remote sensing

MDPI

Article

Mapping the Population Density in Mainland China Using NPP/VIIRS and Points-Of-Interest Data Based on a Random Forests Model

Yunchen Wang [1,2], Chunlin Huang [1,*], Minyan Zhao [3], Jinliang Hou [1], Ying Zhang [1] and Juan Gu [4]

1 Key Laboratory of Remote Sensing of Gansu Province, Heihe Remote Sensing Experimental Research Station, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; wangyunchen@lzb.ac.cn (Y.W.); jlhours@lzb.ac.cn (J.H.); zhangy@lzb.ac.cn (Y.Z.)
2 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100000, China
3 UCD School of Civil Engineering, University College Dublin, D04 V1W8 Dublin, Ireland; zhao.minyan@ucdconnect.ie
4 Key Laboratory of Western China's Environmental Systems (Ministry of Education), Lanzhou University, Lanzhou 730000, China; gujuan@lzu.edu.cn
* Correspondence: huangcl@lzb.ac.cn

Received: 28 August 2020; Accepted: 3 November 2020; Published: 6 November 2020

Abstract: Understanding the spatial distribution of populations at a finer spatial scale has important value for many applications, such as disaster risk rescue operations, business decision-making, and regional planning. In this study, a random forest (RF)-based population density mapping method was proposed in order to generate high-precision population density data with a 100 m × 100 m grid in mainland China in 2015 (hereafter referred to as 'Popi'). Besides the commonly used elevation, slope, Normalized Vegetation Index (NDVI), land use/land cover, roads, and National Polar Orbiting Partnership/Visible Infrared Imaging Radiometer Suite (NPP/VIIRS), 16,101,762 records of points of interest (POIs) and 2867 county-level censuses were used in order to develop the model. Furthermore, 28,505 township-level censuses (74% of the total number of townships) were collected in order to evaluate the accuracy of the Popi product. The results showed that the utilization of multi-source data (especially the combination of POIs and NPP/VIIRS data) can effectively improve the accuracy of population mapping at a finer scale. The feature importances of the POIs and NPP/VIIRS are 0.49 and 0.14, respectively, which are higher values than those obtained for other natural factors. Compared with the Worldpop population dataset, the Popi data exhibited a higher accuracy. The number of accurately-estimated townships was 19,300 (67.7%) in the Popi product and 16,237 (56.9%) in the Worldpop product. The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were 14,839 and 7218, respectively, for Popi, and 18,014 and 8572, respectively, for Worldpop. The research method in this paper could provide a reference for the spatialization of other socioeconomic data (such as GDP).

Keywords: population density; NPP/VIIRS; POI; random forest; China

1. Introduction

Population density data, presenting a realistic description of the spatial distribution of a population, is of great significance for mapping human settlements [1], quantifying the number of populations threatened by infectious diseases [2], calculating the birth rate and mortality [3], quantifying the spatial distribution of poverty and hunger [4], and evaluating Sustainable Development Goals (SDGs) [5]. According to data from the National Bureau of Statistics of China (http://www.stats.gov.cn/tjsj/pcsj/),

at present, China is one of the most populous countries in the world, with a population of over 1.4 billion. Due to its complex natural conditions and large range, fine scale population density mapping is facing challenges, including the way in which to obtain accurate data as a modeling factor, and how to build models to simulate complex nonlinear relationships between the population and heterogeneous geographic covariates. Therefore, the way in which to generate accurate population density data is a concern for scholars [6].

Census data are supported by rigorous statistical theories and methods; however, there are some problems that arise in its application. For example, census data represents the number of populations within the administrative division, which cannot reflect the spatial distribution and heterogeneity of the population within the administrative unit. Next, a finer scale is required for actual research. However, the differences between the census statistical caliber and the research units may cause the Modifiable Areal Unit Problem [7]. Finally, census data are generally difficult to obtain in the amount of time required. It is difficult to obtain statistics in many countries. For example, in China, there is a census every ten years, and it is hard to obtain such data in sparsely populated areas in western China [8]. Therefore, scholars utilize population density data as a supplement to and replacement for census data, as population density data can easily and flexibly be integrated with other spatial data [9].

Although population density data can express spatial details more clearly, the current population density products still have several deficiencies. For example, the Gridded Population of the World (GPW) dataset uses the weighted-average method to estimate the population density [10]. This method makes use of census data divided by the corresponding administrative division area, which is prone to errors when administrative boundaries change. The Global Rural-Urban Mapping Project (GRUMP) dataset is also based on the weighted-average method [11]. Although it is better than the GPW dataset in terms of the estimation of the urban population density, it still does not solve the limitations of the method itself. Landscan products extract relevant information from remote sensing and GIS data to build a population spatialization model [12]. However, Landscan applies provincial-level population data modeling in China, and the population data resolution is coarse. The China 1 km Gridded Population (CnPop) dataset utilizes the least-squares method [13], and the accuracy almost depends on land cover data, due to the lower number of types of modeling factors included. Worldpop employs a random forest (RF) model to calculate the population density [14]. Bai et al. [9] used the relative error (RE) and the 2000 township population data obtained from mainland China to evaluate the accuracy of the above dataset. The results show that, although the accuracy of Worldpop data are better than GPW, GRUMP, and CnPop data, 42.6% of the population still have an absolute RE value that exceeds 0.25. Therefore, it is still challenging to produce population density data.

