IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2019, accepted September 28, 2019, date of publication October 10, 2019,
date of current version October 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945000

Downscaling Census Data for Gridded Population Mapping With Geographically Weighted Area-to-Point Regression Kriging

YUEHONG CHEN[1], RUOJING ZHANG[1], YONG GE[2], (Member, IEEE), YAN JIN[3], AND ZELONG XIA[1]

[1]School of Earth Sciences and Engineering, Hohai University, Nanjing 210098, China
[2]State Key Laboratory of Resources & Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[3]School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Corresponding author: Yong Ge (gey@lreis.ac.cn)

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA 20030302, in part by the Fundamental Research Funds for the Central Universities under Grant 2019B00314, and in part by the National Natural Science Foundation of China under Grant 41701376 and Grant 41725006.

ABSTRACT Understanding human population distribution on the earth at fine scales is an increasingly need to a broad range of geoscience fields, including resource allocation, transport and city planning, infectious disease assessment, disaster risk response, and climate change. Many approaches have been developed to spatially downscale census data to gridded population distribution datasets, which are preferable to integration with natural and socio-economic variables. We present a novel population downscaling approach that geographically weighted area-to-point regression kriging technique is used to downscale census data to gridded population distribution datasets with multisource geospatial and social sensing data. As a case study in Nanjing city, China we evaluated the effectiveness of the proposed population downscaling approach. The experimental results demonstrated that the proposed approach generated more accurate details of population distribution and higher accuracy than existing widely-used gridded population distribution products. Hence, the proposed population downscaling approach is a valuable option in producing gridded population distribution maps.

INDEX TERMS Gridded population distribution, census data, geospatial data, social sensing data, geographically weighted area-to-point regression kriging, downscaling, geographical information science.

I. INTRODUCTION

Human population is a critical indicator in human-environment interactions [1]–[3]. Accurate human population distribution is one of the most important variables for a broad range of geoscience fields, such as resource allocation [4], [5], transport and city planning [6]–[8], infectious disease assessment [9]–[11], disaster risk response [12], [13], and climate change amongst others [14], [15]. Population distribution data provide explicit population knowledge that where and how many people spatially distribute within a region of interest [16], [17]. Generally available information on human population is demographical data, describing population counts, structure and other information within each defined statistical unit (e.g., administrative unit, post-code zone and census tract). Census data are the main source of demographical data and they have some limitations in geoscience applications [16]–[18]. First of all, census data only provide a single value of population counts for each census unit; hence, it cannot specify the spatial population distribution within each census unit and reflect the internal population variation. Furthermore, the unit of census data is sometimes inconsistent with the unit of socioeconomic variables and the zone system of natural variables (e.g., layers derived from remote sensing images), which is known as the change of support problem. Thus, the redistribution of census data is strongly required to generate gridded population distribution data [3], [4], [8]. Gridded population distribution data

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

have three advantages over census data [18]. First, gridded population data are generated by downscaling each unit of irregular census data to several fine regular grid cells or pixels within the census unit and it provides the detailed spatial distribution of population counts within each census unit using these fine regular grid cells. Second, gridded population data are stored using the raster format and are facilitate to compute when comparing with the census data using vector format. Third, gridded population data with raster format are easy to integrate with other geoscience variables in both raster-format and vector-format. For other raster geoscience variables, the simple resample and scale transformation processes can be taken to meet the consistent requirement of data spatial resolution. For other vector geoscience variables, the aggregation process can be employed to summarize the values of grid cells within each vector unit. Consequently, gridded population data are becoming increasingly used in various applications to replace census data [1], [3].

Over the past few decades, various approaches have been developed to downscale the irregular census data to gridded population distribution maps at fine scales. The simple areal weighting method was earlier proposed to assume the population counts of a grid cell is proportional to its land area and redistribute population counts to an uniformly continuous surface within each census unit, as used in Gridded Population of the World (GPW, version 2-4) [19]–[21]. The simple areal weighting method did not use ancillary data except census data and its spatial boundaries, leading to its limited accuracy for large census units. Subsequently, many dasymetric mapping methods were developed to incorporate ancillary data to improve the details of gridded population maps. The Global Rural-Urban Mapping Project (GRUMP) incorporated the rural-urban extents derived from satellite nightlight data into the areal weighting method [2]. The LandScan global database, refreshed annually, used a multivariate dasymetric modeling method with ancillary information extracted from geographic data and remote sensing data [22], [23] Jia, *et al.* [24] proposed a dasymetric method based on the heuristic sampling by integrating tax parcel data. Azar, *et al.* [25] used a classification and regression tree to model the relationship between impervious surface ratio and population for improving the gridded population mapping. With substantially multisource geospatial data (e.g., satellite nightlight data, land cover [26], OpenStreetMap-derived data [27], and topography) as inputs, WorldPop population distribution project developed a semi-automated dasymetric method based on random forest regression to produce gridded population maps at both 30 arc-seconds (∼1 km at the equator) and 3 arc-seconds (∼100 m at the equator) spatial resolutions [1].

Recently, efforts have been made to refine the redistribution process with a proliferation of social sensing big data [28], including mobile phone data, taxi GPS trajectories, geo-located tweets, positioning density of mobile phone Apps, etc. The mobile phone location data, geolocated call records to a signal tower of mobile phone network, were used to model the dynamic gridded population distribution [4], [7], [29]. Yu, *et al.* [30] combined taxi GPS trajectories and satellite nightlight data to yield the gridded population maps at 500 m in Shanghai, China. Patel, *et al.* [31] used the density of geo-located tweets from the social media application of Twitter in 1 × 1 km grid cells as a covariate to improve the accuracy of gridded population maps by the random forest regression algorithm. The real-time Tencent user density, collected from Tencent's Apps with location-based service (e.g., the online map service (https://map.qq.com/), the largest online social network chatting Apps of QQ and WeChat in China, and other smart phone Apps), was fused with points of interest to predict the population distribution at the building scale [32].

Although a variety of gridded population downscaling approaches have been developed and obtained relatively satisfactory performances over the past decades, the spatial dependence and spatial heterogeneity of geographical variables are often ignored in these approaches [3], [33]. The spatial dependence or autocorrelation, the first law of geography, indicates that spatial events are more influenced by near events than distant events [34]. That is to say spatial dependence of geographical variables is a function of distance [35]. Kriging interpolation methods are one type of widely used methods for incorporating the spatial dependence in spatial prediction [36], [37]. To deal with the change of support problem, area-to-point kriging (ATPK), a relatively new geostatistical method, was developed for spatial prediction and downscaling by incorporating the spatial dependence [38], [39] and it has been successfully applied to downscaling different geographical variables [39]–[41]. However, the spatial dependence is not considered in most of gridded population downscaling approaches. For example, the widely used regression-based approaches only use the attribution data of census data and covariates and their spatial locations are not used [1], [25], [31]. The spatial heterogeneity, the second law of geography, refers to the non-stationarity or uncontrolled variance for geographical variables [42]. It leads to the requirement of local models as global models may be inadequate to capture the local behaviors [43]. As the inevitable feature of spatial heterogeneity in geographical variables, geographically weighted regression (GWR) was developed as an effective solution to capture the local behaviors and generate the local coefficients of regression for each unit [43], [44]. But, most of these approaches built a global model in downscaling census data, such as the random forest regression [1], [31] and the classification and regression tree [25]. Geographically weighted area-to-point regression kriging (GWATPRK) is the combination of GWR and ATPK [45]. GWR is used to model the local spatial trend component of correlated variables whereas ATPK is used to model the residual component as it is spatial dependent [45]. Although there are several advantages of GWATPRK, little if any consideration has been given to the application of GWATPRK to downscale socioeconomic variables with irregular units. In addition, social sensing big

data, especially the directly indicative data of people presence from smart phone Apps with location-based service, have demonstrated the large potential in mapping population distribution [4], [30], [31], [46]. However, they are rare combined with geospatial data derived from remote sensing data and geographical information data to make full use of their own merits.

