Disaggregating County-Level Census Data for Population Mapping Using Residential Geo-Objects With Multisource Geo-Spatial Data

Tianjun Wu, Jiancheng Luo, Wen Dong, Lijing Gao, Xiaodong Hu, Zhifeng Wu, Yingwei Sun, and Jinsong Liu

***Abstract*—Accurate spatialization of socioeconomic data is conducive to understand the spatial and temporal distribution of human social development status and, thus, effectively support future scientific decision-making. This study focuses on population mapping, which is a classical spatialization of macroeconomic data of the social economy. Traditional population mapping based on rough grids or administrative divisions such as townships often has deficiencies in the accuracy of spatial pattern and prediction. In this article, hence, we employ residential geo-objects as basic mapping units and formalize the problem as a spatial prediction process using machine-learning (ML) methods with high-spatial-resolution (HSR) satellite remote sensing images and multisource geospatial data. The indicators of population spatial density, including residential geo-objects' area, building existence index, terrain slope, night light intensity, density of point of interest (POI) and road network from Internet electronic maps, and locational factors such as the distances from road and river, are jointly applied to establish the relationship between these multivariable factors and quantitative index of population density using ML algorithms such as Random Forests and XGBoost. The predicated values of population density from the mined nonlinear regression relation are further used to calculate the weights of disaggregation of each unit, and then the population quantity distribution at the scale of residential geo-objects is obtained under the control of the total amount of population statistics. Experiments with a county area show that the methodology has the ability to achieve better results than the traditional deterministic methods by reproducing a more accurate and finer geographic population distribution pattern. Meanwhile, it is found that the optimization of mapping results may benefit from the multisources geospatial data, and thus the methodological framework can be recommended to be extended to other spatialization areas of socioeconomic data.**

***Index Terms*—Census data, machine-learning (ML) algorithms, multisource geospatial data, population mapping, residential geo-objects, spatialization.**

Manuscript received December 3, 2019; revised January 20, 2020; accepted February 15, 2020. Date of publication March 19, 2020; date of current version April 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41631179, Grant 41601437, and Grant 41671138, in part by the Fundamental Research Funds for the Central Universities, CHD under Grant 300102120201 and Grant 300102269205, in part by the National Key Research and Development Program under Grant 2017YFB0503600, in part by the Ningxia Academy of Agricultural and Forestry Sciences Foreign Science and Technology Cooperation Project under Grant 07030002, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ1038, and in part by the Open Projects of Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University under Grant 2018LSDMIS03. *(Corresponding author: Jiancheng Luo.)*

Tianjun Wu is with the School of Geology Engineering and Geomatics, Chang'an University, Xi'an 710064, China, and also with the Key Laboratory of Spatial Data Mining and Information Sharing of the Ministry of Education, Fuzhou University, Fuzhou 350000, China (e-mail: wutianjun1986@163. com).

Jiancheng Luo, Wen Dong, Lijing Gao, Xiaodong Hu, and Yingwei Sun are with the State Key Laboratory of Remote Sensing Science of the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100864, China, and also with the University of Chinese Academy of Sciences, Beijing 100864, China (e-mail: luojc@radi.ac.cn; dongwen01@radi.ac.cn; gaolj@radi. ac.cn; huxd@radi.ac.cn; wfsunyingwei@yeah.net).

Zhifeng Wu is with the School of Geographical Sciences, Guangzhou University in Guangzhou, Guangdong 510000, China (e-mail: gzuwzf@163.com).

Jinsong Liu is with the School of Resources and Environmental Sciences, Hebei Normal University, Hebei 051530, China (e-mail: ljslyx@sohu.com).

Digital Object Identifier 10.1109/JSTARS.2020.2974896

I. Introduction

POPULATION is one of the most important social issues affecting the sustainable development of the world today. Population growth in recent years has increased the bearing pressure of global resources and environment, and, thus, the contradiction between environment, resources, and population has become more prominent [1]. The timely grasp of population information and a clear understanding of population spatial distribution and its changes can provide scientific support for regional sustainable development research and planning [2].

Population census is the main channel of realizing the statistics and analysis of population information in various countries, including sampling survey and general census [3]. Although the census and statistics are supported by rigorous statistical theories and methods and have the advantages of authority and standardization, there are still some defects such as data scarce, low accuracy and time resolution, long update cycle, which are not conducive to visualization and spatial analysis [4]. It is difficult to meet the requirements of the accurate application in population distribution law research and natural disaster response. Therefore, in the current context of rapid urbanization and population migration, the information of population spatial distribution with fine resolution, high accuracy, and reliability would play an important role in studying and explaining the impact of human beings on society, economy, and environment [5]. It is also important to study population problems, environmental problems, and man-land relationships.

After nearly 30 years of development, the research of population data spatialization has gradually matured. Many methods are developed and many achievements have been achieved. The main methods of spatialization of population data are as follows. First, dasymetric mapping divides the spatial distribution of population into small areas that can reflect the spatial

heterogeneity of population based on auxiliary information and to generate fine-scale population distribution data using surface interpolation technology [6]–[8]. Zoning density model has become a widely used population spatialization technology, and its application is mostly found in foreign research [9]–[12]. Second, multifactor integration fuses of different schemes for each single factor and obtains population distribution weights of each unit for disaggregating the total population census of administrative unit. The most typical application of this idea is LandScan, a global population distribution database developed by Oak Ridge National Laboratory in the United States [13]–[15]. Third, multivariate regression constructs the relationship between multiple factors and population statistics by using spatial data containing population distribution information. Linear regression (LR) model with medium resolution data was usually used under the basic hypothesis that the population density of the same land use type is the same in a certain area [16]–[18].

In general, previous studies mostly follow the top-down approach via disaggregating statistical data in administrative units in grid scales according to certain principles. Methodologically, latest studies have shown that there is a trend from the top-down population spatialization method under the influence of natural and economic factors to the bottom-up intelligent multiple regression method [3].

In addition, as for the basic mapping units, grid scales with the different resolution are commonly used in current representative global and national scale population spatial databases such as Gridded Population of the World (WorldPop) [19], [20], Global Rural-Urban Mapping Project (GRUMP) [21], LandScan [13], [14], Global Resource Information Database (UNEP/GRID) [22], China's 1-km grid population database [23]. Although this gridding operation is simple and closer to the actual distribution of population than that of using an administrative unit. However, it is difficult to evaluate the uncertainty of population distribution at the grid scale, especially a mapping is conducted at the scale of 1 km or 500 m. Moreover, the information about population distribution tends to perform poorly in refined scenarios when the degree of refinement is not enough. The main problems are 1) the grid units in the population mapping are not consistent with the natural units (soil type units, vegetation type units, land-use units, etc.) in the actual research, which makes it difficult to verify the accuracy of spatial population information on grid scale, and is not conducive to the research of mining the distribution law of population and its simulation and migration prediction; 2) the population information represented by grids cannot well reflect the different characteristics of population spatial distribution on a small scale, and, thus, its accuracy always cannot meet the requirements of many scientific research works and engineering applications. Therefore, fine-scale population mapping corresponding to geographical entities has realistic research needs [24]. High spatial-temporal resolution products of population spatialization will be more effective in the applications of disaster assessment, resource allocation, and smart city construction.

The residential geo-objects are settlement units where people live together on the surface of the Earth, which can be clearly identified by buildings on remote sensing images with high-spatial-resolution (HSR). The residential geo-objects can reflect not only the distribution of original settlements but also the continuity and aggregation of population residence. High precision estimation of the population may be achieved within buildings polygons [25]. However, at present, the spatialization of population data on this scale is less studied, and it is worth further study as the accuracy of the mapping needs to be improved.

Therefore, this article employs residential geo-objects as the basic unit of population mapping. A multifactor consideration based on multisource geospatial data is conducted with ML methods of non-LR tree algorithms. The proposed method is shown to be effective for disaggregating county-level census data in population mapping. The residential geo-objects are proved to be an appropriate scale of fine mapping and avoid the suitability of the grid scale. Accurate, realistic, and geographically consistent results of population spatial distribution can be obtained by carrying out this spatialization of population census data.

The rest of this article is organized as follows. The proposed method is presented in Section II. The experiments and their analysis are conducted to evaluate the effectiveness of our proposed method are described in Section III. The conclusions of the article are provided in Section IV.

II. Methodology

As a relative fine scale, the spatialization of population data in residential units has attracted more and more attention. To effectively mine the patterns of population distribution in geospace, we aim to solve the mapping problems based on the paradigm of microcosmic residential geo-objects. They are population information granules for mapping because they have similar conditions of natural resources and social economy under a geographical background. In this article, hence, we employ residential geo-objects as a basic mapping unit and formalize the problem as a spatial prediction process using machine-learning (ML) methods with HSR satellite remote sensing images and multisource geospatial data. Note that, we take midnight as the time node and focus on the study of the static spatial distribution of the resident population on residential land. Effectively revealing the difference of population distribution within residential land is the purpose of this article.

Fig. 1 illustrates the whole process with the operations of downscaling, raster aggregation, spatial overlay, model training, prediction, and result exportation. It contains the following four steps. First, the residential geo-objects are extracted from HSR remote sensing images. Second, a structured multidimensional attribute table is constructed by integrating multisource geospatial data into each residential geo-object. Third, we extract a nonlinear relationship between multidimensional attributes from multiple geospatial data (i.e., the predictor/explanatory variables) and census measured population density (i.e., target/response variable) based on residential geo-object units using tree-based ML methods. This relationship (i.e., the training model) is further employed to generate spatial predictions of population density. Then we can obtain the final population quantity mapping on all residential geo-objects by calculating

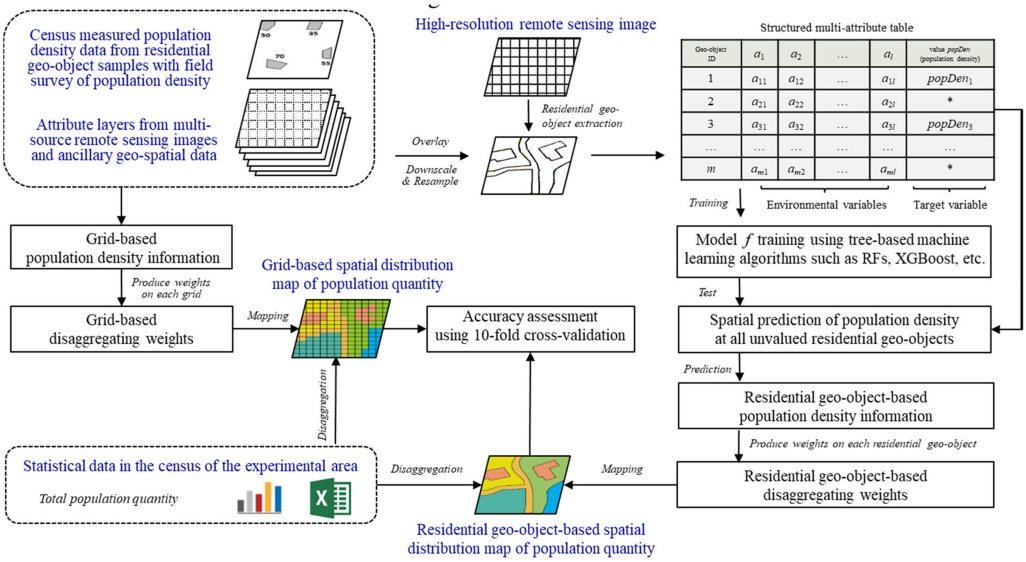

Fig. 1. Flowchart of the proposed procedure for population mapping based on grids and residential geo-objects using multisource geospatial data.

weights of decomposition according to their predicated population densities. Finally, an accuracy assessment is conducted to evaluate the residential geo-object-based result. It is also compared with that of a grid-based process. The following subsections describe the implementation of our proposed procedure.