Modeling factors and models are the two most important aspects of building population density models. On the one hand, common modeling factors include land cover data, the slope, the Digital Elevation Model (DEM), precipitation, and the Normalized Vegetation Index (NDVI), etc. [6]. In addition to the above natural factors, socioeconomic factors are also utilized as modeling factors. Scholars have proved that the shape and brightness contained in the nighttime light (NTL) data can provide reliable information for population density modeling [15], since NTL is highly related to human activities [16]. Commonly-used NTL data include the Defense Meteorological Satellite Program/Operational Linescan System (DMSP/OLS) and National Polar-orbiting Operational Environmental Satellite System Preparatory Project/Visible Infrared Imaging Radiometer (NPP/VIIRS) data, which are normally applied to estimate the spatial distribution of CO_2 [17], GDP [18], and the population density [19]. Since the DMSP/OLS image is not calibrated by the onboard radiation, the digital number (DN) value is the relative brightness radiation value, with the problem of pixel saturation and discontinuity [20]. NPP/VIIRS data has been used for radiometric calibration, and can identify weak light. The resolution of NPP/VIIRS is 15 arc-seconds; thus, is better than the DMSP/OLS data resolution (30 arc-second). Yu et al. [21] estimated the population density in Shanghai, China, via NPP/VIIRS data, proving the feasibility of NPP/VIIRS data in estimating the population density.

Additionally, WeChat, Twitter, points of interest (POIs), and other data provide new sources of socioeconomic factors for population spatialization. Among them, POI is a type of social perception data, with attribute information including the location (latitude and longitude) and type (such as food, company, etc.). If each POI record is regarded as a functional unit, the higher the POI density, the more concentrated the urban function [22]. The urban function directly affects the spatial distribution of the population to a certain extent [23]. For example, Li et al. [24] used POI data to estimate the population density, confirming that POI data can be used as a population spatialization modeling factor. As typical geospatial big data, POI data also has the following advantages. POIs have the characteristics of easy access, rich data volume, high positioning accuracy, and more reflective micro details, making up for the shortcomings of traditional data, such as long update periods and coarse resolutions [25]. Moreover, POI data can be converted into raster layers with different scales, which can be combined with geospatial data and remote sensing data, etc.

However, there will also be an overestimation, underestimation, or distribution to inappropriate locations when a socioeconomic factor is used alone. For example, Yu et al. [21] found that the population was overestimated in a commercial district and transportation hub where the NPP/VIIRS intensity was too high. Similarly, when the POIs are used as the only variable, many people may be allocated to undeveloped areas due to the default assumption that each POI has the same attractiveness [26]. Therefore, the combination of multiple natural factors and social-economic factors has become a new trend. Since NPP/VIIRS and POI data have their own advantages, the combination of these two data types can provide a new approach for the estimation of large-scale and high-precision population densities, which is not implemented at present.

On the other hand, the accuracy of population density estimations is still limited by the algorithm. Common methods include spatial regression models (e.g., the gravity model, the spatial autocorrelation model, and geographically weighted regression (GWR)), and the linear regression model (e.g., logistic regression). However, spatial regression models need to be based on a specific assumption. For example, Sutton et al. [27] utilized the distance weighting method, which is based on the attenuation law between population and distance. This is reasonable, to a certain extent, but the modeling results will produce greater numbers of errors in more complicated situations. Wang et al. [28] used GWR in China from 1990 to 2010. Although this method can solve spatial heterogeneity, the GWR model may lead to a meaningless negative population density in rural areas. The linear model needs to assume that there is a linear relationship between the population and the modeling factors, which is contrary to reality in large-scale research areas. It is difficult for the regression model to reflect the spatial heterogeneity of the population distribution [29].

In recent years, some scholars have used machine learning to estimate the population density [14]. Machine learning algorithms include support vector regression, random forest (RF), and neural networks, etc. Among them, the RF model uses multiple decision trees in order to achieve data classification or regression [30] and is a non-parametric method that can simulate complex nonlinear relationships. The RF model introduces double random processes (sample randomness and feature randomness), making the model less prone to overfitting. Furthermore, Gaughan et al. [6] calculated the population density through the RF model, which proved the feasibility of the RF model in realizing population spatialization.

In summary, large-scale population density mapping needs to consider many natural and social-economic factors. The combination of multiple modeling factors may improve the model's accuracy. The remarkable advancement of machine learning can provide a toolkit for demographers, which is conducive to the simulation of the complex nonlinear relationship between the population and heterogeneous geographic covariates. Therefore, using NPP/VIIRS, POI data, land use/land cover (LULC) data, NDVI, roads, Digital Elevation Model (DEM) data, and censuses, this study constructed the RF regression model in order to calculate the population density data at a 100 m × 100 m resolution in mainland China in 2015.

The remainder of this paper is structured as follows: Sections 2 and 3 introduce the data and methods, respectively; the results, accuracy verification, and feature importance of the modeling factors are outlined in Section 4; in the discussion section of Section 5, the differences between POIs and NPP/VIIRS, the response of POI data, and the error analysis in mapping the population density are analyzed; and finally, Section 6 outlines the conclusions.

2. Data

The data used in this paper include NPP/VIIRS, POI, LULC, NDVI, DEM, road, and census data (Table 1).

Table 1. Data resources.