Therefore, GWATPRK is employed to evaluate its feasibility in spatially downscaling irregular census data to gridded population distribution count data at fine scales with multi-source geospatial and social sensing datasets while considering the spatial dependence and spatial heterogeneity. A case study in Nanjing city, China is taken to demonstrate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section II describes the proposed method. Section III outlines the case study, which is discussed in Section IV. Section V presents the conclusions.

II. METHOD

A. FLOWCHART OF GWATPRK-BASED GRIDDED POPULATION DISTRIBUTION MAPPING

GWATPRK-based gridded population mapping method implements in six main steps: 1) preprocess the collected geospatial data and social sensing data to prepare the inputs of generating covariates (including projection transformation, clip vector data and raster data to match the spatial extent of census data, and other operations); 2) generate covariates at fine grid scale in terms of geospatial data and social sensing data (i.e., positioning density of Tencent users); 3) generate covariates at census unit scale by aggregating the gridded covariates to each census unit; 4) perform GWR on census population data (dependent variable) and covariates (independent variables) firstly and then implement ATPK on the residual component of GWR; 5) combine the ATPK results at the fine grid scale and the volume-preserved GWR results to produce the GWATPRK-based gridded population counts distribution maps; 6) assess the accuracy of the GWATPRK-based gridded population counts distribution maps, as shown in Figure 1.

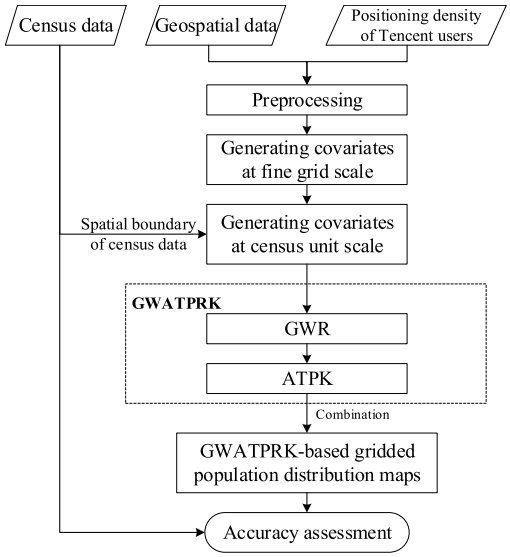


FIGURE 1. Flowchart of GWATPRK-based gridded population distribution mapping.

B. GWATPRK

Let $p(u_i); i = 1, \ldots, m$ and $p(v_j); j = 1, \ldots, n$ to be the m irregular census data and the n population grid cells in the study area, respectively. GWATPRK can be formulated as a combination of spatial trend component and residual component. Thus, the general form of the prediction of population grid cell v_j by GWATPRK model can be expressed as

$$p(v_j) = m(v_j) + e(v_j) \tag{1}$$

where $m(v_j)$ is the spatial trend component at grid cell v_j and $e(v_j)$ is the residual component at grid cell v_j.

C. ESTIMATION OF SPATIAL TREND BY GWR

To consider the spatial heterogeneity, GWR is applied to model the relationship between population data and correlated covariates. The spatial trend component of grid cell v_j can be predicted by

$$m(v_j) = \beta_0(v_j) + \sum_{k=1}^{K} x_k(v_j)\,\beta_k(v_j) \tag{2}$$

where $\beta_0(v_j)$ is the GWR intercept, $\beta_k(v_j)$ is the GWR coefficient of covariate k at grid cell v_j, and $x_k(v_j)$ is the value of covariate k at grid cell v_j.

Due to the lacking of population data at the fine grid scale, the population prediction of fine grid cells in (1) is impossible [45]. To benefit from detailed information of covariates at the fine grid scale, the invariable assumption of regression coefficients at different scales can be taken in GWR model [45], [47] like the existing gridded population mapping using random forest regression [1], [31] and classification and regression tree [25]. Thus, the coefficients of $\beta_0(\cdot)$ and $\beta_k(\cdot)$ in (2) can be replaced with the coefficients of GWR calculated at the irregular census unit scale, that is

$$p(u_i) = \beta_0(u_i) + \sum_{k=1}^{K} x_k(u_i)\,\beta_k(u_i) + e(u_i) \tag{3}$$

where $\beta_0(\cdot)$ and $\beta_k(\cdot)$ are the coefficients of GWR at the census unit scale, $x_k(u_i)$ is the value of covariate k at census unit u_i, and $e(u_i)$ is the residual component at census unit u_i.

After the GWR operation on population data and covariates at the census unit scale, the GWR coefficients can be yielded for each census unit from (3). The direct rasterization on these GWR coefficients is implemented to obtain the gridded GWR coefficients at the fine grid cell scale. The spatial trend component at grid cell v_j can be linear weighted by (2) in terms of the rasterized GWR coefficients and covariates at the fine grid scale. Note that gridded GWR results cannot meet the estimated total population constraint of GWR within each census unit (i.e., the volume preserving property) whereas ATP results can meet this constraint. Thus, the min-max

normalization was used to scale the gridded GWR results to yield the volume-preserved GWR results that meet the estimated total population constraint within each census unit.

D. ESTIMATION OF RESIDUAL BY ATPK

It is similar to the spatial trend component in (1), the residual component in (1) is also impossible to estimate for each fine grid cell directly. Indirectly, the residual component at fine grid cells can be downscaled by ATPK from the known residual component of GWR in (3) because GWR residuals often have spatial dependence [45]. By considering the spatial dependence, the residual component in (1) can be estimated by ATPK as

$$\hat{e}\left(v_j\right) = \sum_{i=1}^{l} \lambda\left(u_i\right) \cdot e\left(u_i\right) \tag{4}$$

where $e\left(u_i\right)$ is the known GWR residual component in (3) at census unit u_i, $\lambda\left(u_i\right)$ is the ATPK weights of residual component at census unit u_i and $\lambda\left(u_i\right)$ is obtained by solving following kriging system

$$\begin{cases} \sum_{i=1}^{l} \lambda\left(u_i\right) \bar{C}\left(u_i, u_{i'}\right) + \mu\left(v_j\right) = \bar{C}\left(u_i, v_j\right) \\ \sum_{i=1}^{l} \lambda\left(u_i\right) = 1 \end{cases} \tag{5}$$

where $\bar{C}\left(u_i, u_{i'}\right)$ is the area-to-area covariance between two areal census units of $u_i, u_{i'}$, $\bar{C}\left(u_i, v_j\right)$ is the area-to-point covariance between census unit u_i and the fine grid cell v_j, l is the number of neighboring census units round the central census unit under consideration. Note that ATPK needs a variogram of residual component at fine grid scale and it can be estimated by a deconvolution technique from the residual component in (3) at the census scale [40], [48], [49].