A. Extraction of Residential Geo-objects

There is a coupling relationship between population distribution pattern and land use/land cover pattern [1], and the rise of global change research has increased the accessibility of land use/land cover data, making it the most commonly used element in population data spatialization. Residential land use/land cover areas are the major areas of population distribution with buildings for people to live, including ordinary houses, apartments, villas, dormitories house-site in the countryside, etc. Hence, in this article, residential areas are used as the modeling control unit, aiming at reducing the influence of spatial scale transformation of the model and ensuring the accuracy of spatialization of fine-scale population data.

In recent years, the application of HSR remote sensing images in the study of population spatial distribution has improved the spatial resolution of a population distribution to a certain extent, and related research focuses on the fine division of spatial recognition units of population distribution [26]. In our study, HSR remote sensing images are selected to carry out a visual interpretation of land use. Meaningful geo-object-based polygons with established land-use types are produced via the combination of human visual experience and manual interpretation. That is, these geo-object-based polygons are controlled by the image edges and in line with the visual format of image-objects reflecting the exact geographic entries.

Among these, first, geo-objects with the types of water, cultivated land, grassland, forest land, and bare rock in land use are unsuitable for human settlements and are removed as masks in spatialization. Second, building areas are given the highest priority or absolute weight in spatialization as the carrier of human settlements. The building geo-objects are obtained establishing the interpretation signs of residential building patches according to the color, shape, related layout, and geographical distribution of residential buildings reflected in remote sensing images. From the perspective of land use and land cover, the types of these geo-objects belong to impervious surface or construction land. Corresponding geo-objects are extracted with contour polygons of building areas from remote sensing images using vectorization method. On the basis of these image-generated geo-objects, some geo-objects of public building, industrial building, and agricultural building are further removed with the assistance of points of interest (POI), building three-dimensional landscape map, and street view map from electronic maps. This method can minimize the impact of nonresidential construction land on population spatial distribution.

The filtered out residential geo-objects expressed using geometric polygons are referred as the core area of population distribution. We assume that the population density of each residential geo-object is identical for this study and the population densities of various geo-objects will be calculated according to multiscale features extracted in the following stages by overlaying multisource data on these basic interpreted units.

B. Construction of a Structured Multiattribute Table

After extracting the target residential geo-objects, we collect multiple corresponding geospatial data to assign to each geo-objects. We overlay multisource data on these basic mapping units RO_i $(i = 1, 2, \ldots, m)$ to construct a structured multiattribute table for ML-based regression analysis. The following steps are the operations in the process of generating a structured multiattribute table.

On the one hand, the variable *popDen* of population density is the target/response variable of concern in the model. The

value of each residential geo-object is unknown and should be predicated according to a reasonable method. A bottom-up sample-based learning scheme is employed as it is shown to be effective in inverting population information [3]. We annotate a small amount of residential geo-objects using field survey values of population density in several sample areas, which are converted by the population number of surveys with the precise boundaries of residential geo-objects. The field survey is more laborious, and we only get the population information of some typical regions. Their surveyed values are transferred to polygons and employed as the values of the target variable (i.e., response variable) $popDen_{\mathrm{train},i}$ ($i = 1, 2, \ldots, m_{\mathrm{train}}$) of in the regression and the missing values of geo-objects $popDen_{\mathrm{test},i}$ ($i = 1, 2, \ldots, m_{\mathrm{test}}$) should be predicted.

On the other hand, related natural and social-economic factors are comprehensively considered as environmental variables. The variables that affect the distribution of population need to be analyzed. We review on previous studies and determine the influencing factors of population distribution [18]. Multisource geospatial data from the natural environment and social economic activity are comprehensively collected and transferred to each residential geo-object RO_i ($i = 1, 2, \ldots, m$) using spatial analysis technologies such as downscaling using area weighted interpolation, bicubic resampling, and overlay analysis based on polygons [27], [28]. The indicators of population density, including residential geo-objects' spectral and texture characteristics, area, building existence index, terrain slope, night light intensity, density of POI and road network from Internet electronic maps, locational factors such as the distances from road and river, and existing grid-based population data are jointly applied as the environmental variables of each geo-object a_j ($j = 1, \ldots, l$) for multidimensional analyses.

Thus, a structured attribute table with m_{train} valued residential geo-objects for training and m_{test} unvalued residential geo-objects for prediction is prepared to mine the relationships between l environmental variables (i.e., $a_1, a_2, \ldots, a_l$) and one target variable (i.e., the value of *popDen*).

C. Spatial Prediction Using Tree-Based Algorithms

Next, we implement the mapping framework by producing spatial predictions of population density. The relationship between these multivariable factors and quantitative index of population density will be established using ML algorithms. Thus, first, we establish a regression model by analyzing the relationships between l environmental variables and the population density

$$popDen = f(a_1, a_2, \ldots, a_l) + \varepsilon \tag{1}$$

where f is the regression function and ε is a spatially autocorrelated residual. The function f in the prediction model can be fitted using linear or non-LR methods, and the residuals are interpolated using ordinary kriging interpolation methods [29], [30].

Once the relationships contained in function f are established according to the m_{train} train samples, we predict the values of population density at all m_{test} unvalued residential geo-objects in the field. The predicated values of population density from the mined regression relationship are further used to calculate the weights of disaggregation of each unit and then the population quantity distribution at the scale of residential geo-objects can be obtained under the control of the total amount of population statistics. Thus, in this framework, it is important to accurately fit the regression function f as a spatial prediction model. Hence, we employ advisable non-LR methods using ML algorithms in this step [31]. Besides the support vector regression (SVR) [32], [33], Random Forests (RFs) [34], [35], and its gradient boosted algorithms, i.e., XGBoost [36], are commonly used tree-based methods. They have been proved effective in learning a model for prediction within a high dimensional variable set [37]–[41]. The principles of these algorithms are not detailed in this article. Detailed descriptions and relevant explanations of these algorithms can be found in the literature [34]–[36].

The tree-based regression algorithms selected for fitting in our framework have been proved to be robust presenting high prediction accuracies when modeling nonlinear relationship among a large number of variables with discrete and continuous values [37], [39]. For example, in the field of population spatialization, Azar *et al*. [5] successfully conduct the spatialization of Pakistan's population data in 2010 by using the regression decision tree algorithm. Meanwhile, Stevens *et al*. [17] pointed out that the use of RFs model in population spatialization needs census data well matched with the administrative boundaries, which can be guaranteed in our research. These applications of tree-based algorithms provide a good inspiration for our study. However, they were both achieved on the grid scale. While, in our study, when a spatial prediction model is nonlinearly fitted by these tree-based algorithms at the residential geo-object level, we apply it to all the unvalued residential geo-objects to spatially predict their values for the target variable (i.e., the population density) under the constraints of various environmental variables. Final residential geo-object maps of the variation of population density and population quantity can be globally generated by the learned tree-based predication model.

D. Quantitative Accuracy Assessment

Predicated population density/quantity mapping is then achieved on all the residential geo-objects. To evaluate the prediction performance and compare the differences in grid-based and geo-object-based mapping results, we assessed the accuracies using a 10-fold spatial cross-validation (CV) to deal with the issues of small sample size and spatial autocorrelation. Spatial partitioning is carried out to reduce the effect of spatial autocorrelation. The number of folds chosen in spatial partitioning equal to spatial clusters contained in the dataset and a spatial Euclidean distance between training and test sets is introduced to relieve the effect of spatial autocorrelation [42]–[45].

Based on the 10-fold spatial CV, four quantitative measures, coefficient of determination (R^2), the mean absolute error (MAE), root mean squared error (RMSE), and relative RMSE (%RMSE), are calculated to evaluate the prediction performance

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \tag{2}$$

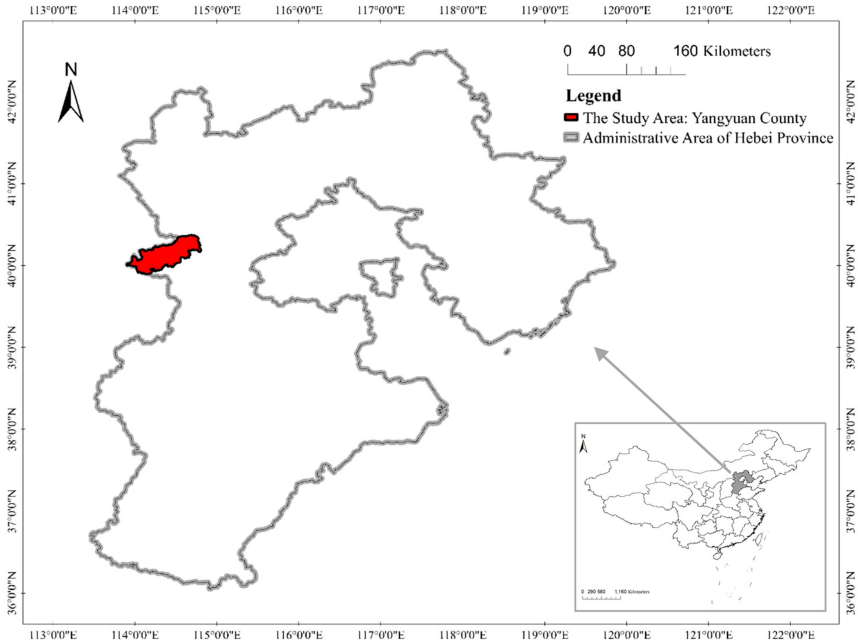

Fig. 2. Geographic location of the study area.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{popDen}_i - popDen_i\right| \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{popDen}_i - popDen_i\right)^2} \tag{4}$$

$$\%\mathrm{RMSE} = \frac{\mathrm{RMSE}}{\frac{1}{n}\sum_{i=1}^{n} popDen_i} \tag{5}$$

where SSE is the sum of the squared errors at CV residential geo-objects, and SST is the sum of the squares of the original observations; n is the number of test observations (i.e., 10% of the size of CV residential geo-objects); $\hat{popDen}_i$ is the estimated value of the ith data, i.e., the estimated population density obtained after population spatialization in this article; $popDen_i$ is the corresponding reference value, i.e., the population density obtained from field survey. The R^2 value indicates the amount of variation explained by the model. An R^2 value close to 1 indicates a perfect model. The MAE is the measure of the model bias and a smaller value corresponds to smaller model prediction bias. The RMSE is a measure of the accuracy of the model, and a lower RMSE indicates more accurate predictions during global mapping. The %RMSE is calculated by dividing the RMSE by the average of the census values, which can also reflect the accuracy of the model [17].

III. Experiments and Analysis

A. Study Area

The experimental study area is selected in Yangyuan County, Hebei Province, China, which is located in North China with the geographical coordinates 39°53'–40°22' in the North latitude and 113°54'–114°48' in the East longitude. It is 82 km long from east to west, 27 km wide from north to south, with a total area of 1849 square km. As shown in Fig. 2, Yangyuan County is surrounded by mountains in the north and south. The Sanggan River runs across the whole territory from west to east. Its general characteristics are high in the southwest and low in the northeast, high in the South and low in the north, and there is a narrow basin with two mountains and one river. The region is a transitional zone between the Loess Plateau, Inner Mongolia Plateau, and North China Plain. Its landform includes mountains, hilly plains in front of mountains and rivers. The average altitude is about 1100 m, the lowest is 770 m, and the highest is 2045.9 m.