Datasets	Data Declaration	Time	Sources
NPP/VIIRS	750 m	2015	National Oceanic and Atmospheric Administration, (NOAA) (https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html)
POIs	Point features	2015	Baidu Maps API (http://lbsyun.baidu.com/)
Land use/ land cover	100 m	2015	Chinese Academy of Sciences Resource and Environmental Science Data Center (http://www.resdc.cn/data.aspx?DATAID=99)
NDVI	250 m	2015	Moderate-resolution Imaging Spectroradiometer (MODIS) (http://ladsweb.modaps.eosdis.nasa.gov/)
DEM	90 m	2000	National Aeronautics and Space Administration (NASA) (http://srtm.csi.cgiar.org/srtmdata/)
Roads	Line features	2015	Baidu Maps Application Programming Interface (API) (http://lbsyun.baidu.com/)
Census	County Township	2015	2016 Statistical Yearbook of Provinces and Cities, National Bureau of Statistics of China (http://www.stats.gov.cn/tjsj/pcsj/, http://www.ngcc.cn/ngcc/)
Worldpop	100 m	2015	University of Southampton (https://www.worldpop.org/geodata/listing?id=16)
Administrative boundary map	County Township	2015	National Fundamental Geography Information System (http://www.ngcc.cn/ngcc/)

The NPP/VIIRS satellite imagery was obtained from a new generation of polar-orbiting satellite system preparation projects. The first version of the 'vcm-orm-ntl' annual synthetic product was used in this paper, which excludes stray light and filters out transient lights, background values, and abnormal values, forming cloudless average radiation value data.

POI is a type of social perception data, and the duplicate records need to be deleted. Within mainland China, a total of 16,101,762 records of POI data were obtained in 2015. All of the POIs were classified into 10 categories based on their type (Table 2).

LULC data, representing one of the most accurate long-term products in China, was obtained by the visual interpretation of Landsat satellite images, and the accuracy exceeds 75% [31,32]. The types of LULC data used in this paper include four classifications: cultivated land, forest land, grassland, and construction land.

The NDVI data include the MOD13Q1 product, with a spatial resolution of 250 m. The data are 16-day synthetic data. It is necessary to mosaic all of the NDVI on the same day to generate a full map of China.

The DEM data were employed in order to obtain elevation data and slope data.

The road data includes seven types (railways, national roads, provincial roads, high-speed roads, urban roads, county roads, and village roads). A total of 13,429,160 road data sets for mainland China were obtained.

Table 2. 10 categories of the Point of Interest (POI) data downloaded from Baidu Application Programming Interface.

Classification	Content
Transportation	Airports, railway stations, bus stations, terminals, ferries, etc.
Government	Government agencies
Village	Villages
Education	Universities, research institutes, middle schools, kindergartens, etc.
Food	Restaurants, canteens, etc.
Medical	Hospitals, pharmacies, health service stations, clinics, etc.
Entertainment	Business centers, supermarkets, shops, wholesale markets, etc.
Accommodation	Guest, hotels, etc.
Working place	Companies, factories, etc.
Financial	ATMs, banks, savings centers, etc.

Additionally, 2867 county-level censuses, with a total of 1,374,620,000 people, were obtained and were used as the dependent variable in order to establish the RF model. The township level is the smallest unit of the census in China, and is a lower administrative unit than the county level. This study also obtained 28,505 township censuses (accounting for 74% of all of the townships in China), which were employed in order to analyze the model's accuracy (Figure 1).

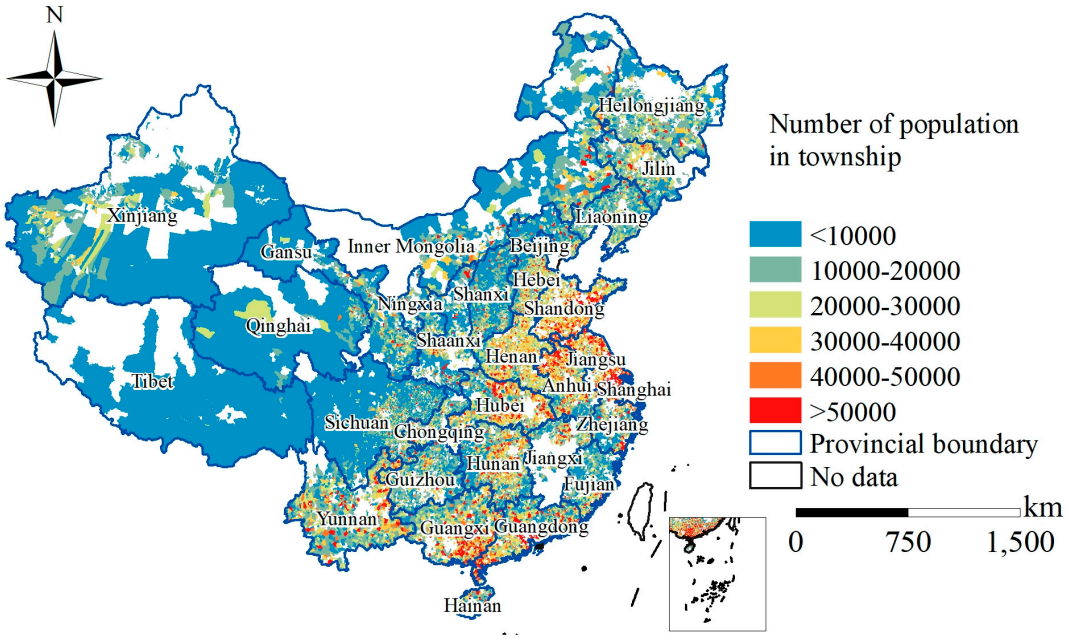

Figure 1. The classification of 28,505 township-level census data sets in China. The township census data of the Tianjin, Hong Kong, Macau, and Taiwan provinces are missing.