After the two process of GWR and ATPK, the population prediction of grid cell v_j can be calculated by (1).

III. CASE STUDY

To evaluate the performance of GWATPRK-based population counts distribution mapping method, a case study in Nanjing city, China was carried out to produce the gridded population distribution maps at two scales of the WGS 84 geographical coordinate system (i.e., 30 arc-seconds (∼1000 m at the equator) and 3 arc-seconds (∼100 m at the equator)).

A. STUDY AREA AND DATA

The study area, Nanjing city (31°14′ –32°37′ N, 118°22′ – 119°14′ E), locates in the lower Yangtze River basin. Nanjing city is the capital of the Jiangsu province in eastern China and it has 11 districts and 130 administrative street zones with a total population of 8.24 million in 2015. It is a representative city of the rapidly growing and globalizing in the Yangtze River Delta economic zone. The gridded population distribution data at fine scales are useful to the urban planning and development of Nanjing city; hence, it is selected as the case study area.

The census data in 2015 of Nanjing city were collected from the yearbook and the official website of each district. Ancillary data that may have correlation with population distribution were collected from different ways to derive covariates [3]. They included points of interest (POIs), roads, water, digital elevation model (DEM), Luojia 1-01 nighttime light image, impervious surface ratio, height of buildings, and positioning density of Tencent users, as shown in Figure 2. The covariates were derived from the ancillary data as follows.

(1) POIs are geographical points that identify the significant places of human activity [50]. POIs in Nanjing were acquired from one of the biggest online maps in China (https://map.qq.com/) on 6 July 2017 [50] and the total number was 306,517, as shown in Figure 2(a). POIs were used to derive the covariate of gridded kernel density maps at two scales of 30 arc-seconds and 3 arc-seconds by the Kernel Density tool in ArcMap 10.2 software.

(2) Roads were obtained from OpenStreetMap (http://download.geofabrik.de/index.html), as shown in Figure 2(b). Roads were employed to derive two covariates of road density and road accessibility at the two scales by the Line Density and Euclidean Distance tools in ArcMap 10.2 software, respectively.

(3) Water was obtained from OpenStreetMap, as shown in Figure 2(c). The gridded water accessibility maps at the two scales were derived as a covariate by the Euclidean Distance tool in ArcMap 10.2 software.

(4) DEM data with the spatial resolution of 30 m (see Figure 2(d)) were the ASTER GDEM data and were collected from Geospatial Data Cloud platform (http://www.gscloud.cn/). DEM and its derived slope data were aggregated to yield two covariates at the two scales.

(5) Luojia 1-01 nighttime light image, acquired on July 15, 2018, was used as a covariate, as shown in Figure 2(e). Luojia 1-01, launched on June 2, 2018, is a CubeSat (6U) sized earth observation satellite designed by the Wuhan University and it features an image with about 130 m ground resolution [51]. Luojia 1-01 nightlight images have been proved to be more potential in population mapping than the Suomi National Polar-Orbiting Partnership Visible Infrared Imaging Radiometer Suite images (NPP-VIIRS) [51]. Therefore, a Luojia 1-01 nighttime light image was chosen as a covariate at the two scales. The original Luojia 1-01 nighttime light image was first conveted to an image at the scale of 3 arc-seconds by projection transformation and then the converted Luojia 1-01 image was aggregated to an image at the scale 30 arc-seconds.

(6) An impervious surface ratio image at 10 m spatial resolution in our previous work was selected to produce a covariate, as shown in Figure 2(f). The impervious surface ratio image was derived from S2A images taken on April 2, 2017 [50] and it was converted and aggregated to yield two images at the two scales.

(7) Height image of buildings was used as a covariate, as shown in Figure 2(g). The building Esri shapefile with the attribute of height in 2015 was collected from Jiangsu Land Surveying and Planning (http://www.jslsp.com/).