There are 301 administrative villages in 5 towns and 9 townships under the jurisdiction of Yangyuan County in 2018. The residential buildings in this area are mainly multistorey buildings and flat buildings, which are distributed in the central part of the county and the countryside, respectively. It is a state-level poverty-stricken county in China. Population is one of the concerns of the local government, which is concentrated mainly in the plains on both sides of the Sanggan River due to the topography and landform. Population distribution reflects the difference between natural geographical conditions and the level of economic development in a region. The significance of studying the population distribution in this area lies in revealing the regional characteristics of population distribution and further grasping the regularity of population spatial distribution. Hence, fine population quantity/density maps are important for policy formulation for local sustainable development and precise poverty alleviation.

According to the latest census data in 2018, the total population of the region is 288 713. This number is used to control constrain the total population quantity in our experiment. The goal of our research is to disaggregate such a census population quantity into microscale areas where the population distribution may exist in the whole county.

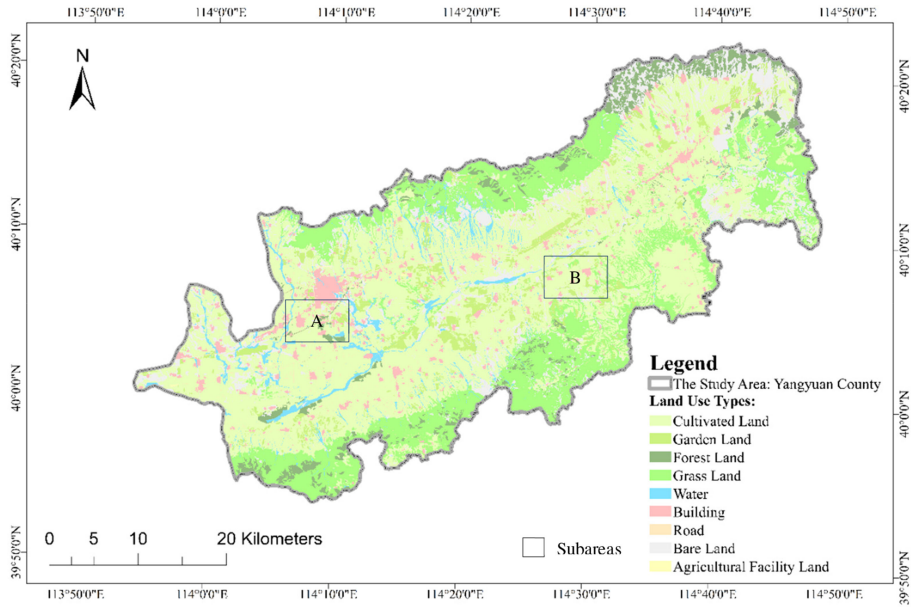

Fig. 3. Land use map and examples of residential geo-objects.

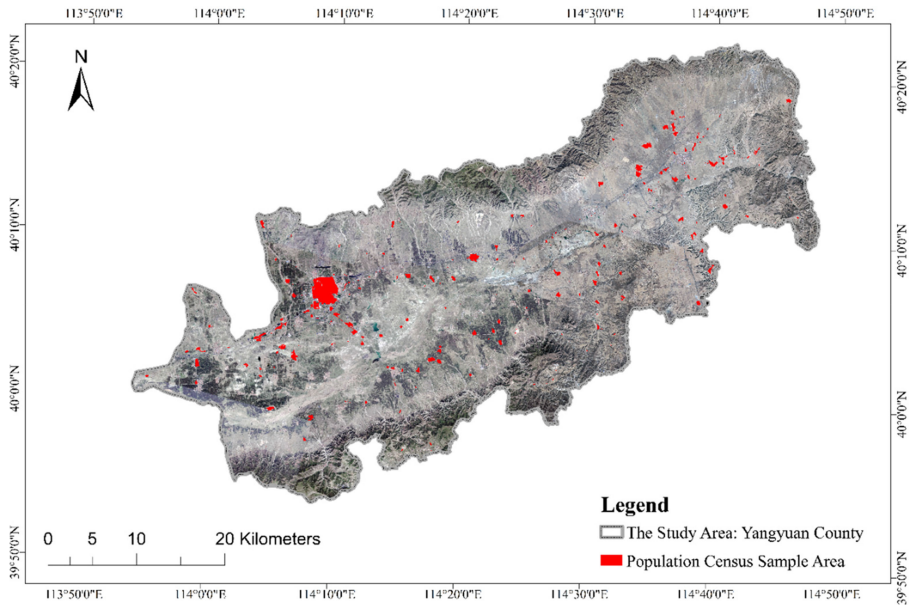

Fig. 4. GF-2 remote sensing image (March 24, 2018, true color band combination) of the study area and the distribution of census sample areas.

B. Data Sets and Preprocessing

Although the mechanism of population spatial distribution has not been thoroughly studied, the conditions affecting its distribution are probably clear. Its related indicators are complex and diverse. Multisource and new-type geospatial data make the modeling factors pluralistic. To construct abundant covariate layers for a comprehensive multiattribute table, we collected multisource auxiliary data as follows.

1) Land Use Data: The classification accuracy and detail of land use data determine the accuracy and scale of spatialization. First, as shown in Fig. 3, a geo-object-based land use map was produced based on the HSR remote sensing images in Fig. 4. To distinguish the influence of different land-use types on population distribution, we masked cultivated land, garden land, grassland, forest land, water, road, and only filtered out the geo-objects with the type of building. On the basis of these image-generated geo-objects, some geo-objects of public building, industrial building, and agricultural building were further removed with the assistance of an electronic map. The remaining geo-objects were considered possible settlement-based areas of population distribution, which were calle residential geo-objects

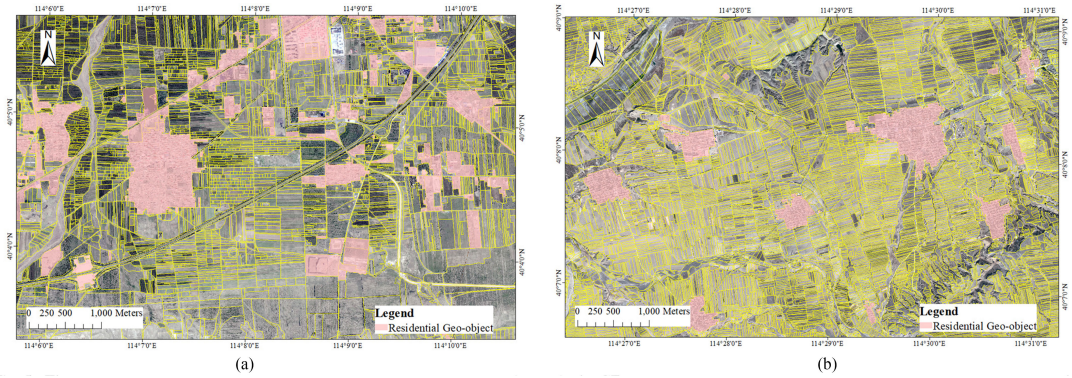

Fig. 5. Two subareas of the extracted residential geo-objects (the bottom image is the GF-2 remote sensing image, the polygons with yellow boundary are the extracted geo-objects, and the colored polygons are residential geo-objects). (a) Residential geo-objects in the subarea A of Fig. 3. (b) Residential geo-objects in the subarea B of Fig. 3.

in this study and referred to as the appropriate entities that can be obtained from HSR remote sensing images by identifying areas where the population is concentrated. Fig. 5 shows two subareas of the extracted residential geo-objects with visually precise boundaries. These geo-objects with clear polygon edges were recognized as the basic spatial mapping units in our methodology. That is, we only disaggregated the total county population census data in these areas with impervious surface cover.

In addition, their distances to road, water, and the county center also can be measured by calculating European distance between the center points of residential geo-objects' polygons and those of the geo-objects with corresponding land-use types. These index derived from this land use map were further integrated into residential geo-objects. They characterize the accessibility of road, water, and the county center. It is reasonable to use them as environmental covariates for population density predication as the traffic corridor is the channel of human activities and has a cohesive effect on population distribution within a certain buffer zone [46].

2) Population Census Sample Data: Since this article adopts the framework of bottom-up approaches [3], we need some samples to support our calculation. Thus, in the first half of 2018, we collected the vector data of administrative divisions and conducted a household-based survey by questionnaire and collected information on the number of permanent residents, population structure, and mobility. We completed the sampling in two stages. Before the survey, 10 sample villages were randomly selected from 288 rural communities according to the judgment principle of sampling effectiveness: each endowment area should be sampled to sample villages; the proportion of sample quantity in each district should be equal to the percentage of quantity and area of administrative villages; and the comprehensive endowment characteristics of sample villages can represent the average endowment characteristics of the area. After determining the sample villages, we selected several peasant households in our extracted residential geo-objects for investigation. The questionnaire of household mainly includes basic information of family (family type, population, age, education level, employment status, etc.); land status (conversion of farmland, abandoned land, farmland transfer, basic information of family operating land); lifestyle in food, clothing, housing, transportation, entertainment; and livelihood (income, expenditure) in 2017. The population information of each household was reprocessed to avoid confidentiality issues. They were organized and spatialized into corresponding residential geo-objects for sample production. The total population number of households in each residential geo-objects and the population density obtained by dividing it by the area were calculated by reorganizing census data at corresponding residential geo-objects. Then, three hundred samples with the census population density and population quantity were collected in the study area through ensuring that the space coverage of most of the peasant households and screening of some residential geo-objects with generally high support, confidence, and typicality. The distribution of these sample areas is shown in Fig. 4. Ninety percent of these assigned geo-objects are used to train the prediction model, and the remaining percent is used for performance verification with a rotation way for CV.

3) HSR Remote Sensing Images: Remote sensing image provides data sources of variable visual factors in the spatialization of population data, which promotes the population spatialization with its advantages of fast acquisition speed and good comprehensiveness [47]. Hence, adjacent pixels in Fig. 4 can be merged into corresponding polygons of geo-objects, we further extract their multiple visual image-features into three aspects, namely, spectrum features, shape features, and texture features [48], from the Chinese Gao Fen No. 2 (GF-2) satellite images (i.e., the HSR image of Fig. 4 acquired on March 24, 2018 with a 0.8-m spatial resolution).

First, multispectral reflectance from remote sensing images is an effective information factor for estimating population distribution [49]–[52]. Thus, we exacted the spectral reflectance of each band (i.e., average and standard deviation of spectral signals from all the pixels located in a residential geo-object), brightness (i.e., the average spectral signals from all the bands), maximum differences (i.e., the maximum variation between the spectral signals of bands), spectral indices (i.e., the normalized difference water index and normalized difference vegetation index).

Second, the shape features, including the area, length-width ratio, main direction, shape index [53], and pixel shape index [54] were extracted based on the geometric features of residential geo-objects.

Third, the texture features of remote sensing images are related to the distribution of population density. Especially, texture representation is continuous and indirectly reflect the density of buildings on HSR satellite images and, thus, has a certain potential in population spatial modeling. Hence, in this study, we exacted the texture features consist of measures of the Gray-Level Co-occurrence Matrices [55] and texture-derived built-up presence index [56], [57].