The Worldpop global population data set was provided by the University of Southampton (https://www.worldpop.org/geodata/listing?id=16), with a spatial resolution of 100 m. The Worldpop program provide high resolution, open and contemporary data in human population distributions, which makes the measurement of the population density more accurate.

In this paper, the data were resampled to 100 m × 100 m.

3. Methods

3.1. Mapping the Population Density with the RF Model

This paper first constructs a non-linear relationship between the modeling factors and censuses at the county level. Then, the non-linear relationship at the county level is applied to the grid level,

in order to achieve the spatialization of the population data. As shown in Figure 2, Step 1 is the establishment of the RF model: the 10 NDVI, slope, elevation, NPP/VIIRS, POI density layer, road distance layer, and LULC (cultivated land, forest, grassland, and construction land) preprocessed layers are aggregated at the county level (the preprocessing of independent variables is described in Section 3.2). Taking the above-10 county-level summary values as independent variables, and the natural logarithm of the county-level census as the dependent variable, all of the variables need to be divided into train sets and test sets. A total of 70% of the variables are randomly selected as train sets for the training of the RF model. The remaining 30% form the test set, which is used to calculate the model's accuracy. Step 2 involves predicting the population density. By taking 10 independent variable raster layers as prediction sets, the prediction sets are input into the RF model to obtain the population density data. Step 3 includes the dasymetric population mapping, where the population density data are adjusted according to the relationship between the census and the population density calculated by the RF model. The final population density data (named Popi) is calculated by Equation (1):

$$p_g = \frac{p'_g \times p_c}{p'_c}, \tag{1}$$

where p_g is the Popi data, p'_g is the population density data calculated by the RF model, p_c is the census data at the county level, and p'_c is the aggerated value of population density data calculated by the RF model at the county level.

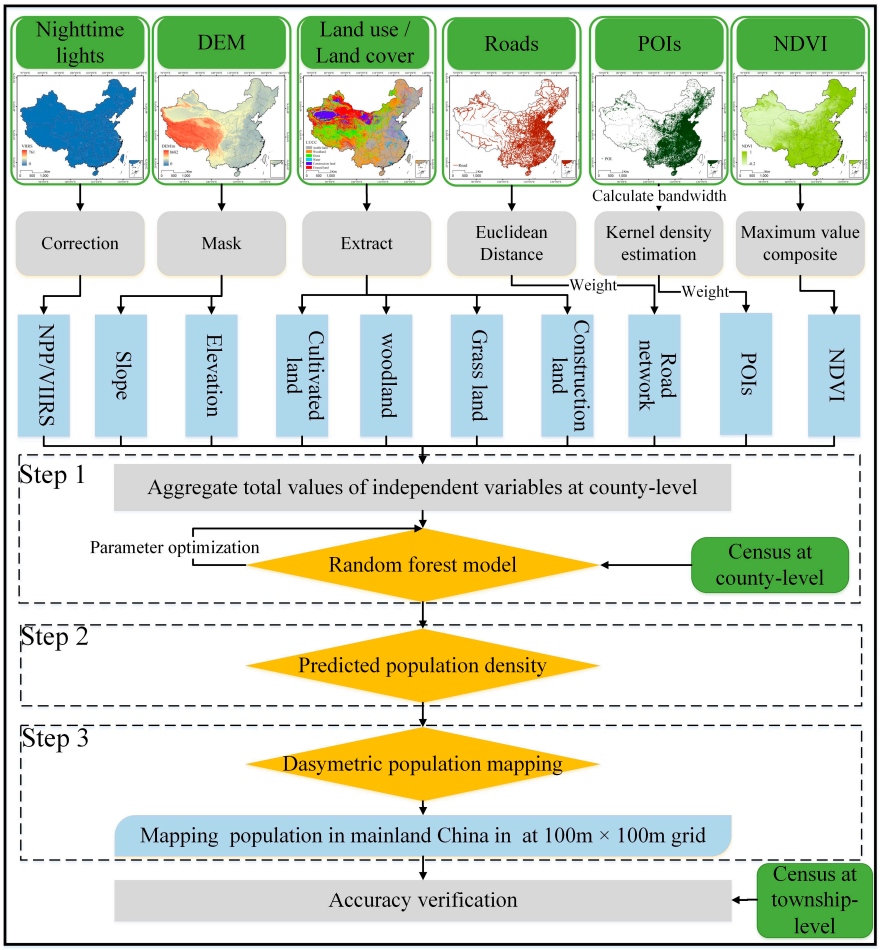

Figure 2. A flowchart of the mapping of the population density.

3.2. Preprocessing of the Remote Sensing Data and Geographic Big Data

3.2.1. Eliminating the Background Noise and Extreme Values of the NPP/VIIRS Data

Due to the strong sensitivity of the sensor, it is easy to detect objects such as instant lights and aurora. Therefore, the NPP/VIIRS data requires the following corrections. First, the background noise needs to be eliminated: the non-zero DN value in 2013 the DMSP/OLS image is used as a mask and the area outside the mask is regarded as a noise area. Then, the 2015 NPP/VIIRS data are extracted by the mask. Additionally, the maximum value needs to be excluded: according to the assumption proposed by Shi et al. [33], the maximum DN value in NPP/VIIRS cannot exceed the maximum DN value in the most developed cities in mainland China. This paper takes the maximum DN value (400) of the most economically developed cities (Beijing, Shanghai, and Guangzhou) in the study area as the threshold. The eight-neighborhood algorithm is utilized for the smoothing of the pixels which are larger than the threshold.