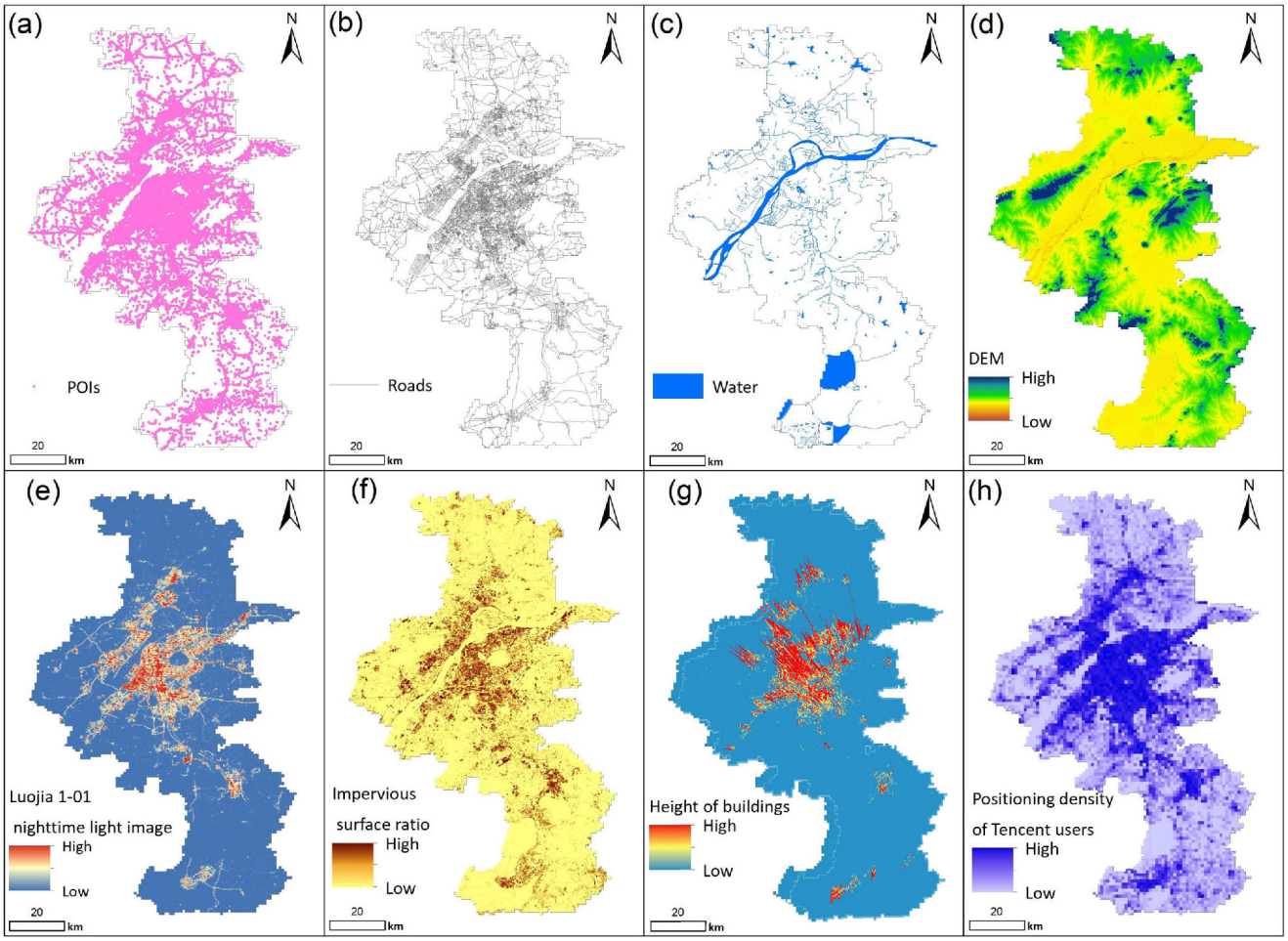


FIGURE 2. Ancillary data. (a) POIs, (b) Roads, (c) Water, (d) DEM, (e) Luojia 1-01 nighttime light image, (f) Impervious surface ratio, (g) Height image of buildings, and (h) Positioning density image of Tencent users.

The shapefile of building height was first rasterized to an image at 10 m spatial resolution and then it was converted and aggregated to yield two height images at the two scales.

(8) The real-time Tencent user positioning density image, requested by Tencent's Apps of smart phones, was collected to yield a social sensing covariate, as shown in Figure 2(h). Tencent company is one of the biggest companies in China. It has the largest online social network chatting Apps of QQ and WeChat in China and there are over 94.6% of the Chinese people who installed Tencent social network chatting Apps in 2015 [52]. Once the installed Tencent' Apps were used, their geographical locations could be recorded to calculate the number of people who have used these Apps within a cell (∼1000 m at the equator) during a time interval. This information was used to yield the real-time Tencent user positioning density image and it could be directly crawled from the website: https://heat.qq.com/. The real-time Tencent user positioning density image has been used several applications [32], [52]. Compared with aforementioned geospatial data, the social sensing data of Tencent user positioning density is able to directly indicate the detailed presence of people with the locations. Such social sensing data therefore were used to test its potential in gridded population distribution mapping. We crawled the real-time Tencent user positioning density images every five minutes from https://heat.qq.com/ between September 1 to September 30 in 2018. To avoid the effect of population mobility, the final positioning density image at the scale of 30 arc-seconds was calculated by averaging all the real-time Tencent user positioning density images in this month. The positioning density image at the scale of 30 arc-seconds was converted to the positioning density image at the scale of 3 arc-seconds.

The above generated geospatial and social sensing covariates at each scale were aggregated to census units. The Pearson correlation coefficients between census data and aggregated covariates at the scale of 30 arc-seconds were calculated in Table 1. It can be found from Table 1 that the six covariates including density of POIs, density of roads, Luojia 1-01 nighttime light image, impervious surface ratio image, height image of buildings, and positioning density of Tencent users have relatively higher Pearson correlation coefficients with census data than other four covariates of

TABLE 1. Pearson correlation coefficients between census data and covariates.

Covariate	Correlation	Covariate	Correlation
Density of POIs	0.83	Positioning density of Tencent users	0.43
Density of roads	0.51	Distance to roads	-0.01
Luojia 1-01 nighttime light image	0.56	Distance to water	0.09
Impervious surface ratio	0.43	DEM	0.15
Height of buildings	0.71	Slope	0.16

distance to roads, distance to water, DEM, and slope. Thus, the six covariates with relatively high Pearson correlation coefficients were chosen as covariates in building the model of GWATPRK.

The widely-used gridded population counts data in 2015 including GPW version 4 at the scale of 30 arc-seconds (http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/sets/browse), LandScan at the scale of 30 arc-seconds (https://landscan.ornl.gov/landscan-datasets), and WorldPop at the two scales of 30 arc-seconds (https://www.worldpop.org/geodata/listing?id = 17) and 3 arc-seconds (https://www.worldpop.org/geodata/listing?id = 16) were employed to compare with the GWATPRK-based gridded population distribution results. In addition, the GWR result at the scale of 30 arc-seconds was generated to compare with GWATPRK for using the same ancillary data.

Accuracy assessment of each gridded population counts distribution map featured a suite of metrics, including the R^2, the root mean squared error (RMSE) and the mean absolute error (MAE). During the accuracy assessment, each gridded population counts distribution map was first aggregated to census units to get the population sum of each census unit and then it was used to compare with the actual census data to calculate above metrics.

B. EXPERIMENTAL RESULTS

Figure 3(a)-(e) present the five gridded population counts distribution maps in Nanjing city at the scale of 30 arc-seconds obtained from GPW version 4, WorldPop, LandScan, GWR and GWATPRK, respectively. Figure 3(f) and (g) are the gridded population counts distribution maps at the scale of 3 arc-seconds for WorldPop and GWATPRK, respectively. Figure 3(h) shows the census data of Nanjing city in 2015. The seven maps in Figure 3(a)-(g) show that high population counts mainly distribute in the downtown areas (e.g., Gulou and Xuanwu districts), which is basically consistent with the census data and real population distribution status. In the five maps at the scale of 30 arc-seconds, GPW map and WorldPop map are similar and have relatively homogeneous spatial distribution pattern whereas other three maps of LandScan, GWR and GWATPRK are similar and have relatively heterogeneous spatial distribution pattern of population. On visual inspection of Figure 3(a)-(e), LandScan, GWR and GWATPRK maps provide more population distribution details and more accurate distribution on low population counts (<100 in one pixel) than GPW and WorldPop maps because low population counts mainly distribute in mountain areas (see Figure 2(e)) in LandScan, GWR and GWATPRK maps whereas GPW and WorldPop maps do not have these distributions. When examining LandScan, GWR and GWATPRK maps in Figure 3(c)-(e), their distribution is very close except that LandScan map has more scattered pixels with high population counts than GWR and GWATPRK maps. For the two maps at the scale of 3 arc-seconds, it can be observed from Figure 3(f) and (g) that the overall population counts of pixels are significantly lower than those of the five maps at the scale of 30 arc-seconds and that GWATPRK generated more population distribution details than WorldPop, especially in urban center and suburban areas.

TABLE 2. Accuracy comparison of seven population maps.