4) Topographic and Geomorphic Data: Topography affects the distribution of population to some extent as human beings always live in the areas meeting certain topographic and geomorphological conditions. One of the conventional understandings is that the population density decreases rapidly with the increase of terrain gradient, and the population mainly distributes on the sunny slope. Therefore, four commonly used modeling factors, including topographic elevation, slope, aspect, and landform class, were derived from a conterminous 30-m advanced spaceborne thermal emission and reflection radiometer global digital elevation model[1] dataset and a public geomorphic dataset with 1-km spatial resolution provided by the Data Center for Resources and Environmental Sciences (DCRES) of the Chinese Academy of Sciences[2]. These topographic feature layers were generated by averaging the values of covered pixels. That is, the mean value in each residential geo-objects was used as their description of topographic features.

5) Meteorological and Climate Data: Small variations in meteorological and climate environment may affect human habitation at a microlevel. Thus, the features of the annual average rainfall, annual average sunshine hours, cumulative annual accumulated temperature, mean temperature, minimum temperature, and maximum temperature were obtained from the datasets of DCRES[3]. Dryness and wetness index products with a 1-km spatial resolution were also used to calculate the corresponding features of the geo-objects.

6) Nighttime Light Data and Impervious Surface Area Data: Nighttime light data and impervious surface area (ISA) data can provide comprehensive information of traffic corridors, land use, and residential areas, which have a good comprehensive indicator for human activity intensity and socioeconomic development [58]–[61].

For nighttime light data, early studies mostly used the data acquired by the Operational Line Scan (OLS) System on the US Defense Meteorological Satellite Program (DMSP) [62]–[64]. Because the spatial resolution of this data is low (about 1 km), the related research on a large and medium scale (national, provincial, and state) is more suitable. Compared with DMSP/OLS nighttime light data, the Visible Infrared Imaging Radiometer Suite (VIIRS) carried by National Polar-orbiting Operational Environmental Satellite System Preparatory Project has a 500-m spatial resolution and is more applicability in small-scale estimation [65], which had been applied in the fields such as GDP simulation [66], [67] and power consumption estimation [68].

In this study, we employed a latest similar data with a 130-m resolution, Luojia-01 nighttime light remote sensing data[4]. The data was produced by Wuhan University and the Hubei Data and Application Center of the High-Resolution Earth Observation System, China, via utilizing the sensing images from June 2018 to December 2018. This raster was collected to derive the nighttime light intensity of each residential geo-object in the study areas. In addition, an ISA product with a 1-km spatial resolution, global distribution, and density of constructed impervious surfaces 2010 (EstISA: Estimate the density of constructed ISA[5]), was employed to extract an indicator of impervious surface coverage and strength in the residential geo-objects.

7) POI, River, and Road Data From Electronic Maps: The densities of POI (including the residential quarters, shops, etc.), river, and roads also indicate the density of the population to a certain extent. We crawled vector data of POI, rivers, and roads from Baidu electronic map[6], and used them to produce raster-layers of point density for POI, line density for rivers, and roads. These layers were also used to extract geo-objects' features.

8) Other Socioeconomic Data: Historical spatialized data of China GDP in 2016 with 1-km spatial resolution provided by the DCRES were further used to express in the features of geo-objects. The grids in these original rasters were downscaled to a 0.8-m resolution using bicubic resampling in GDAL [63] and averaged within each geo-object using their mean for covariate stack. In addition, other public population data sets, including the 2016 global Landscape product with a 1-km spatial resolution[7], the global 2010 WorldPop product with a 100-m spatial resolution[8], and the 2010 China population distribution product with a 1-km spatial resolution[9], were also collected for accuracy comparison.

All of the above data were selected to calculate environmental variables of population formation with the operations of reprojection, resampling, mosaicking, and overlay analysis. Fifty environmental variables were extracted to construct a high-dimensional description of each residential geo-object. A structured multiattribute table in our experimental area was then constructed to train a population density predication model for the geo-objects without observed values.

C. Results and Analysis

1) Residential Geo-object-Based Mapping Results: The processed data were input to the mapping procedure of Fig. 1 and a population density map can be achieved via the predication based on the created relationship on the irregular residential geo-objects. Fig. 6 shows the geo-object-based population density

[1]Online. [Available]: http://www.gdem.aster.ersdac.or.jp/
[2]Online. [Available]: http://www.resdc.cn
[3]Online. [Available]: http://www.resdc.cn
[4]Online. [Available]: http://www.hbeos.org.cn/
[5]Online. [Available]: http://www.ngdc.noaa.gov/dmsp/download_global_isa.html
[6]Online. [Available]: https://map.baidu.com
[7]Online. [Available]: https://landscan.ornl.gov
[8]Online. [Available]: https://www.worldpop.org
[9]Online. [Available]: http://www.geodata.cn

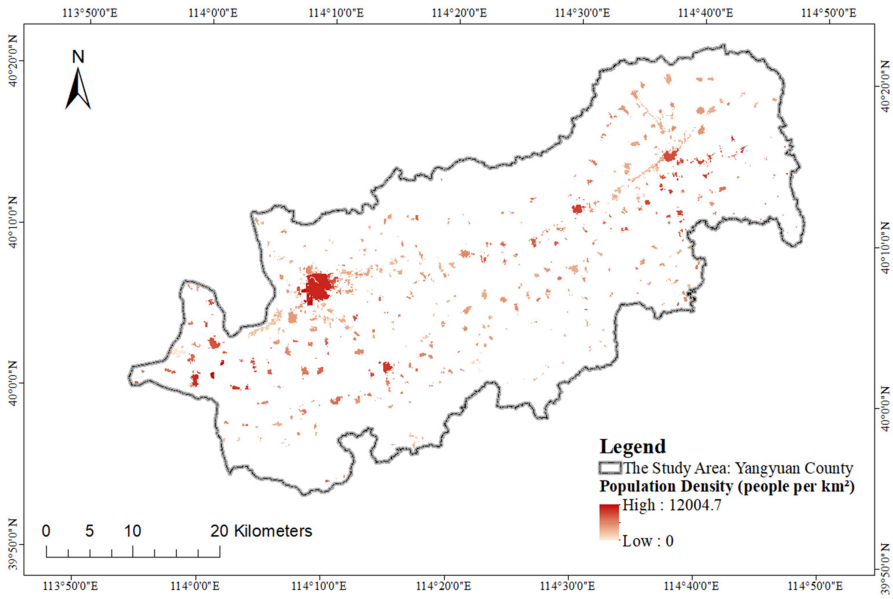

Fig. 6. Residential geo-object-based mapping result of population density.

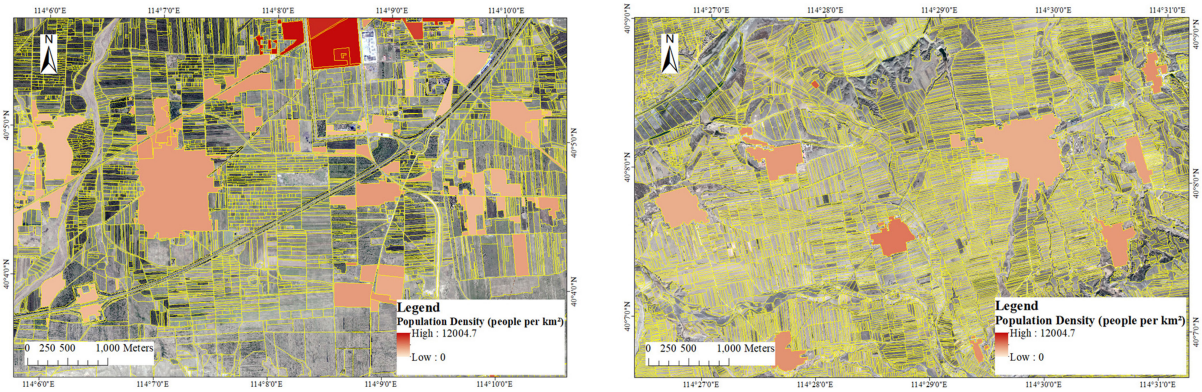

Fig. 7. Predicated values of population density in the residential geo-objects of the two subareas (the bottom image is the GF-2 remote sensing image, the polygons with yellow boundary are the extracted geo-objects, and the colored polygons are residential geo-objects). (a) Predicated population density in the subarea A of Fig. 3. (b) Predicated population density in the subarea B of Fig. 3.

mapping result obtained by using our proposed methodology. Fig. 7 presents the predicated values of population density in the two subareas of Fig. 3. While Figs. 8 and 9 are their disaggregation results of population quantity via weight calculation based on the density.

First, from a global perspective of Figs. 6 and 8, the population of the county is mainly distributed in hilly plains in front of mountains and rivers, especially on both sides of the Sanggan River. While the population distribution is relatively small on both sides of the upper slopes with high altitudes. Macroscopically, these spatial patterns produced on the geo-objects visually match our expert knowledge of the study area.

Second, the bright red areas in these figures indicate regions with high population density and quantity, while the light red areas represent regions with low values. That is, as shown in Fig. 6, the population density in the study area changes in different geo-objects. The figure demonstrates population-concentrated areas. High population density areas are mainly located in the county center and several large town centers. These regions are relatively highly urbanized. The highest population density appeared in the county center with a 12004.7 people per km^2. In addition, there is a pattern of antidistance attenuation from these centers. Especially, in the villages farther away from centers, the population is less distributed. The production and living conditions in these areas are generally poor and not suitable for human habitation. According to the survey, in villages farther away from town centers and rivers, the permanent population has been shrinking in recent years due to the inconvenience of life and employment. They are generally a poverty-stricken zone, which needs the attention and support of the local government. The above findings are generally consistent with the previous cognition on the population distribution in this area.

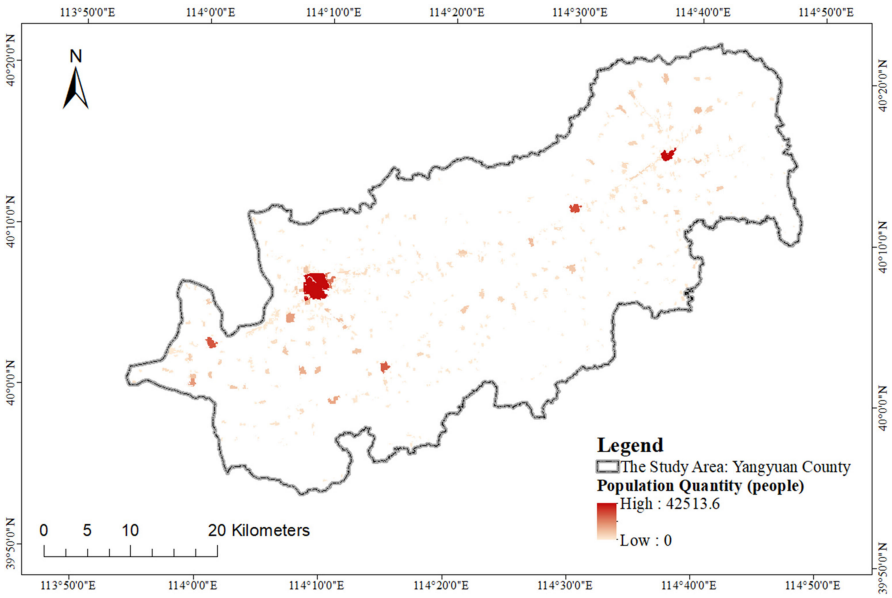

Fig. 8. Residential geo-object-based mapping result of a population quantity.