It can be seen from Figure 3 that the corrected NPP/VIIRS data eliminates the background noise and DN values greater than 400. In addition, since the resolution of the NPP/VIIRS data is 750 m, it needs to be matched with other data. Considering that NPP/VIIRS is the current high-resolution global open source night light data, and that we also have other high-resolution data to provide detailed information, we use the nearest neighbor method to resample the NPP/VIIRS data to 100 m.

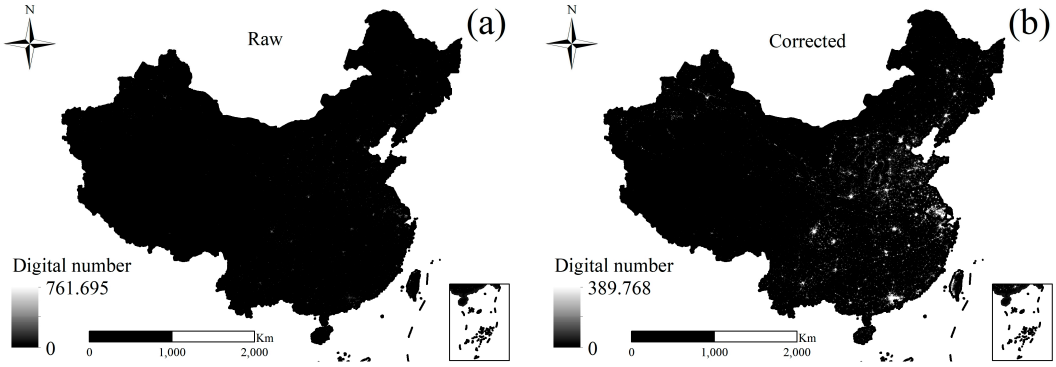

Figure 3. The NPP/VIIRS data in mainland China in 2015. (**a**) The raw NPP/VIIRS data; (**b**) the corrected NPP/VIIRS data.

3.2.2. Producing NDVI Annual Synthetic Data

The maximum value composite (MVC) method is used to generate annual synthetic NDVI images from multi-temporal NDVI images in order to separate human settlements from other land cover types, and to eliminate the effects of cloud pollution [2]. The MVC method is shown in Equation (2):

$$NDVI_{max} = MAX\ (NDVI_1, NDVI_2, \ldots, NDVI_{23}), \quad (2)$$

$$NDVI = NDVI_{max}/10000, \quad (3)$$

where NDVI is the annual synthetic image; $NDVI_{max}$ is an unscaled annual synthetic image; and $NDVI_1$, $NDVI_2$, ... , $NDVI_{23}$ are images obtained after the mosaic process. Since the $NDVI_{max}$ value range is −10000 to 10000, Equation (3) can scale the $NDVI_{max}$ value between −1 and 1.

3.2.3. Generating the POI Density Layers

The POI density layer is generated from POI point data by the kernel density estimation (KDE) method. In order to ease the burden of the calculation and make full use of POI, the entropy method is used to merge 10 POI density layers into one layer.

First, the 10 types of discrete POI point data (vector) are utilized to generate 10 smooth surfaces (the raster layer) of 100 m × 100 m through the KDE with a 6500 m bandwidth, respectively. KDE can calculate the density of a certain type of POI point within a given range. The optimal bandwidth is determined as follows: when setting the bandwidth, the step length is set to 100 m in the range of 1000–10,000 m, and the step length is set to 1000 m in the range of 10,000–24,000 m. Then, a total of 105 different bandwidths are set, and 1050 corresponding POI density layers (105 times × 10 layers) are generated in sequence. Next, the county-level 10 POI density layer is generated by different bandwidths, respectively. Finally, taking 10 POI county-level sums as the independent variables, and the county-level census as the dependent variable, an RF regression model is built. A total of 105 RF regression models were constructed in this study. The bandwidth corresponding to the minimum out-of-bag error value of the 105th RF model is the optimal bandwidth. 'Out-of-bag' is an unbiased estimate of the RF generalization error [30]. The smaller the out-of-bag value, the smaller the model error. The out-of-bag experienced a process of decreasing first and then increasing (Figure 4). When the bandwidth is 6500, the out-of-bag value is the lowest.

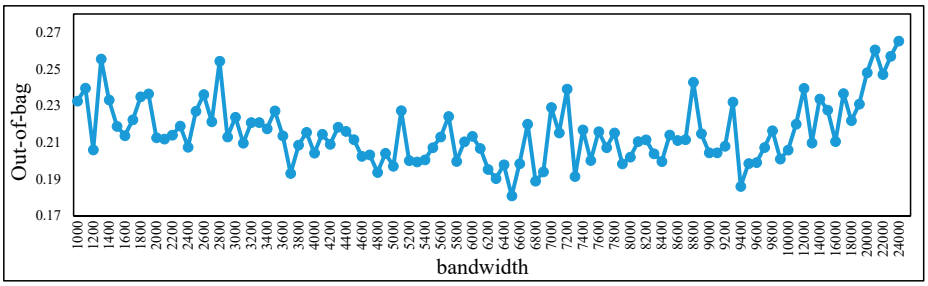

Figure 4. The out-of-bag values at different bandwidths.