	R^2	RMSE	MAE
GPW (30 arc-seconds)	0.5225	41879	28088
WorldPop (30 arc-seconds)	0.6751	28750	18954
LandScan (30 arc-seconds)	0.5446	37489	24985
GWR (30 arc-seconds)	0.8621	19223	14121
GWATPRK (30 arc-seconds)	0.9983	1998	349
WorldPop (3 arc-seconds)	0.6755	29984	20413
GWATPRK (3 arc-seconds)	0.9995	1198	524

Table 2 shows the accuracy assessment for the seven gridded population distribution maps. For the three gridded population products at the scale of 30 arc-seconds, GPW has the lowest accuracy, followed by LandScan, and WorldPop is higher than GPW and LandScan, which is consistent with the accuracy assessment results in a previous study [18]. With the multisource geospatial data and social sensing data as inputs, GWR has higher accuracy than the three existing products at the scale of 30 arc-seconds while its accuracy is slightly lower than GWATPRK. The accuracy of GWATPRK is obviously greater than existing gridded population products. Specifically, the R^2 of two GWATPRK results are over 0.99, the R^2 of GWATPRK for the scale of 30 arc-seconds is 0.4177 higher than the average R^2 of GPW, WorldPop and LandScan, and the R^2 of GWATPRK for the scale of 3 arc-seconds is 0.324 greater than that of WorldPop product.

IV. DISCUSSION

A. INTERCOMPARISON OF SEVEN GRIDDED POPULATION MAPS

The performance of GWATPRK was further discussed and compared for the spatial distribution details of population counts in a subarea of urban center marked by a black rectangle in Figure 3(h). Figure 4(a)-(e) are the GPW,

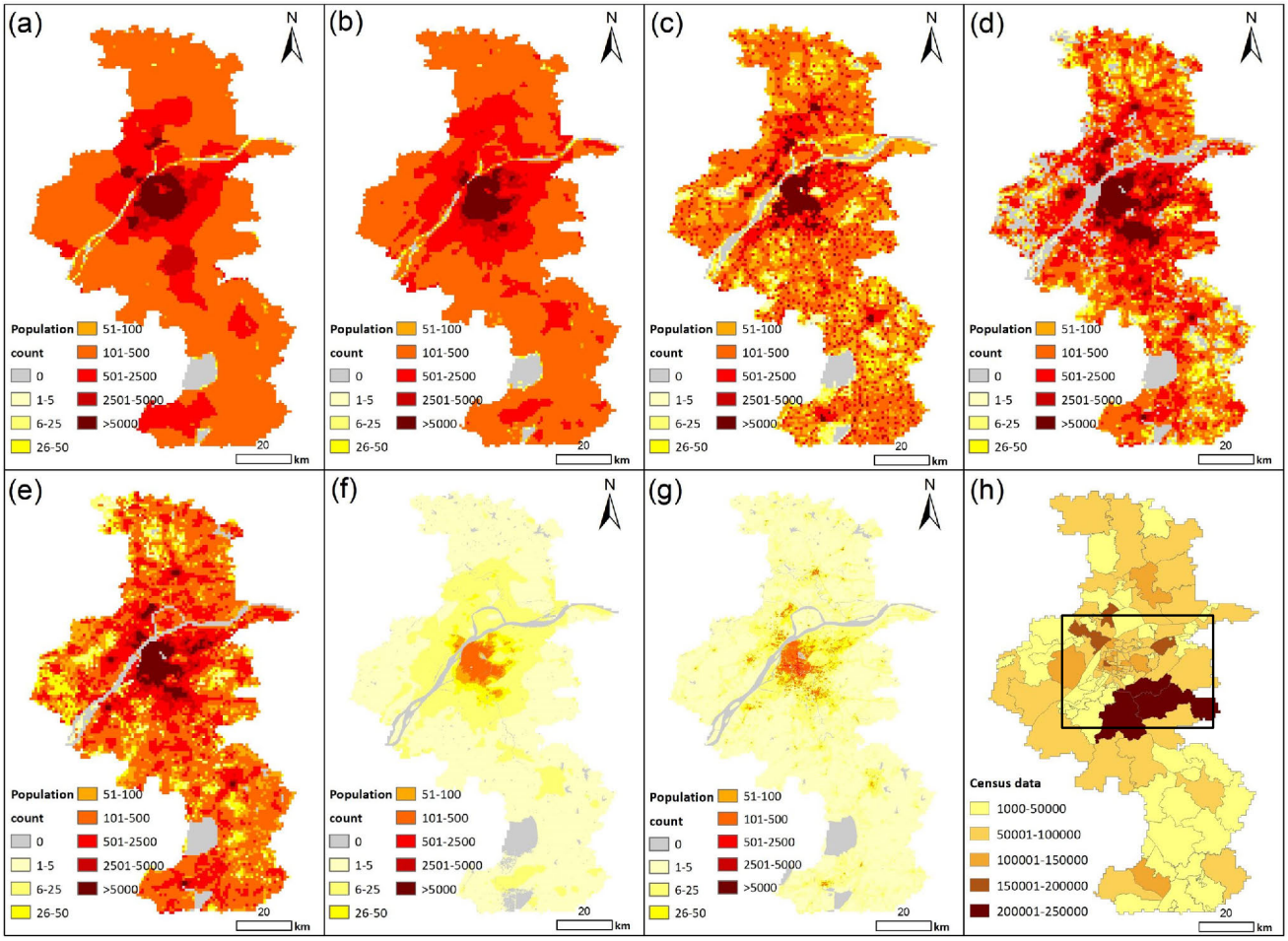


FIGURE 3. Gridded population distribution maps of Nanjing in 2015. (a) GPW version 4 (30 arc-seconds), (b) WorldPop (30 arc-seconds), (c) LandScan (30 arc-seconds), (d) GWR (30 arc-seconds), (e) GWATPRK-based method (30 arc-seconds), (f) WorldPop (3 arc-seconds), (g) GWATPRK-based method (3 arc-seconds), and (h) Census data of Nanjing in 2015.

WorldPop, LandScan, GWR and GWATPRK results at the scale of 30 arc-seconds in the subarea, respectively. Figure 4(f) and (g) are the WorldPop and GWATPRK results at the scale of 3 arc-seconds in the subarea, respectively. Figure 4(h) is the high-resolution remote sensing image in the subarea. For the five results (30 arc-seconds), Figure 4(c)-(e) have more spatial distribution details of population counts than Figure 4(a) and (b). Specifically, Figure 4(a) and (b) have smoother spatial distribution of population counts than other three results and the spatial pattern shape of high population counts (>5000) in GPW looks like a circle while the shapes in other four results are close to the real shape of the administrative boundary in the downtown area. In Qixia district center marked by the upper blue circle and Jiangning district center marked by the lower blue circle in Figure 4, GPW did not predict the high population counts for the two centers and both WorldPop and LandScan only predicted the high population counts in the center of Jiangning district, whereas GWR and GWATPRK predicted the high population counts in the two centers. When examining GWR and GWATPRK results in Figure 4(d) and (e), GWR produced a few under-estimated results (e.g., some gray pixels) and some over-estimated results (e.g., the lower blue circle area). The building height information of the lower blue circle area in Figure 2(g) shows that the average height of this circle area is lower than that of the downtown area. That is to say the average population count of this circle area in GWR results should be lower than that of the downtown area. While the circle area presents a large area high population counts, which is obviously over-estimated. GWATPRK result in Figure 4(e) has less over-estimated and under-estimated results than GWR result because the ATPK result of GWR residual component was combined with gridded GWR estimations to decrease these over- and under-estimated results. The result in Figure 4(g) shows significantly more population distribution details than those of Figure 4(a)-(f). Especially, Figure 4(g) presents more relatively high population counts in the upper blue circle area than Figure 4(f). The upper blue circle contains the business center (e.g., office buildings, Outlets and other shopping malls) and the Xianlin higher education mega center (more than fifteen university campuses). Therefore, this area has real high population counts, which

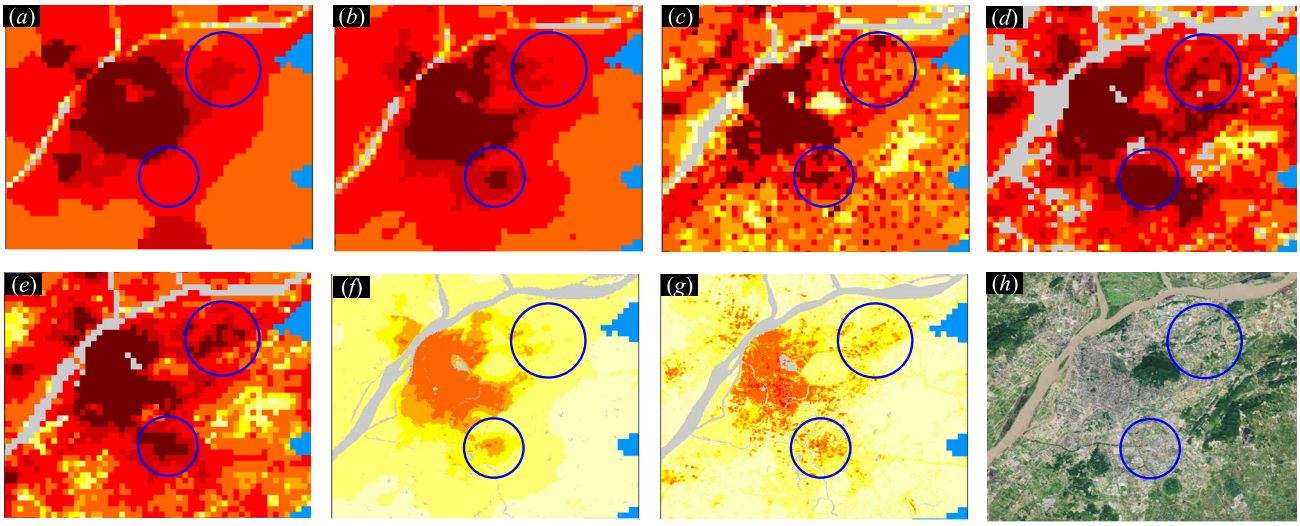


FIGURE 4. Gridded population distribution maps within a subarea of Nanjing in 2015. (a) GPW version 4 (30 arc-seconds), (b) WorldPop (30 arc-seconds), (c) LandScan (30 arc-seconds), (d) GWR (30 arc-seconds), (e) GWATPRK-based method (30 arc-seconds), (f) WorldPop (3 arc-seconds), (g) GWATPRK-based method (3 arc-seconds), and (h) High-resolution remote sensing image.

can also be proved by the high building density from the high-resolution remote sensing image in Figure 4(h). The lower blue circle has the business center (e.g., Baijiahu shopping center and Jiulonghu shopping center) and Jiangning higher education mega center (more than sixteen university campuses); hence it has real high population counts (see another evidence from the high-resolution remote sensing image in Figure 4(h)). Therefore, GWATPRK generated more spatial details and more accurate spatial distribution of population counts than other methods. One of the most important reasons for the apparent improvements of GWATPRK is that GWR generated local regression coefficients to capture the local behaviors by considering spatial heterogeneity while ATPK used the spatial dependence to redistribute the residual component of GWR instead of deleting it like WorldPop to decrease the over- and under-estimated grids in GWR.

B. IMPACT OF ANCILLARY DATA

The ancillary data play a critical role in predicting spatial details for downscaling census data to gridded population distribution maps. The GPW mainly used the census data and a few ancillary data, resulted in the worst performance among these methods. Compared with GPW, LandScan applied more ancillary data derived from remote sensing images and had better performance than GPW in both visual and quantitative. Although Worldpop employed random forest regression and more multisource geospatial data to produce higher accuracy than GPW and LandScan, its spatial distribution details of population are less than LandScan. Compared with GPW, LandScan, and WorldPop, the another important reason for the significant improvements of GWATPRK is largely due to the fact that GWATPRK made full use of both multisource geospatial and social sensing covariates, especially the covariate of positioning density of Tencent users. Compared with WorldPop used substantial geospatial data, GWATPRK employed more detailed and new ancillary data in improving the accurate spatial detail prediction of population. POIs used in GWATPRK were obtained from one of the biggest commercial online maps in China (https://map.qq.com/), which provides more POIs than OpenStreetMap whose POIs were used in WorldPop. Meanwhile, Luojia 1-01 nighttime light image used in GWATPRK is finer than NPP-VIIRS nighttime image used in WorldPop. Last, GWATPRK employed the positioning density of Tencent users, a new type of social sensing data, which provides useful information on the detailed presence of people. Comparing the positioning density of Tencent users in Figure 2(h) with the GWATPRK result in Figure 3(e), it can be found that their spatial pattern is very close to each other. It suggests that the positioning density of Tencent users play a crucial role in predicting the population distribution details, Hence, social sensing data (e.g., the positioning density of Tencent users) have large potential to predict gridded population distribution maps.

Although GWATPRK with both geospatial data and social sensing data generated more accurate gridded population distribution maps than existing gridded population products, the availability of fine ancillary data is a critical factor. For example, the positioning density of Tencent users mainly indicates the information on the detailed presence of Chinese and other areas (over China) may be not applied. But, other alternative social sensing data (e.g., geo-located tweets [31]) can replace it. In future, the proposed method will be employed to generate gridded population count distribution maps over some large areas in China.

V. CONCLUSION

This paper presents a new population downscaling approach. It aimed to take advantage of geographically

weighted area-to-point regression kriging technique to consider the spatial dependence and spatial heterogeneity in downscaling census data to gridded population distribution datasets. Meanwhile, it used both geospatial data and social sensing data to improve its performance. Increased accuracy and more spatial distribution details of population were generated. Therefore, the new population downscaling approach is an effective option for predicting the gridded population distribution data at regional to global scales from census data, geospatial data, and social sensing data.

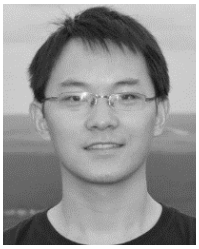
ACKNOWLEDGMENT

The authors would like to thank the GPW, LandScan and WorldPop research groups for providing gridded population distribution products.

REFERENCES

[1] F. R. Stevens, A. E. Gaughan, C. Linard, and A. J. Tatem, "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," *PLoS ONE*, vol. 10, Feb. 2015, Art. no. e0107042.
[2] D. L. Balk, U. Deichmann, G. Yetman, F. Pozzi, S. L. Hay, and A. Nelson, "Determining global population distribution: Methods, applications and data," *Adv. Parasitol.*, vol. 62, pp. 119–156, Jan. 2006.
[3] N. A. Wardrop, W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. J. Tatem, "Spatially disaggregated population estimates in the absence of national population and housing census data," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 14, pp. 3529–3537, 2018.
[4] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proc. Nat. Acad. Sci. USA*, vol. 111, pp. 15888–15893, Nov. 2014.
[5] A. J. Tatem, "Mapping the denominator: Spatial demography in the measurement of progress," *Int. Health*, vol. 6, no. 3, pp. 153–155, 2014.
[6] M. Chen, L. I. Yang, Y. Gong, L. U. Dadao, and H. Zhang, "The population distribution and trend of urbanization pattern on two sides of Hu Huanyong population line: A tentative response to Premier Li Keqiang," *Acta Geograph. Sinica*, vol. 71, no. 2, pp. 179–193, Feb. 2016.
[7] Z. D. Fan, T. Pei, T. Ma, Y. Y. Du, C. Song, and Z. Liu, "Estimation of urban crowd flux based on mobile phone location data: A case study of Beijing, China," *Comput. Environ. Urban*, vol. 69, pp. 114–123, May 2018.
[8] G. R. Watmough, C. L. J. Marcinko, C. Sullivan, K. Tschirhart, P. K. Mutuo, C. A. Palm, and J.-C. Svenning, "Socioecologically informed use of remote sensing data to predict rural household poverty," *Proc. Nat. Acad. Sci. USA*, vol. 116, pp. 1213–1218, Jan. 2019.
[9] S. I. Hay, A. M. Noor, A. Nelson, and A. J. Tatem, "The accuracy of human population maps for public health application," *Tropical Med. Int. Health*, vol. 10, no. 10, pp. 1073–1086, 2010.
[10] C. Linard, V. A. Alegana, A. M. Noor, R. W. Snow, and A. J. Tatem, "A high resolution spatial population database of Somalia for disease risk mapping," *Int. J. Health Geograph.*, vol. 9, p. 45, Sep. 2010.
[11] A. J. Tatem, S. Adamo, N. Bharti, C. R. Burgert, M. Castro, and A. Dorelien, "Mapping populations at risk: Improving spatial demographic data for infectious disease modeling and metric derivation," *Population Health Metrics*, vol. 10, no. 1, p. 8, 2012.
[12] J. Fang, S. Sun, P. Shi, and J. Wang, "Assessment and mapping of potential storm surge impacts on global population and economy," *Int. J. Disaster Risk Sci.*, vol. 5, no. 4, pp. 323–331, 2014.
[13] F. Nadim, O. Kjekstad, P. Peduzzi, C. Herold, and C. Jaedicke, "Global landslide and avalanche hotspots," *Landslides*, vol. 3, no. 2, pp. 159–173, May 2006.
[14] G. Mcgranahan, D. Balk, and B. Anderson, "The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones," *Environ. Urbanization*, vol. 19, pp. 17–37, Apr. 2007.
[15] K. C. Samir and W. Lutz, "The human core of the shared socioeconomic pathways: Population scenarios by age, sex and level of education for all countries to 2100," *Global Environ. Change*, vol. 42, pp. 181–192, Jan. 2017.
[16] D. Martin, "Directions in population GIS," *Geogr. Compass*, vol. 5, no. 9, pp. 655–665, 2011.
[17] S.-S. Wu, X. Qiu, and L. Wang, "Population estimation methods in GIS and remote sensing: A review," *Mapping Sci. Remote Sens.*, vol. 42, no. 1, pp. 80–96, 2005.
[18] Z. Bai, J. Wang, M. Wang, M. Gao, and J. Sun, "Accuracy assessment of multi-source gridded population distribution datasets in China," *Sustainability*, vol. 10, p. 1363, May 2018.
[19] C. T. Lloyd, A. Sorichetta, and A. J. Tatem, "High resolution global gridded data for use in population studies," *Sci. Data*, vol. 4, Jan. 2017, Art. no. 170001.
[20] E. Doxsey-Whitfield, K. MacManus, S. B. Adamo, L. Pistolesi, J. Squires, O. Borkovska, and S. R. Baptista, "Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4," *Papers Appl. Geograph.*, vol. 1, no. 3, pp. 226–234, 2015.
[21] F. J. Reed, A. E. Gaughan, F. R. Stevens, G. Yetman, A. Sorichetta, and A. J. Tatem, "Gridded population maps informed by different built settlement products," *Data*, vol. 3, p. 33, Sep. 2018.
[22] B. Bhaduri, E. Bright, P. Coleman, and M. L. Urban, "LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics," *GeoJournal*, vol. 69, pp. 103–117, Jun. 2007.
[23] J. E. Dobson, E. A. Bright, P. R. Coleman, R. C. Durfee, and B. A. Worley, "LandScan: A global population database for estimating populations at risk," *Photogramm. Eng. Remote Sens.*, vol. 66, pp. 849–858, Jul. 2000.
[24] P. Jia, Y. Qiu, and A. E. Gaughan, "A fine-scale spatial population distribution on the high-resolution gridded population surface and application in Alachua county, Florida," *Appl. Geograph.*, vol. 50, pp. 99–107, Jun. 2014.
[25] D. Azar, J. Graesser, R. Engstrom, J. Comenetz, R. M. Leddy, Jr., N. G. Schechtman, and T. Andrews, "Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti," *Int. J. Remote Sens.*, vol. 31, pp. 5635–5655, Nov. 2010.
[26] D. J. Briggs, J. Gulliver, D. Fecht, and D. M. Vienneau, "Dasymetric modelling of small-area population distribution using land cover and light emissions data," *Remote Sens. Environ.*, vol. 108, pp. 451–466, Jun. 2007.
[27] M. Bakillah, S. Liang, A. Mobasheri, J. J. Arsanjani, and A. Zipf, "Fine-resolution population mapping using OpenStreetMap points-of-interest," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 9, pp. 1940–1963, 2014.
[28] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi, "Social sensing: A new approach to understanding our socioeconomic environments," *Ann. Assoc. Amer. Geograph.*, vol. 105, no. 3, pp. 512–530, 2015.
[29] Z. Liu, T. Ma, Y. Du, T. Pei, J. Yi, and H. Peng, "Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records," *Trans. GIS*, vol. 22, no. 2, pp. 494–513, 2018.
[30] B. Yu, T. Lian, Y. Huang, S. Yao, X. Ye, Z. Chen, C. Yang, and J. Wu, "Integration of nighttime light remote sensing images and taxi GPS tracking data for population surface enhancement," *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 4, pp. 687–706, 2019.
[31] N. N. Patel, F. R. Stevens, Z. Huang, A. E. Gaughan, I. Elyazar, and A. J. Tatem, "Improving large area population mapping using geotweet densities," *Trans. GIS*, vol. 21, no. 2, pp. 317–331, 2017.
[32] Y. Yao, X. Liu, X. Li, J. Zhang, Z. Liang, K. Mai, and Y. Zhang, "Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data," *Int. J. Geograph. Inf. Sci.*, vol. 31, pp. 1220–1244, Jun. 2017.
[33] W. U. Xun, J. Yang, and H. Zhang, "Analyzing spatial autocorrelation of population distribution in different spatial weights: A case of China," *Geomatics World*, vol. 24, no. 2, pp. 32–38, 2017.
[34] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Econ. Geogr.*, vol. 46, pp. 234–240, Jun. 1970.
[35] Y. Ge, Y. Jin, A. Stein, Y. Chen, J. Wang, J. Wang, Q. Cheng, H. Bai, M. Liu, and P. M. Atkinson, "Principles and methods of scaling geospatial earth science data," *Earth-Sci. Rev.*, vol. 197, Jul. 2019, Art. no. 102897.
[36] T. Hengl, G. B. Heuvelink, and M. P. Tadić, and E. J. Pebesma, "Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images," *Theor. Appl. Climatol.*, vol. 107, nos. 1–2, pp. 265–277, Jan. 2012.
[37] P. Goovaerts, *Geostatistics for Natural Resources Evaluation*. Oxford, U.K.: Oxford Univ. Press, 1997.
[38] D. Murakami and M. Tsutsumi, "Area-to-point parameter estimation with geographically weighted regression," *J. Geograph. Syst.*, vol. 17, pp. 207–225, Jul. 2015.

[39] P. C. Kyriakidis, "A geostatistical framework for area-to-point spatial interpolation," *Geograph. Anal.*, vol. 36, no. 3, pp. 259–289, 2004.
[40] Y. Chen, Y. Ge, Y. Chen, Y. Jin, and R. An, "Subpixel land cover mapping using multiscale spatial dependence," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5097–5106, Mar. 2018.
[41] Q. Wang, W. Shi, P. M. Atkinson, and Y. Zhao, "Downscaling MODIS images with area-to-point regression Kriging," *Remote Sens. Environ.*, vol. 166, pp. 191–204, Sep. 2015.
[42] M. F. Goodchild, "The validity and usefulness of laws in geographic information science and geography," *Ann. Assoc. Amer. Geogr.*, vol. 94, no. 2, pp. 300–303, 2004.
[43] A. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ, USA: Wiley, 2002.
[44] P. Harris, A. S. Fotheringham, R. Crespo, and M. Charlton, "The use of geographically weighted regression for spatial prediction: An evaluation of models using simulated data sets," *Math. Geosci.*, vol. 42, pp. 657–680, Aug. 2010.
[45] Y. Jin, Y. Ge, J. Wang, Y. Chen, G. B. Heuvelink, and P. M. Atkinson, "Downscaling amsr-2 soil moisture data with geographically weighted area-to-area regression Kriging," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2362–2376, Apr. 2018.
[46] Y. Yao, J. Zhang, Y. Hong, H. Liang, and J. He, "Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data," *Trans. GIS*, vol. 22, no. 2, pp. 561–581, 2018.
[47] Y. Jin, Y. Ge, J. Wang, G. Heuvelink, and L. Wang, "Geographically weighted area-to-point regression Kriging for spatial downscaling in remote sensing," *Remote Sens.*, vol. 10, no. 4, p. 579, 2018.
[48] Y. Chen, Y. Ge, G. B. M. Heuvelink, R. An, and Y. Chen, "Object-based superresolution land-cover mapping from remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 328–340, Jan. 2018.
[49] P. Goovaerts, "Kriging and semivariogram deconvolution in the presence of irregular geographical units," *Math. Geosci.*, vol. 40, no. 1, pp. 101–128, 2008.
[50] Y. Chen, Y. Ge, R. An, and Y. Chen, "Super-resolution mapping of impervious surfaces from remotely sensed imagery with points-of-interest," *Remote Sens.*, vol. 10, p. 242, Feb. 2018.
[51] W. Jiang, G. He, T. Long, H. Guo, R. Yin, W. Leng, H. Liu, and G. Wang, "Potentiality of using Luojia 1-01 nighttime light imagery to investigate artificial light pollution," *Sensors*, vol. 18, no. 9, p. 2900, 2018.
[52] Y. Wang, L. Dong, Y. Liu, Z. Huang, and Y. Liu, "Migration patterns in China extracted from mobile positioning data," *Habitat Int.*, vol. 86, pp. 71–80, Apr. 2019.



YUEHONG CHEN received the B.S. degree from Hohai University, Nanjing, China, in 2010, and the M.Sc. and Ph.D. degrees from the State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, University of Chinese Academy of Sciences, China, in 2013 and 2016, respectively. He is currently an Associate Professor with the School of Earth Sciences and Engineering, Hohai University. His current research interests include geospatial data analysis, geostatistics, and scale transformation.

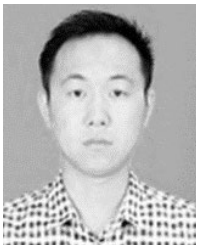
RUOJING ZHANG received the B.S. degree from Anhui Normal University, Wuhu, China, in 2019. She is currently pursuing the master's degree in cartography and geographical information system with the School of Earth Sciences and Engineering, Hohai University, Nanjing, China. Her current research interest includes geospatial data analysis.

YONG GE (M'14) received the Ph.D. degree in cartography and geographical information system from the Chinese Academy of Sciences (CAS), Beijing, China, in 2001.

She is currently a Professor with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS. She has directed research in more than ten national projects. She is the author or coauthor of over 80 scientific articles published in refereed journals, one book, and six chapters in books. She is the editor of one book, and she holds three granted patents in improving the accuracy of information extraction from remotely sensed imagery. Her research interests include spatial data analysis and data quality assessment.

Dr. Ge is a member of the Theory and Methodology Committee of the Cartography and Geographic Information Society, the International Association of Mathematical Geosciences, and the Editorial Board of *Spatial Statistics* (Elsevier). She has been involved in the organization of several international conferences and workshops.

YAN JIN received the B.S. degree in information and computation science and the M.S. degree in applied mathematics from Chang'an University, Xi'an, China, in 2011 and 2014, respectively, and the Ph.D. degree in cartography and geographical information system with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, in 2018. She is currently a Lecturer with the School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing China. Her research interests include scale transformation, data fusion, geostatistics, and remote sensing applications.



ZELONG XIA was born in Taizhou, Jiangsu, China, in 1988. He received the M.S. degree in surveying and mapping engineering from the Chengdu University of Technology, Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Earth Sciences and Engineering, Hohai University, Nanjing, China.

His research interests include urban geography research and spatiotemporal data mining.