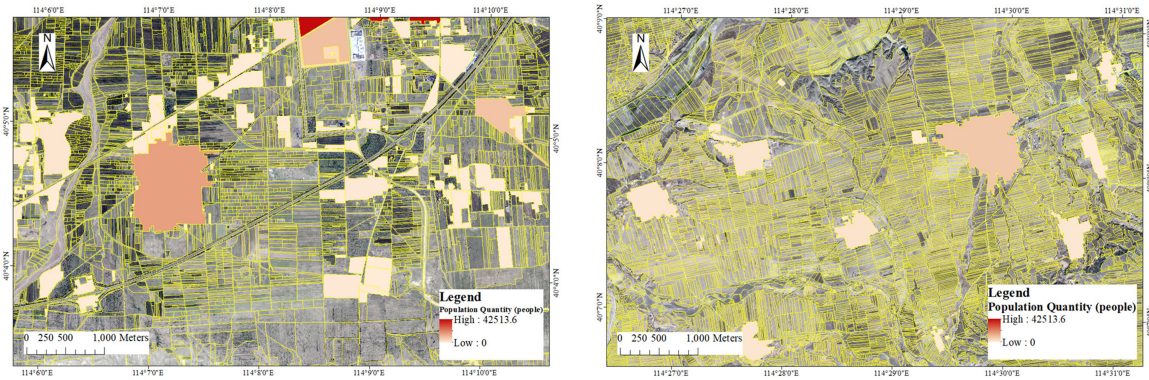

Fig. 9. Predicated values of population quantity in the residential geo-objects of the two subareas (the bottom image is the GF-2 remote sensing image, the polygons with yellow boundary are the extracted geo-objects, and the colored polygons are residential geo-objects). (a) Predicated population quantity in the subarea A of Fig. 3. (b) Predicated population quantity in the subarea B of Fig. 3.

Third, as shown in Figs. 7 and 9, the information on population distribution is clearly apparent in the spatial distribution diagram based on the residential geo-object units. Traditional population spatialization maps are limited by the cartographic units with grid or administrative division, the spatial distribution of population information cannot be delineated in detail. While the geo-object-based maps predicted by our method are spatially visual-detailed with a satisfactory appearance of a continuous surface and gradual transitions. Their fine polygon boundaries of population mapping units are useful to accuracy verification, regional planning, and decision making for solving the local population problems. Effective population management policy and measures should be conducted based on these elaborate mapping results.

2) Accuracy Assessment: Quantitative accuracy assessments in terms of the MAE, RMSE, and R^2 were conducted according to the spatial CV scheme. Table I summarizes the accuracy of the results obtained using our different regression methods based on our framework of Fig. 1 with residential geo-objects. The results indicate that the spatial predication accuracies of population density vary among four methods and those who showing lower MAE and RMSE and a higher R^2 are regarded as a more accurate model in this case study.

TABLE I
COMPARISON OF THE ACCURACY ACHIEVED USING DIFFERENT METHODS

Our regression framework based on residential geo-objects	Measures of evaluating accuracy		
	MAE (people per km^2)	*RMSE* (people per km^2)	R^2
LR	1485.3441	1557.2090	0.54
SVR	1193.4347	1278.8070	0.69
RFs	964.5332	1192.1613	0.72
XGBoost	**894.5761**	**1003.3887**	**0.75**

First, thus, our framework combining multisource data with geo-objects using a regression-based learning is superior in the

terms of mapping accuracy. This is a kind of technical framework that can be relied on in the context of big data.

Second, among the four ML methods, i.e., one LR algorithm and three non-LR algorithms (SVR, RFs, and XGBoost), the performance of nonlinear learning is generally better than that of the LR learning. This is because the relationship between the explanatory variables and response variable is complex in this study, and it is not well described by a simple linear model. In addition, the tree-based methods, namely RFs and XGBoost, outperform than others with statistical significance as their R^2 values are both higher than 0.7. They can improve the prediction accuracy by avoiding over-fitting based on bootstrap sampling, which provides a strong support for rapid and accurate population spatialization.

Third, the values of MAE and RMSE of population density predication are 894.5761 people per km^2 and 1003.3887 people per km^2, respectively, for the XGBoost-based method. These lowest values indicate that the predictions based on XGBoost are better than those based on other algorithms. It is more reliable as its R^2 value is the highest, which also indicates its effectiveness in regression and yields the most accurate results among these typical methods. Meanwhile, this also shows that the gradient boosting scheme implemented in the XGBoost algorithm can build a solid non-LR model [36]. The reasons why XGBoost algorithm provided the best fitting regression model are: 1) first, it uses RFs method for reference to support column sampling, which can not only prevent overfitting, but also reduce calculation; 2) second, it introduces regularization term and pruning technology based on traditional boosting to control the complexity of the model via reducing the variance, compared with the traditional RFs; 3) moreover, it combines the first-order and second-order derivatives in the Taylor expansion, which makes the expansion of cost function in the model learning process more approximate to the loss. These advantages in the XGBoost algorithm make it performed the best results and thus explains the majority of spatial variation of population density.

In conclusion, the nonlinear ML methods in our proposed framework can play a better prediction ability by establishing an appropriate relationship between population density and different environmental factors. The XGBoost algorithm provided the best fitting regression model for accurately obtaining the spatial distribution of population information in our case study.

D. Further Discussion

On the basis of the above experimental analysis, it is necessary to further discuss the following issues.

1) Comparison With Grid-Based Results: Previous studies on fine-scale population spatialization are few, and the spatial resolutions of mapping results are mostly based on a grid with a 1-km or 100-m spatial resolution, which is difficult to meet the requirements of fine management. For this study area, Table II presents the comparison of results with different mapping units. There are nine mapping results, including three public population data sets (i.e., Landscape with 1000-m grids, WorldPop with 100-m grids, and China population distribution product with 1000-m grids), five results using our procedure with grid mapping units of 1000-, 500-, 200-, 100-, and 30-m spatial resolution, and the residential geo-object-based results shown above. Note that, it is difficult to verify the accuracy of spatialized population data on the grid scale as they are inconsistent with geographical entities. We transform the mean values of corresponding pixels and superimpose them on geo-objects for overlay analysis and accuracy verification. In addition, the population quantity at the scale of residential geo-objects was obtained under the control of the total amount of population statistics. The similar control ways were used for the grid-based predictions via the conversion in our procedure, which ensure that the usability and advantages of the proposed procedure for population mapping were not due to the control of the total amount of population statistics. That is, population information mappings were all carried out under the control of the total amount, no matter on the scale of residential geo-objects or grids. Therefore, the comparison here can be ensured that this control factor was consistent.

TABLE II
COMPARISON OF THE ACCURACY ACHIEVED USING DIFFERENT MAPPING UNITS IN DIFFERENT RESULTS

Mapping unit	Measures of evaluating accuracy		
	MAE (people per km^2)	*RMSE* (people per km^2)	*%RMSE* (%)
Landscape grid1000	2824.1624	3934.4741	98.39
WorldPop grid100	1702.2346	2564.6562	64.14
China grid1000	2934.3561	3995.3653	99.91
Our grid1000	2791.4572	3784.5564	94.64
Our grid500	2498.8049	3383.8228	84.62
Our grid200	1976.5756	2973.3167	74.36
Our grid100	1517.2090	2346.2764	58.67
Our grid30	1078.8070	1382.9482	34.58
Our geo-objects	**894.5761**	**1003.3887**	**25.09**

On the one hand, from Table II, we can see that the grid-based mapping results using our framework are better than public data set for the same spatial resolution. Although this is partly caused by the advantage of our method, it is undeniable that this is somewhat unfair for these public products, which were produced several years ago based on a global or national scale, while our results are based on the latest survey in 2018.

On the other hand, with the increase of the spatial resolution of the grid, the mapping accuracies are also improving within our framework. However, even at the 30-m grid scale, the accuracies are still lower than those of the geo-object-based results. Moreover, the improvement in accuracy for geo-object-based mapping is remarkable with respect to grid-based results.

Figs. 10 and 11 show that the population density mapping result based on 500-m grids and 100- m grids, respectively. The details of their results in the two subareas of Fig. 3 (i.e., the subarea A and subarea B) are further presented in Figs. 12 and 13. Obviously, the maps with grid units are visually coarse and cannot produce the information consisting of the geographical entities (i.e., the human residential settlements). On the contrary, our mapping based on residential geo-objects can provides fine population information by using homogeneous polygons with irregular boundaries. The phenomenon of jagged edges in grid-based mapping is also avoided via a more natural and realistic

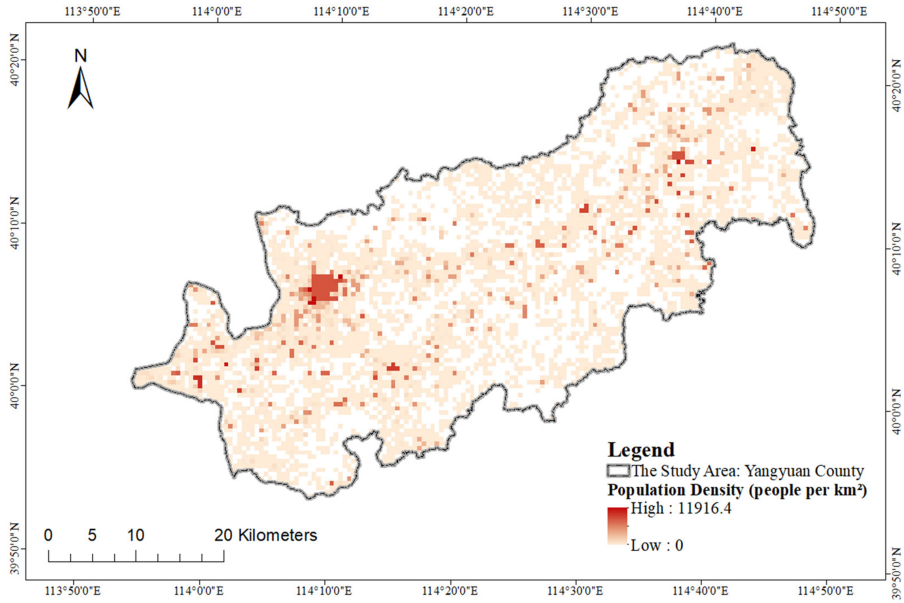

Fig. 10. Grid-based mapping result of population density with a 500-m spatial resolution.

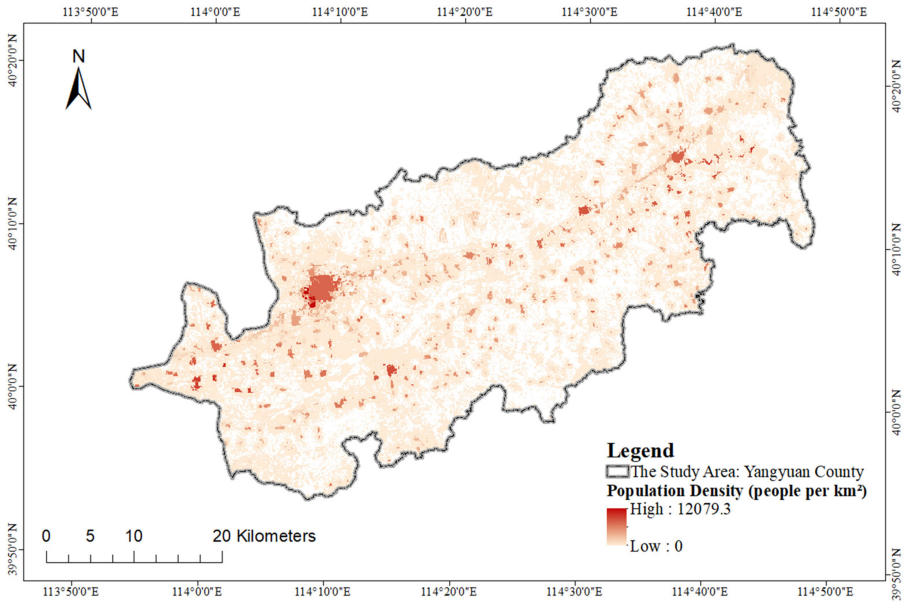

Fig. 11. Grid-based mapping result of population density with a 100-m spatial resolution.

transition. In addition, the grid-based mapping cannot show the synergistic change between the population density and a pure land-use type of residential buildings as many unrelated land classes may be included in one cell at coarse grid resolutions. While microcosmic changes in population distribution in the mapping can be extracted with fine geo-object-based polygons. Therefore, compared with conventional grid-based mapping, our geo-object-based mapping can provide visually superior maps by showing more abundant population information and spatial details in pattern differences. It is conducive to discovery potential population problem and geographic pattern knowledge of population law.

Here, based on the discussion on the comparison with grid results, we would like to further elaborate the different definitions of the population on residential geo-objects and country-level census data. The disaggregation predictions in this article is based on residential geo-objects, which is considered that there is no population in other areas. This assumption is completely acceptable when clear geographical entities (i.e., residential geo-objects) can be obtained by using HSR remote sensing images. While it cannot be established in the mapping products with the coarse grids or administrative units. Among them, the defects of grid-based mapping have been described in detail above. Generally speaking, it is considered that grids with the

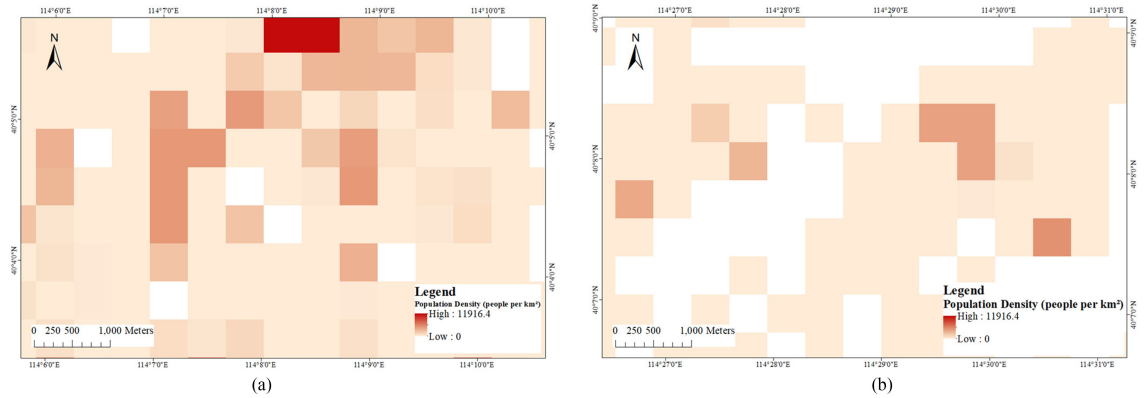

Fig. 12. Details of the grid-based mapping result of population density in the two subareas with a 500-m spatial resolution. (a) Predicated population density in the subarea A of Fig. 3. (b) Predicated population density in the subarea B of Fig. 3.

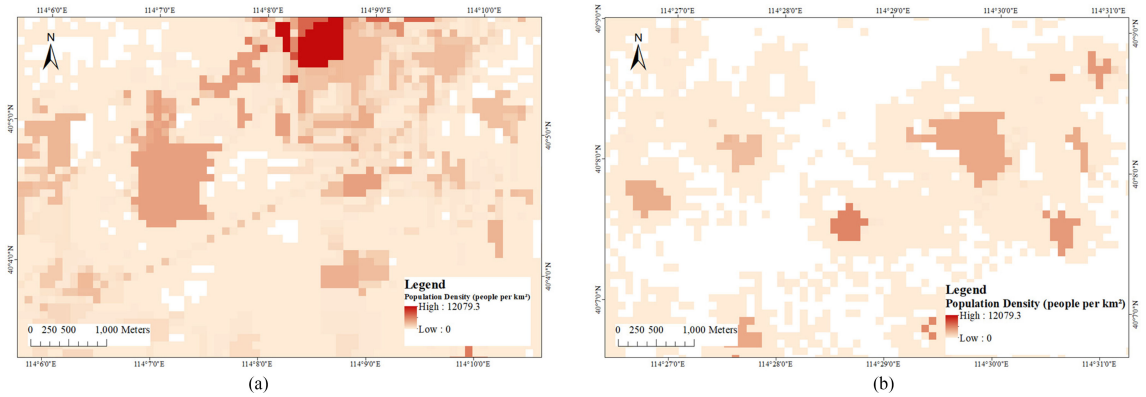

Fig. 13. Details of the grid-based mapping result of population density in the two subareas with a 100-m spatial resolution. (a) Predicated population density in the subarea A of Fig. 3. (b) Predicated population density in the subarea B of Fig. 3.

coarse spatial resolution are mixed pixels of multiple land cover, which cannot be refined as they are not pure population distribution units. As for the census data of administrative units, it shows how many people are distributed in a certain spatial range (e.g., a county or a township), and where the specific distribution cannot be presented. That is, census data is not designed for the purpose of spatialization, and the mapping based on residential geo-object units can achieve the fine spatialization information of population distribution. Printed maps can be used as guides to identify the correct population condition in geographical areas, based on the physical landscape characteristics of residential geo-objects. This is the significance of our study.

2) Relative Importance Analysis of Environmental Variables: As many concerned environmental variables were input in our modeling, the analysis of their importance has an obvious positive effect on understanding the mechanism of influencing population distribution in this study area. Hence, according to the importance measure produced by the XGBoost algorithm based on information entropy, we further extracted the relative importance of different environmental variables in the modeling. The estimated importance of the top 15 significant environmental variables is shown in Fig. 14, from which the following conclusions can be drawn.

First, the water-related factors are important for the distribution of the local population as the most important variable for mapping is the river network density (rank 1st) and the factor of the distance to water is also critical (rank 13th). It is understood that this northern county in China has a problem of

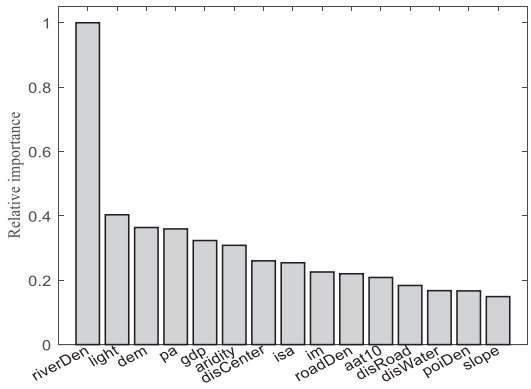

Fig. 14. Relative importance of the top 15 significant environmental variables used in the XGBoost model for prediction of population density (1-riverDen: river network density; 2-light: the feature from nighttime light data; 3-dem: the elevation from DEM data; 4-pa: the feature of annual average rainfall; 5-gdp: the feature from the spatialized data of China GDP; 6-aridity: the dryness index; 7-disCenter: the distance to county center; 8-isa: the feature from EstISA data; 9-im: the wetting index; 10-roadDen: road network density; 11-aat10: the feature of annual accumulated temperature; 12-disRoad: the distance to road; 13-disWater: the distance to water; 14-poiDen: POI density; 15-slope: the slope from DEM data).

water shortage. Surface water cannot be stored. Thus, the rivers are very important for the local people's life and agricultural production. In addition, the inherited idea of living by rivers also influences the choice of residence for generations. Therefore, the water-related factors have become an important cause affecting the population distribution in the study area.

Second, many socioeconomic factors are highly correlated with population distribution. The features from nighttime light data (rank 2nd), the spatialized data of China GDP (rank 5th), EstISA data (rank 8th), have great significance for predicting population information. These indicators show the degree of economic development, indirectly related to the distribution of the population.

Third, the terrain-related factors, including the features of elevation (rank 3rd) and slope (rank 15th) from DEM data, are also important factors influencing the prediction. The reason is that, according to the survey, the local mountains are relatively barren and not conducive to human survival. Local people usually live on flat areas under mountains or on sunny slopes at lower altitudes.

Fourth, among the top 15 important factors, 4 variables are climate-related metrics. There are the features of annual average rainfall (rank 4th), the dryness index (rank 6th), the wetting index (rank 9th), and annual accumulated temperature (rank 11st). This can be explained by the fact that the natural conditions of poor land, drought and water shortage, and cold climate also indirectly affect people's agriculture production and living environment.

Fifth, factors related to urbanization play an important role in population distribution. A shorter distance to county center (rank 7th) and road (rank 12nd), a high density of road network density (rank 10th), and POI (rank 14th) account for a large population distribution. Urban development has the effect of agglomeration and construction landforms built-up areas in succession. There are always more municipal facilities, government organs, health and education institutions, households, and businesses near the county centers. Meanwhile, unobstructed roads are conducive to the communication and transportation of materials and residents. These factors obviously attract more people to live in more urbanized areas. Therefore, the population density in highly urbanized areas is higher. According to Martin's theory of distance attenuation [69], the farther away from these areas, the sparser the population distribution is. The population distribution in the study area also conforms to this model.

Finally, the importance of the remaining variables is relatively tiny. Although they are somewhat less important, they marginally improve the accuracy of prediction. Overall, except for some local particularities such as high terrain, scarce rainfall, serious soil erosion, weak industrial foundation, restricted/prohibited development, the analysis results of the importance of variables are basically consistent with previous studies [17], [18]. The lack of natural resources and protection of the ecological environment will be important considerations for population adjustment.

In addition, we would like to explain how the importance of these environmental variables was calculated. Generally speaking, the score of importance measures the value of features in the construction of decision trees in the model. The more a variable is used to build a decision tree in a model, the more important it is. Therefore, obviously, the importance analysis could be done in both of RFs and XGBoost algorithms. The idea of using RFs to evaluate the importance of features is to see how much contribution each feature has made to each tree, then take an average value of the contribution measured by Gini index or out-of-bag (OOB) error rate. For instance, if random noise is added, the accuracy of OOB data will be greatly decreased, which shows that this variable has a great impact on the prediction results, and thus it is of high importance. While the contribution in the XGBoost algorithm is calculated by the sum of the number of times the feature splits in each tree. That is, the important measures of a variable in all the trees are averaged to get the final score. All these evaluation indexes have their respective interpretability and acceptability. Their different usages is due to the fact that RFs algorithm is an extended variant of bagging, while XGBoost belongs to the framework of boosting. However, the results of the important evaluation with different measure indexes are still similar with a small relative value difference. In our experiment, although the variable importance obtained by XGBoost is relatively more balanced than that of RFs, there is not too much difference in the overall ranking of their orders. This also shows that the importance ranking of these features is robust and credible in our experiment.

3) Novelty of This Research: The novelty of our research can be summarized as follows. First, the basic mapping units in this study are residential geo-objects with finer polygon boundaries, which are more consistent with our cognition on the population aggregation unit. It can reflect the continuity and spatial gradient characteristics of the population distribution and enables us to conduct an analysis on the microspatial scale to meet decision-making requirements for applications. In addition, the detailed and natural boundaries of the geo-objects provide a unified spatio-temporal benchmark for integrating multisource data. Thus, they present high sensitivity to the spatial variability of environmental factors and effectively reflect the impacts of environment on the population distribution on a fine scale.

Second, multisource auxiliary data can be applied synergistically in the mapping of population information via data association on these geo-object units. Potential correlation factors can be identified to improve the accuracy of mapping by designing high-dimensional environmental variables. Such a strategy has the potential to extract the population distribution with spatial variation induced by microenvironmental factors. The factors that are highly correlated with the population property may be quickly detected.

Third, nonlinear modeling based on tree-based ML algorithms, including RFs and XGBoost, are employed in the domain of population mapping, which is demonstrated to be robustness for multivariable relational analysis in accurate predication by effectively mining implicit non-LR relationships. This bottom-up population estimation approach is an important direction for future research due to the growing demand of spatially disaggregated population data with increasingly multiple data sources. In fact, the proposed procedure using microcensus surveys has become popular and highly praised in the field of population mapping [3], [17].

Finally, the proposed methodology provides an effective mapping procedure for a fine population mapping in a rapid way. The learned model can be transferred into adjacent counties with similar environmental conditions. Consequently, compared with a rigorous census, this formalized mapping strategy is cost-effective by playing the role of HSR remote sensing images and various spatial data. Furthermore, the geo-object-level population information data via this rapid production provides an important reference for fine management of population, environment, and social economy and, thus, has practical significance. The popularization of this way will be further enhanced with the wide application of various wearable sensing devices (e.g., smart phones) or social media in different structure of the population as they can be used to ensure enumeration in the correct locations. That is, with the technology development of recording of geographical coordinates, the mapping procedure of this article will be more popular under the advancement of urbanization.

4) Explanations for Further Research: One of the disadvantages of this article is that the model assumes that the population density of each residential geo-object is the same, which leads to the bias of modeling and then affects the accuracy of mapping. In the follow-up study, we can focus on subdividing the type, volume, profile, and height of residential buildings in the same geo-object to further improve the accuracy of the model [8], [26], [70], [71], revealing the similarities and differences of the residential areas with different types of buildings.

In addition, it is urgent to enrich the data sources of population spatial research on a fine scale. More sources of data, such as urban public facilities data [72], mobile phone communication data [73], internet login data of social contact tools [74], and LiDAR data [25], [75], [76], can provide a new database for the bottom-up approach. Generally speaking, the modeling factors tend to be multisource and new type. The diversification of modeling factors poses a challenge for us to reasonably select modeling parameters. Meanwhile, special attention should be paid to the data biases. Hence, in order to improve the accuracy of the model, we can further deeply study the indicative mechanism of various factors on population distribution and then seek reasonable and suitable factors for modeling. These issues are worthy of further investigation.

IV. Conclusion

Population spatialization is an effective way to realize the integration and analysis of population statistics and other spatial data of environmental resources. In this article, fine-mapping units from HSR remote sensing images, called residential geo-objects, were utilized with multisource geospatial data influencing the spatial distribution of the population. Under the framework of bottom-up learning, rich variable factors were selected and further trained a nonlinear model via tree-based ML algorithms. The final spatial map of population quantity can be obtained via a weighting based on the predicated population density generated by the model. According to the designed methodology, the population census data of a county-level study area is successfully disaggregated into the geo-object-based units that may have people to live in. The experimental results show that the overall accuracy of the spatialized map is excellent and better than those grid-based data sets. In addition, by measuring the importance of environmental variables, several types of factors, including the related factors of water, socioeconomic, terrain, climate, and urbanization, are identified as important indicators of population distribution in our study area. It is concluded that the combination of our proposed model and multisource information can robustly achieve population spatialization with high spatial accuracy and thus provide important data sources for fine urban management and scientific decision-making.

Acknowledgment

The authors would like to thank the Population Geography Research Group of Hebei Normal University led by Prof. Jinsong Liu for providing their population survey data in the study area.

References

[1] M. Alahmadi, P. Atkinson, and D. Martin, "Estimating the spatial distribution of the population of Riyadh, Saudi Arabia using remotely sensed built land cover and height data," *Comput. Environ. Urban Syst.*, vol. 41, no. 1, pp. 167–176, Sep. 2013.

[2] J. Geoghegan, L. Pritchard, Y. Ogneva-Himmelberger, R. R. Chowdhury, S. Sanderson, and B. L. Turner, "Socializing the pixel' and 'pixelizing the Social' in land-use and land-cover change," in *People and Pixels: Linking Remote Sensing and Social Science*. Washington, DC, USA: National Academies Press, 1998, pp. 51–69.

[3] N. A. Wardrop *et al.*, "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, Apr. 2018.

[4] G. C. Gallopin, "Human dimensions of global change: Linking the global and local processes," *Int. Soc. Sci. J.*, vol. 130, no. 1, pp. 707–718, Jan. 1991.

[5] D. Azar, R. Engstrom, J. Graesser, and J. Comenetz, "Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data," *Remote Sens. Environ.*, vol. 130, no. 1, pp. 219–232, Mar. 2013.

[6] J. Mennis, "Generating surface models of population using dasymetric mapping," *Prof. Geographer*, vol. 5, no. 1, pp. 31–42, Feb. 2003.

[7] F. J. Gallego, "A population density grid of the European Union," *Population Environ.*, vol. 31, no. 6, pp. 460–473, Jul. 2010.

[8] K. Lwin and Y. Murayama, "A GIS approach to estimation of building population for micro-spatial analysis," *Trans. Geographical Inf. Syst.*, vol. 13, no. 4, pp. 401–414, Aug. 2009.

[9] A. Dmowska and T. F. Stepinski, "High resolution dasymetric model of U.S. demographics with application to spatial distribution of racial diversity," *Appl. Geography*, vol. 53, no. 1, pp. 417–426, Sep. 2014.

[10] M. Langford, "Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps," *Comput. Environ. Urban Syst.*, vol. 31, no. 1, pp. 19–32, Jan. 2007.

[11] J. B. Holt, C. P. Lo, and T. W. Hodler, "Dasymetric estimation of population density and areal interpolation of census data," *Cartography Geographical Inf. Sci.*, vol. 34, no. 2, pp. 103–121, Apr. 2004.

[12] M. D. Su, M. C. Lin, H. I. Hsieh, B. W. Tsasi, and C. H. Lin, "Multi-layer multi-class dasymetric mapping to estimate population distribution," *Sci. Total Environ.*, vol. 408, no. 20, pp. 4807–4816, Sep. 2010.

[13] B. Bhaduri, E. Brigh, P. Coleman, and M. L. Urban, "LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics," *Geo J.*, vol. 69, no. 1, pp. 103–117, Jun. 2007.

[14] J. E. Dobson, E. A. Brlght, 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, no. 7, pp. 849–857, Jul. 2000.

[15] Y. Zhang, C. Dong, J. P. Liu, S. Z. Xu, T. H. Ai, and F. G. Kang, "Gridded population distribution map for the Hebei Province of China," *Environ. Eng. Manag. J.*, vol. 14, no. 3, pp. 673–680, Mar. 2015.

[16] X. H. Yang, Y. H. Huang, P. L. Dong, D. Jiang, and H. H. Liu, "An updating system for the gridded population database of china based on remote sensing, GIS and spatial database technologies," *Sensors*, vol. 9, no. 1, pp. 1128–1140, Feb. 2009.

[17] 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, no. 2, pp. 1–22, Feb. 2015.

[18] A. E. Gaughan *et al.*, "Spatiotemporal patterns of population in mainland China, 1990 to 2010," *Sci. Data*, vol. 3, no. 1, Feb. 2016.

[19] D. L. Balk, U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay, and A. Nelson, "Determining global population distribution: Methods, applications and data," *Adv. Parasitology*, vol. 62, no. 1, pp. 119–156, Apr. 2006.

[20] U. Deichmann, D. L. Balk, and G. Yetman, "Transforming population data for interdisciplinary usages: From census to grid," Center Int. Earth Sci. Inf. Netw., Washington, DC, USA, 2001. [Online]. Available: http://sedac.ciesin.org/gpw-v2/GPWdocumentation.pdf

[21] Center for International Earth Science Information Network. Global Rural-Urban Mapping Project (GRUMP), Alpha version: Urban extents. New York: Center for International Earth Science Information Network (CIESIN), Columbia Univ. Chicago Mag., 2004.

[22] UNEP, "Global resource information database," 1993. [Online]. Available: http://na.unep.net/siouxfalls/datasets/datalist.php

[23] D. Jiang, X. H. Yang, N. B. Wang, and H. H. Liu, "Study on spatial distribution of population based on remote sensing and GIS," *Adv. Earth Sci.*, vol. 17, no. 5, pp. 734–738, May 2002.

[24] P. Jia and A. E. Gaughan, "Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida," *Appl. Geography*, vol. 66, pp. 100–108, Jan. 2016.

[25] J. L. Silvan, L. Wang, P. Rogerson, C. S. Wu, T. T. Feng, and B. D. Kamphaus, "Assessing fine spatial resolution remote sensing for small area population estimation," *Int. J. Remote Sens.*, vol. 31, no. 21, pp. 5605–5634, Nov. 2010.

[26] T. Lung, T. Lübker, J. K. Ngochoch, and J. Schaab, "Human population distribution modelling at regional level using very high resolution satellite imagery," *Appl. Geography*, vol. 41, no.1, pp. 36–45, Jul. 2013.

[27] T. J. Wu *et al.*, "Geo-parcel-based geographical thematic mapping using C5.0 decision tree: A case study of evaluating sugarcane planting suitability," *Earth Sci. Inf.*, vol. 12, no. 12, pp. 57–70, Mar. 2019.

[28] T. J. Wu, J. C. Luo, W. Dong, Y. W. Sun, L. G. Xia, and X. J. Zhang, "Geo-object-based soil organic matter mapping using machine learning algorithms with multi-source spatial data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 12, pp. 1901–1106, Mar. 2019.

[29] T. J. Sauer and D. W. Meek, "Spatial variation of plant-available phosphorus in pastures with contrasting management," *Soil Sci. Soc. Amer. J.*, vol. 67, no. 3, pp. 826–836, May 2003.

[30] C. Göl, S. Bulut, and F. Bolat, "Comparison of different interpolation methods for spatial distribution of soil organic carbon and some soil properties in the Black Sea backward region of Turkey," *J. African Earth Sci.*, vol. 134, no. 1, pp. 85–91, Oct. 2017.

[31] G. B. M. Heuvelink, D. Brus, T. Hengl, B. Kempen, J. G. B. Leenaars, and M. Ruiperez-Gonzalez, "Uncertainty quantification of interpolated maps derived from observations with different accuracy levels," in *Proc. Spatial Accuracy12th Int. Symp. Spatial Accuracy Assessment Natural Res. Environ. Sci.*, 2016, pp. 49–51.

[32] H. K. Zhang and B. Huang, "Support vector regression-based downscaling for intercalibration of multiresolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1114–1123, Mar. 2013.

[33] K. Were, D. T. Bui, B. Ø. Dick, and B. R. Singh, "A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape," *Ecol. Indicators*, vol. 52, no.1, pp. 394–403, May 2015.

[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[35] L. Breiman, "Statistical modeling: The two cultures," *Statist. Sci.*, vol. 16, no. 3, pp. 199–231, 2001.

[36] T. Q. Chen and C. Guestrin, "'XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, San Francisco, USA, Aug. 13–17, 2016, pp. 785–794.

[37] T. Hengl *et al.*, "Mapping soil properties of Africa at 250m resolution: Random forests significantly improve current predictions," *PLoS One*, vol. 10, no. 6, Jun. 2015, Art. no. e0125814.

[38] R. M. Yang *et al.*, "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem," *Ecol. Indicators*, vol. 60, no. 1, pp. 870–878, Jan. 2016.

[39] T. Hengl *et al.*, "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS One*, vol. 12, no. 2, Feb. 2017, Art. no. e0169748.

[40] T. Hengl *et al.*, "Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250m spatial resolution using machine learning," *Nutrient Cycle Agroecosyst.*, vol. 109, no. 1, pp. 77–102, Sep. 2017.

[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.

[42] A. Brenning, "Spatial prediction models for landslide hazards: review, comparison and evaluation," *Natural Hazards Earth Syst. Sci.*, vol. 5, pp. 853–862, Nov. 2005.

[43] A. Brenning, "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package 'sperrorest'," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 5372–5375.

[44] S. Mannel, P. Maribeth, and H. Dong, "Impact of reference datasets and autocorrelation on classification accuracy," *Int. J. Remote Sens.*, vol. 32, no. 19, pp. 5321–5330, Jul. 2011.

[45] J. Pohjankukka, P. Tapio, N. Paavo, and H. Jukka, "Estimating the prediction performance of spatial models via spatial k-fold cross validation," *Int. J. Geographical Inf. Sci.*, vol. 31, no. 10, pp. 2001–2019, Jul. 2017.

[46] T. X. Yue *et al.*, "Surface modeling of the human population distribution in China," *Ecol. Model.*, vol. 181, no. 4, pp. 461–478, Feb. 2005.

[47] W. Qi, S. H. Liu, X. L. Gao, and M. F. Zhao, "Modeling the spatial distribution of urban population during the daytime and at night based on land use: A case study in Beijing, China," *J. Geographical Sci.*, vol. 25, no. 6, pp. 756–768, Jun. 2015.

[48] T. J. Wu, J. C. Luo, L. G. Xia, Z. F. Shen, and X. D. Hu, "Prior knowledge-based automatic object-oriented hierarchical classification for updating detailed land cover maps," *J. Indian Soc. Remote Sens.*, vol. 43, no. 4, pp. 653–669, Dec. 2015.

[49] C. P. Lo, "Automated population and dwelling unit estimation from high-resolution satellite images: A GIS approach," *Int. J. Remote Sens.*, vol. 16, no. 1, pp. 17–34, Mar. 1995.

[50] C. J. Webster, "Population and dwelling unit estimation from space," *Third World Plann Rev.*, vol. 18, no. 2, pp. 155–176, Jun. 1996.

[51] J. T. Harvey, "Population estimation models based on individual TM pixels," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 11, pp. 1181–1192, Nov. 2002.

[52] X. H. Liu, C. Keith, and H. Martin, "Population density and image texture: A comparison study," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 2, pp. 187–196, Feb. 2006.

[53] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *Int. J. Remote Sens.*, vol. 28, no. 5, pp. 823–870, Mar. 2007.

[54] L. P. Zhang, X. Huang, and B. Huang, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.

[55] D. J. Marceau, P. J. Howarth, J. M. Dubois, and D. J. Gratton, "Evaluation of the grey-level co-occurrence matrix method for land-cover classification using spot imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 513–519, Jul. 1990.

[56] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A robust built-up area presence index by anisotropic rotation-invariant textural measure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 3, pp. 180–192, Sep. 2008.

[57] M. Pesaresi and A. Gerhardinger, "Improved textural built-up presence index for automatic recognition of human settlements in arid regions with scattered vegetation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 16–26, Mar. 2011.

[58] P. Sutton, "Modeling population density with night-time satellite imagery and GIS," *Comput. Environ. Urban. Syst.*, vol. 21, no. 3, pp. 227–244, May 1997.

[59] C. P. Lo, "Modeling the population of China using DMSP operational linescan system nighttime data," *Photogramm. Eng. Remote Sens.*, vol. 67, no. 9, pp. 1037–1047, Sep. 2001.

[60] L. Prosperie and R. Eyton, "The relationship between brightness values from a nighttime satellite image and Texas County population," *Southwest Geographical*, vol. 92, no. 2, pp. 224–240, Dec. 2002.

[61] 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, no. 4, pp. 451–466, Jun. 2007.

[62] C. Small, F. Pozzi, and C. D. Elvidge, "Spatial analysis of global urban extent from DMSP- OLS night lights," *Remote Sens. Environ.*, vol. 96, no. 3–4, pp. 277–291, Jun. 2005.
[63] P. C. Sutton, D. A. Roberts, C. D. Elvidge, and H. Melj, "A comparison of nighttime satellite imagery and population density for the Continental United States," *Photogramm. Remote Sens.*, vol. 63, no. 11, pp. 1303–1313, Nov. 1997.
[64] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/ OLS nighttime light data," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011.
[65] K. Baugh, F-C. Hsu, C. D. Elvidge, and M. N. Zhizhin, "Nighttime lights compositing using the VIIRS day-night band: Preliminary results," *Proc. Asia-Pac. Adv. Netw.*, vol. 35, pp. 70–86, Jun. 2013.
[66] Z. Dai, Y. Hu, and G. Zhao, "The suitability of different nighttime light data for GDP estimation at different spatial scales and regional levels," *Sustainability*, vol. 9, no. 2, pp. 305–320, Feb. 2017.
[67] X. Li, H. M. Xu, X. L. Chen, and C. Li, "Potential of NPP-VIIRS nighttime light imagery for modeling the regional economy of China," *Remote Sens.*, vol. 5, no. 6, pp. 3057–3081, Jun. 2013.
[68] K. F. Shi *et al.*, "Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data," *Remote Sens.*, vol. 6, no. 2, pp. 1705–1724, Jan. 2014.
[69] D. Martin, "Mapping population data from zone centroid locations," *Trans. Inst. British Geographers*, vol. 14, no. 1, pp. 90–97, Feb. 1989.
[70] S. Ural, E. Hussain, and J. Shan, "Building population mapping with aerial imagery and GIS data," *Int. J. Appl. Earth Observ. GeoInf.*, vol. 13, no. 6, pp. 841–852, Dec. 2011.
[71] F. Biljecki *et al.*, "Population estimation using a 3D city model: a multi-scale country-wide study in the Netherlands," *PLoS ONE*, vol. 11, no. 6, Jun. 2016, Art. no. e0156808.
[72] M. Bakillah, S. Liang, A. Mobasheri, J. J. Arsanjani, and A. Zipf, "Fine-resolution population mapping using OpenStreetMap points-of-interest," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 9, pp. 1940–1963, Sep. 2014.
[73] C. G. Kang, Y. Liu, X. J. Ma, and L. Wu, "Towards estimating urban population distributions from mobile call data," *J. Urban Technol.*, vol. 19, no. 4, pp. 3–21, Oct. 2012.
[74] N. N. Patel, F. R. Stevens, Z. J. Huang, A. E. Gaughan, I. Elyazar, and A. J. Tatem, "Improving large area population mapping using geotweet densities," *Trans. GIS*, vol. 21, pp. 317–331, Jun. 2016. doi: 10.1111/tgis.12214.
[75] P. L. Dong, S. Ramesh, and A. Nepalia, "Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data," *Int. J. Remote Sens.*, vol. 31, no. 21, pp. 5571–5586, Nov. 2010.
[76] L. Tomás, L. Fonseca, C. Almeida, F. Leonardi, and M. Pereira, "Urban population estimation based on residential buildings volume using IKONOS 2 images and LiDAR data," *Int. J. Remote Sens.*, vol. 37, no. 1, pp. 1–28, Dec. 2015.

Tianjun Wu received the B.S. degree in information science and the M.S. degree in applied mathematics from Chang'an University, Xi'an, China, in 2009 and 2012, respectively, and the Ph.D. degree in cartography and geographical information system from Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS), Beijing, China, in 2015.

He is currently an Associate Professor with the School of Geology Engineering and Geomatics, Chang'an University. His research interests include the intelligent information extraction from remote sensing images and spatiotemporal data mining.

Jiancheng Luo received the B.S. degree in remote sensing from Zhejiang University, Hangzhou, China, in 1991, the M.S. degree in GIS from Chinese Academy of Sciences (CAS), Beijing, China, in 1996, and the Ph.D. degree in GIS from the Institute of Geographical Sciences and Natural Resources, CAS, in 1999.

He is currently a Professor with the Institute of Remote Sensing and Digital Earth, CAS. His research interests include artificial intelligence techniques in remote sensing.

Wen Dong received the B.S. degree in environment engineering from China Agricultural University, Beijing, China, in 2005, and the Ph.D. degree in cartography and geographical information system from Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS), Beijing, China, in 2010.

She is currently a Research Associate with the State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, CAS. Her research interests include geographic spatio-temporal data mining and application of geographic information system.

Lijing Gao received the B.S. degree in GIS from Shandong University of Science and Technology, Qingdao, China, in 2006, and the M.S. degree in GIS from Institute of Remote Sensing Applications, CAS, Beijing, China, in 2009.

She is currently a Research Assistant with the Institute of Remote Sensing and Digital Earth, CAS. Her research interest include vegetation classification and mapping in mountain areas.

Xiaodong Hu received the B.S. degree in computer science and technology and the M.S. degree in computer software and theory from Zhejiang University of Technology, Hangzhou, China, in 2005 and 2008, the Ph.D. degree in GIS from the Institute of Remote Sensing Application, CAS, Beijing, China, in 2011.

He is currently a Research Assistant with the Institute of Remote Sensing and Digital Earth, CAS. His research interests include remote sensing information extraction and adaptive computing.

Zhifeng Wu received the B.S. degree in geography education from Hunan Normal University, Changsha, China, in 1992, the M.S. degree in physical geography from South China Normal University, Guangzhou, China, in 1995, and the Ph.D. degree in GIS from the Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences, Beijing, China, in 2002.

He is currently a Professor with the School of Geographical Sciences, Guangzhou University, Guangzhou, China. His research interests include urban remote sensing, land ecological remote sensing, spatiotemporal data analysis, and monitoring and assessment of natural resources.

Yingwei Sun received the B.S. degree in remote sensing science and technology from Shandong University of Science and Technology, Qingdao, China, in 2013, and the M.S. degree in geological engineering from China University of geosciences, Wuhan, China, in 2017. He is currently working for the Ph.D. degree in cartography and geographical information system from Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS), Beijing, China.

His research interests include the intelligent information calculation from remote sensing images.

Jinsong Liu received the B.S. degree in geography education, the M.S. degree in physical geography, and the Ph.D. degree in ecology from Hebei Normal University, Shijiazhuang, China, in 1992, 1995, and 2011, respectively.

He is currently a Professor with the School of Resources and Environmental Sciences, Hebei Normal University. His research interest focuses on population geography.