Then, the 10 different types of POI density layers need to be combined into one layer. The entropy value can be utilized to calculate the weight of the 10 POI density layers. The entropy method only depends on the data itself and is an objective weighting method [34]. The smaller the entropy value, the greater the weight of the indicator, and vice versa. Therefore, in order to eliminate the difference in the dimension and range between different indicators, this study utilized Equation (4) to normalize the 10 POI density layers, respectively. Then, Equations (5) and (6) were utilized to calculate the entropy value (e_j) and weight value (w_j) of the j index. The POI weight is shown in Table 3. Finally, Equation (7) was used to merge the 10 POI layers into one layer (POI_{sum}).

$$b_{ij} = \frac{a_{ij} - \min\{a_{1j}, \cdots, a_{nj}\}}{\max\{a_{1j}, \cdots, a_{ij}\} - \min\{a_{1j}, \cdots, a_{ij}\}}, \quad (4)$$

$$e_j = -\frac{1}{\ln(n)} \times \sum_{i=1}^{n} \left(\frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}} \right) \times \ln \left(\frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}} \right), \quad (5)$$

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{m} (1 - e_j)}, \quad (6)$$

$$POI_{sum} = \sum_{j=1}^{m} w_j b_{ij}, \quad (7)$$

where b_{ij} is the pixel value of the POI density layer after normalization; a_{ij} is the pixel value of the POI density layer; e_j is the entropy value; w_j is the weight value; POI_{sum} is the merged POI layer; m is the number of layers, j = 1, ..., m, m = 10; and I = 1,..., n, n is the number of pixels in the raster layer, with a total of 2,047,689,100 pixels.

Table 3. The weight of the POI layer.

Categories	Transportation	Government	Village	Education	Food
Weight	0.096	0.103	0.105	0.106	0.094
Categories	**Medical**	**Entertainment**	**Accommodation**	**Working Place**	**Financial**
Weight	0.104	0.098	0.099	0.092	0.105

3.2.4. Calculating the Road Distance Layers

The distance from the center point of each grid to the nearest road of each type was calculated in order to generate seven road distance rasters. Then, to simplify the subsequent calculations, the entropy method was utilized to merge the seven road raster layers into one raster layer, and the process of merging the layers was the same as that in Section 3.2.3. The road data weights are shown in Table 4.

Table 4. The weight of the road layer.

Categories	Railway	National Road	Provincial Road	Highway
Weight	0.140	0.151	0.138	0.139
Categories	**Urban Road**	**County Road**	**Village Road**	
Weight	0.161	0.128	0.143	

3.3. Accuracy Assessment

Indirect verification was utilized as a compromise, since the 100 m × 100 m population density data were inaccessible. Therefore, we compared the Popi data and Worldpop data with the census at the township level. The indicators, including the Relative Error (RE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), were used to measure the accuracy. The formulae are as follows:

$$RE = \frac{P_i - C_i}{C_i}, \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum (P_i - C_i)^2}, \tag{9}$$

$$MAE = \frac{1}{n}\sum |P_i - C_i|, \tag{10}$$

where P_i is the sum of population density products at the township level, and C_i is the township population.

4. Results

4.1. Results of the Population Density Mapping

In order to improve the model's accuracy, the parameters of the RF model need to be optimized. Considering that the random search parameter is time-consuming, we preset the possible value range of the parameter and utilized the out-of-bag value to access the optimal parameters (Figure 5). The range and optimal values are shown in Table 5. The goodness of fit (R^2) of the training set is 0.98, and the R^2 of the test set is 0.88.

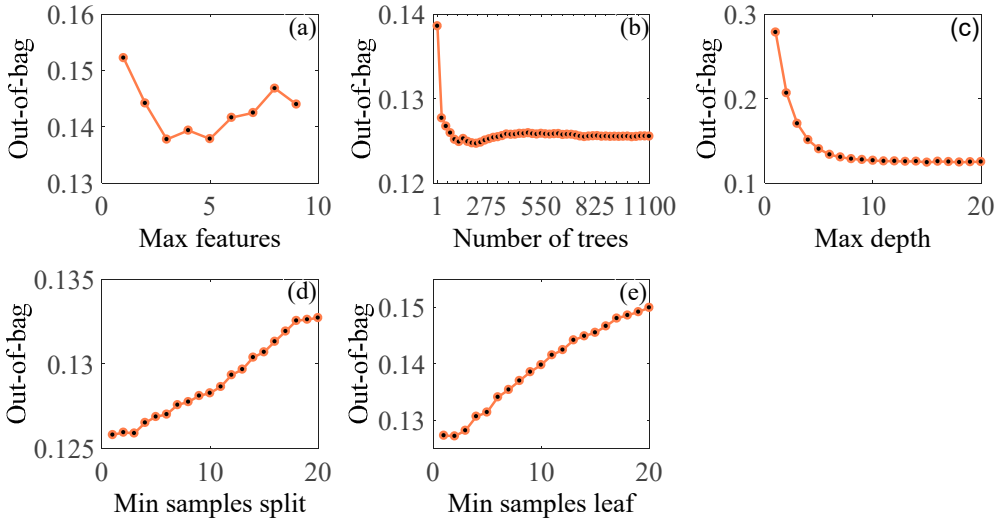

Figure 5. The out-of-bag values with different parameters. (**a**) Max features; (**b**) number of trees; (**c**) max depth; (**d**) min samples split; (**e**) min samples leaf.

Table 5. The parameter range and optimal value.

Parameter Type	Range	Optimal Value
Max. features	1–10	5
Number of trees	1–1100	200
Max. depth	1–20	15
Min. samples split	2–20	2
Min. samples leaf	1–20	1

The population density at the 100 m × 100 m resolution in 2015 is shown in Figure 6a. The population density shows a strong imbalance in mainland China. The population of China is densely distributed in the southeast and sparsely distributed in the northwest. In the east, the population distribution is concentrated. For example, in four of the most densely-populated cities, including Beijing, Shanghai, Guangzhou, and Shenzhen, the population density is more than 1000 people/km^2. In the less densely populated cities, such as Lanzhou, Kunming, Shenyang, and Urumqi, which contain more rural areas compared with the densely populated cities, the population density is 200–300 people/km^2. Due to natural factors, the population density in certain western regions is less than 10 people/km^2.

The spatial distribution of the population is affected by the altitude, landform, vegetation, land cover, transportation, and social economy, etc. The natural conditions of the river valleys, plains, and hills with middle and low altitudes provide superior land conditions for the development of human activities. For example, there are the population is more concentrated in marine plains (such as Shanghai), alluvial plains (such as central Shaanxi), and low-elevation hilly areas (such as Guangzhou). Next, the different combinations of terrain, climate, and vegetation lead to different land cover types. The population will first gather in areas with advantageous natural resource endowments. Since land is the carrier of the population distribution, the population is more likely to be gathered in areas with superior conditions (such as built-up areas). Moreover, the road accessibility affects the level of communication between regions, causing the flow of various resource elements, including the population. Since major traffic arteries become the axis of the population distribution, the population is concentrated in areas with developed transportation (such as Beijing and Shanghai). The intensity of the economic development represented by the level of urbanization redistributes the population density based on resources and natural conditions, forming a 'core–periphery' effect of the population's spatial distribution within the city, reflected in municipality cities, such as Shenzhen, or provincial capitals, including Shenyang and Kunming, etc. (Figure 6b).

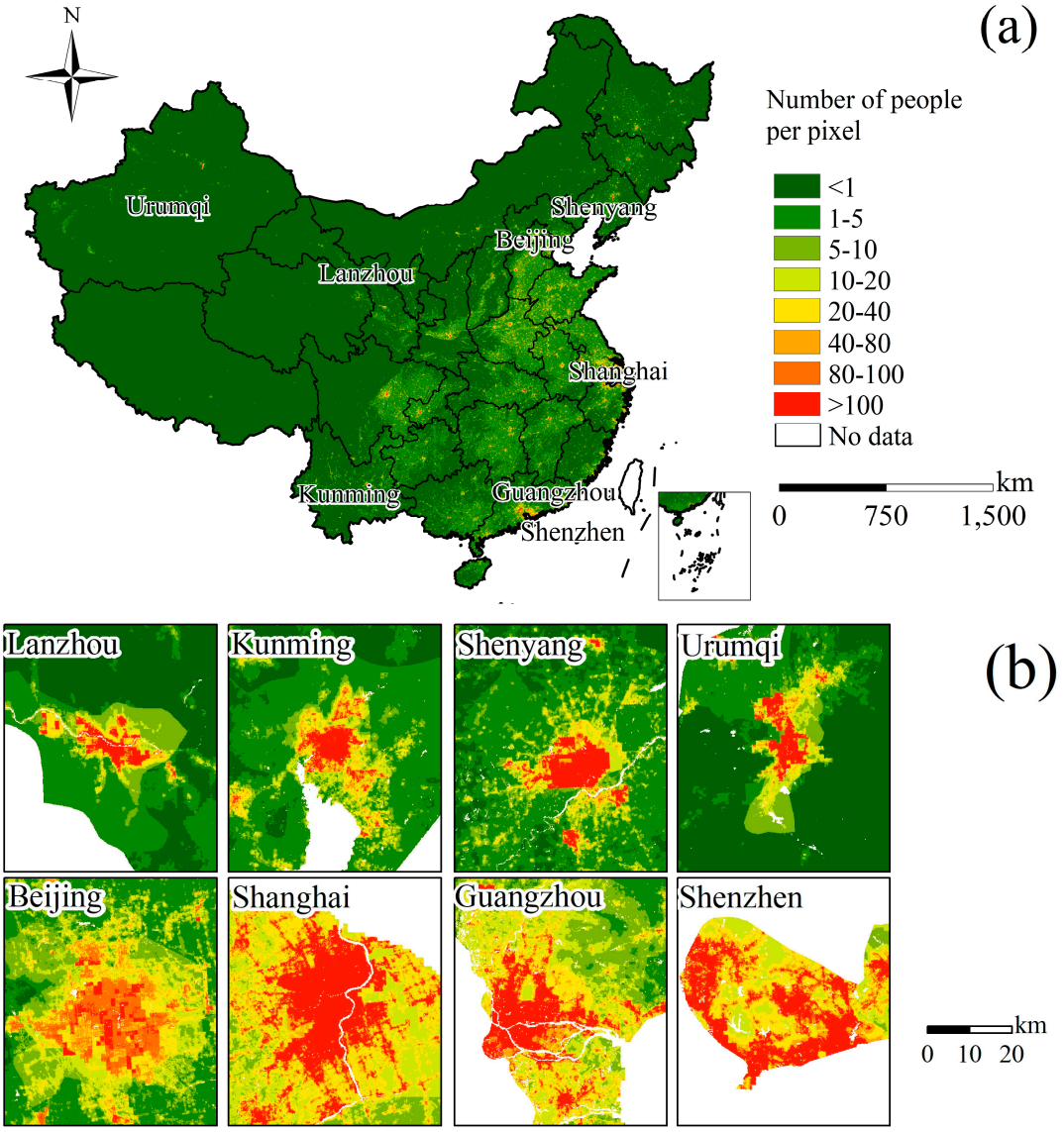

Figure 6. Population density maps with a 100 m × 100 m grid for 2015 in (**a**) mainland China; and (**b**) typical cities.

4.2. Accuracy Assessment

Furthermore, we utilized the 28,505 township censuses to verify the Popi product and compared it with the Worldpop product. The RMSE of the Worldpop and Popi products is 18,014 and 14,839, respectively, and the MAE is 8572 and 7218, respectively. The Worldpop and Popi products have a higher accuracy in mainland China, while the Popi data are slightly better than the Worldpop data. As seen from Figure 7, Popi products can reduce the overestimation and underestimation to a certain extent. For example, Popi products have reduced overestimation in the western provinces (such as southern Xinjiang). Meanwhile, Popi products have reduced underestimation in the central and eastern provinces (such as Shaanxi, Sichuan, Hunan, Anhui, and Zhejiang, etc.).

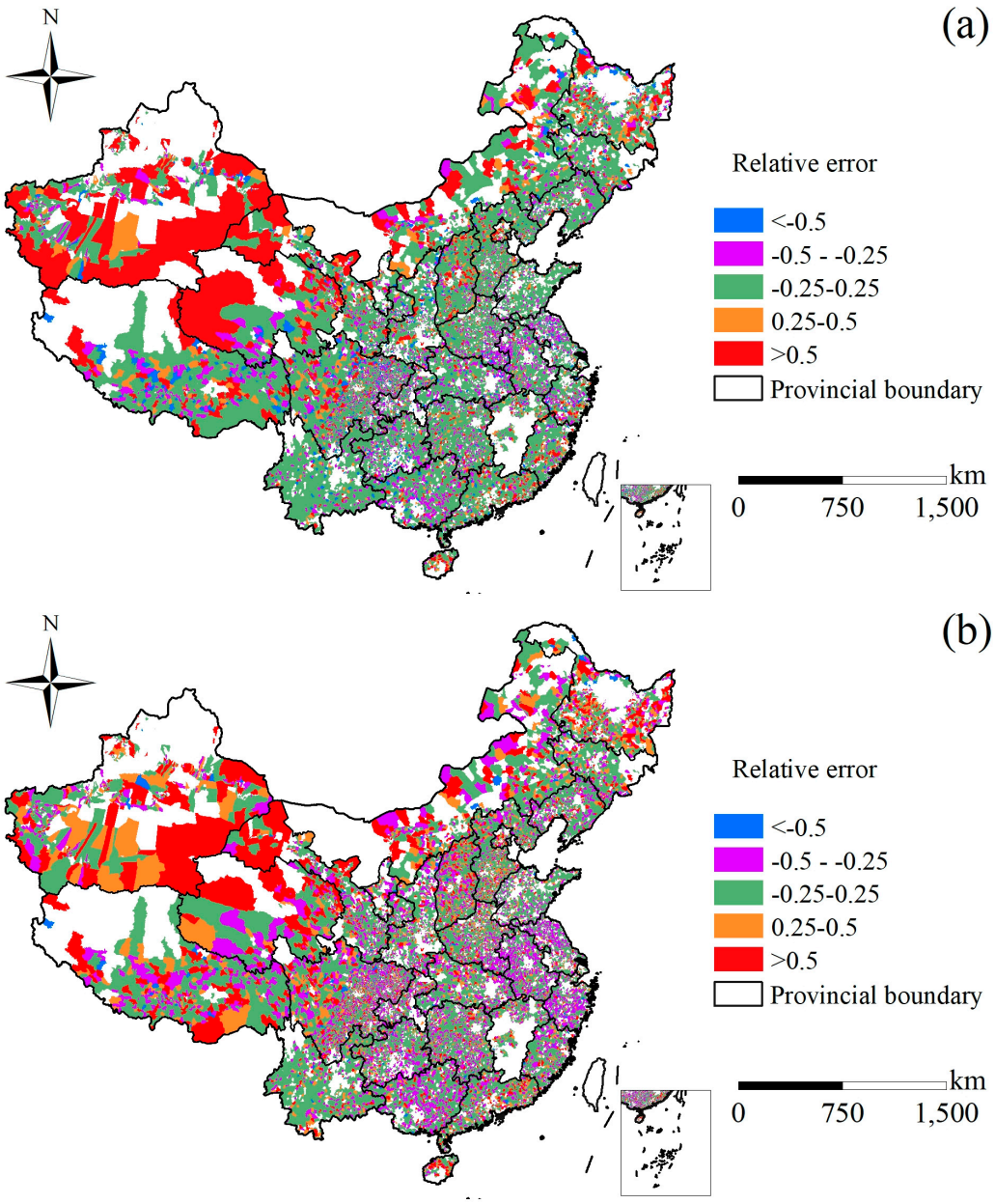

Figure 7. The spatial distribution of the relative error values at the township level for the (**a**) Popi product and (**b**) Worldpop product. The township censuses of Tianjin, Hong Kong, