. Banff, Alberta, Canada. <https://doi.org/10.1145/1015330.1015432>.
- Chen, B., Song, Y., Huang, B., & Xu, B. (2020). A novel method to extract urban human settlements by integrating remote sensing and mobile phone locations. *Science of Remote Sensing*, 1, 100003. <https://doi.org/10.1016/j.srs.2020.100003>.
- Chen, B., Song, Y., Jiang, T., Chen, Z., Huang, B., & Xu, B. (2018). Real-time estimation of population exposure to PM2.5 using Mobile- and station-based big data. *International Journal of Environmental Research and Public Health*, 15(4), 573 <https://www.mdpi.com/1660-4601/15/4/573>.
- Chen, B., Song, Y., Kwan, M.-P., Huang, B., & Xu, B. (2018). How do people in different places experience different levels of air pollution? Using worldwide Chinese as a lens. *Environmental Pollution*, 238, 874–883. <https://doi.org/10.1016/j.envpol.2018.03.093>.
- Chen, B., Wu, S., Jin, Y., Song, Y., Wu, C., Venevsky, S., ... Gong, P. (2022). *Wildfire risk for global wildland–urban interface (WUI) areas*. Natural Sustainability, urban review. <https://doi.org/10.21203/rs.3.rs-2147308>.
- Chen, B., Wu, S., Song, Y., Webster, C., Xu, B., & Gong, P. (2022). Contrasting inequality in human exposure to greenspace between cities of global north and global south. *Nature Communications*, 13(1), 4636. <https://doi.org/10.1038/s41467-022-32258-4>.
- Cheng, Z., Wang, J., & Ge, Y. (2022). Mapping monthly population distribution and variation at 1-km resolution across China. *International Journal of Geographical Information Science*, 36(6), 1166–1184. <https://doi.org/10.1080/13658816.2020.1854767>.
- Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), 15888–15893. <https://doi.org/10.1073/pnas.1408439111>.
- Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistoletti, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3), 226–234. <https://doi.org/10.1080/23754931.2015.1014272>.

- Elvidge, C. D., Baugh, K. E., Dietz, J. B., Bland, T., Sutton, P. C., & Kroehl, H. W. (1999). Radiance calibration of DMSP-OLS low-light imaging data of human settlements. *Remote Sensing of Environment*, 68(1), 77–88. [https://doi.org/10.1016/S0034-4257\(98\)00098-4](https://doi.org/10.1016/S0034-4257(98)00098-4).
- Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505.
- Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., ... Alsdorf, D. (2007). The shuttle radar topography Mission. *Reviews of Geophysics*, 45(2). <https://doi.org/10.1029/2005RG000183>.
- Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., & Mills, J. (2016). *Development of new open and free multi-temporal global population grids at 250 m resolution*.
- Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., ... Xu, B. (2020). Mapping essential urban land use categories in China (EULUC-China): Preliminary results for 2018. *Science Bulletin*, 65(3), 182–187. <https://doi.org/10.1016/j.scib.2019.12.007>.
- He, L., Shen, J., & Zhang, Y. (2018). Ecological vulnerability assessment for ecological conservation and environmental management. *Journal of Environmental Management*, 206, 1115–1125. <https://doi.org/10.1016/j.jenvman.2017.11.059>.
- He, Q., Gao, K., Zhang, L., Song, Y., & Zhang, M. (2021). Satellite-derived 1-km estimates and long-term trends of PM_{2.5} concentrations in China from 2000 to 2018. *Environment International*, 156, 106726. <https://doi.org/10.1016/j.envint.2021.106726>.
- Li, J., Ye, Q., Deng, X., Liu, Y., & Liu, Y. (2016). Spatial-temporal analysis on spring festival travel rush in China based on multisource big data. *Sustainability*, 8(11), 1184 <https://www.mdpi.com/2071-1050/8/11/1184>.
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490), 489–493. <https://doi.org/10.1126/science.abb3221>.
- Linard, C., Gilbert, M., & Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal*, 76(5), 525–538. <https://doi.org/10.1007/s10708-010-9364-8>.
- Liu, X., Huang, B., Li, R., & Wang, J. (2021). Characterizing the complex influence of the urban built environment on the dynamic population distribution of Shenzhen, China, using geographically and temporally weighted regression. *Environment and Planning B: Urban Analytics and City Science*, 48(6), 1445–1462. <https://doi.org/10.1177/23998083211017909>.
- Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3), 512–530. <https://doi.org/10.1080/00045608.2015.1018773>.
- Liu, Z., Ma, T., Du, Y., Pei, T., Yi, J., & Peng, H. (2018). Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS*, 22(2), 494–513. <https://doi.org/10.1111/tgis.12323>.
- Long, Y., Song, Y., & Chen, L. (2022). Identifying subcenters with a nonparametric method and ubiquitous point-of-interest data: A case study of 284 Chinese cities. *Environment and Planning B: Urban Analytics and City Science*, 49(1), 58–75. <https://doi.org/10.1177/2399808321996705>.
- Ma, Y., Xu, W., Zhao, X., & Li, Y. (2017). Modeling the hourly distribution of population at a high spatiotemporal resolution using Subway smart card data: A case study in the central area of Beijing. *ISPRS International Journal of Geo-Information*, 6(5), 128 <https://www.mdpi.com/2220-9964/6/5/128>.
- National Bureau of Statistics of China (2016). *China statistical yearbook 2016*. China Statistics Press <http://www.stats.gov.cn/tjsj/ndsj/2016/index.htm>.
- O'Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K., ... Pichs-Madruga, R. (2020). Achievements and needs for the climate change scenario framework. *Nature Climate Change*, 10(12), 1074–1084. <https://doi.org/10.1038/s41558-020-00952-0>.
- Pan, J., & Lai, J. (2019). Spatial pattern of population mobility among cities in China: Case study of the National day plus mid-autumn festival based on Tencent migration data. *Cities*, 94, 55–69. <https://doi.org/10.1016/j.cities.2019.05.022>.
- 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. *Transactions in GIS*, 21(2), 317–331. <https://doi.org/10.1111/tgis.12214>.
- Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., ... Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145, 154–172. <https://doi.org/10.1016/j.rse.2014.02.001>.
- Song, Y. (2019). Dynamic exposure. In *Inequality and urbanization effects: A multidimensional evaluation of Urban greenspace in China*. The Chinese University of Hong Kong (Hong Kong).
- Song, Y., Chen, B., Ho, H. C., Kwan, M.-P., Liu, D., Wang, F., ... Song, Y. (2021). Observed inequality in urban greenspace exposure in China. *Environment International*, 156, 106778. <https://doi.org/10.1016/j.envint.2021.106778>.
- Song, Y., Chen, B., & Kwan, M.-P. (2020). How does urban expansion impact people's exposure to green environments? A comparative study of 290 Chinese cities. *Journal of Cleaner Production*, 246, 119018. <https://doi.org/10.1016/j.jclepro.2019.119018>.
- Song, Y., Huang, B., Cai, J., & Chen, B. (2018). Dynamic assessments of population exposure to urban greenspace using multi-source big data. *Science of the Total Environment*, 634, 1315–1325. <https://doi.org/10.1016/j.scitotenv.2018.04.061>.
- Song, Y., Huang, B., He, Q., Chen, B., Wei, J., & Mahmood, R. (2019). Dynamic assessment of PM_{2.5} exposure and health risk using remote sensing and geo-spatial big data. *Environmental Pollution*, 253, 288–296. <https://doi.org/10.1016/j.envpol.2019.06.057>.
- Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS One*, 10(2), e0107042. <https://doi.org/10.1371/journal.pone.0107042>.
- Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, 4(1), 170004. <https://doi.org/10.1038/sdata.2017.4>.
- Tian, H., Liu, Y., Li, Y., Wu, C.-H., Chen, B., Kraemer, M. U. G., ... Dye, C. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, 368(6491), 638–642. <https://doi.org/10.1126/science.abb6105>.
- Tu, W., Liu, Z., Du, Y., Yi, J., Liang, F., Wang, N., ... Wang, H. (2022). An ensemble method to generate high-resolution gridded population data for China from digital footprint and ancillary geospatial data. *International Journal of Applied Earth Observation and Geoinformation*, 107, 102709. <https://doi.org/10.1016/j.jag.2022.102709>.
- Wang, L., Wang, S., Zhou, Y., Liu, W., Hou, Y., Zhu, J., & Wang, F. (2018). Mapping population density in China between 1990 and 2010 using remote sensing. *Remote Sensing of Environment*, 210, 269–281. <https://doi.org/10.1016/j.rse.2018.03.007>.
- Wang, Z., Yue, Y., He, B., Nie, K., Tu, W., Du, Q., & Li, Q. (2021). A Bayesian spatio-temporal model to analyzing the stability of patterns of population distribution in an urban space using mobile phone data. *International Journal of Geographical Information Science*, 35(1), 116–134. <https://doi.org/10.1080/13658816.2020.1798967>.
- Xu, Y., Song, Y., Cai, J., & Zhu, H. (2021). Population mapping in China with Tencent social user and remote sensing data. *Applied Geography*, 130, 102450. <https://doi.org/10.1016/j.apgeog.2021.102450>.
- Yu, S., Zhang, Z., & Liu, F. (2018). Monitoring population evolution in China using time-series DMSP/OLS nighttime imagery. *Remote Sensing*, 10(2), 194 <https://www.mdpi.com/2072-4292/10/2/194>.
- Zeng, C., Zhou, Y., Wang, S., Yan, F., & Zhao, Q. (2011). Population spatialization in China based on night-time imagery and land use data. *International Journal of Remote Sensing*, 32(24), 9599–9620. <https://doi.org/10.1080/01431161.2011.569581>.
- Zhang, H. K., Roy, D. P., Yan, L., Li, Z., Huang, H., Vermote, E., ... Roger, J.-C. (2018). Characterization of sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. *Remote Sensing of Environment*, 215, 482–494. <https://doi.org/10.1016/j.rse.2018.04.031>.
- Zhou, T., Huang, B., Liu, X., He, G., Gou, Q., Huang, Z., & Xie, C. (2020). Spatiotemporal exploration of Chinese spring festival population flow patterns and their determinants based on spatial interaction model. *ISPRS International Journal of Geo-Information*, 9(11), 670 <https://www.mdpi.com/2220-9964/9/11/670>.

Zhu, D., Huang, Z., Shi, L., Wu, L., & Liu, Y. (2018). Inferring spatial interaction patterns from sequential snapshots of spatial distributions. *International Journal of Geographical Information Science*, 32(4), 783–805. <https://doi.org/10.1080/13658816.2017.1413192>.

Zhu, R., Wang, Y., Lin, D., Jendryke, M., Xie, M., Guo, J., & Meng, L. (2021). Exploring the rich-club characteristic in internal migration: Evidence from Chinese Chunyun migration. *Cities*, 114, 103198. <https://doi.org/10.1016/j.cities.2021.103198>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.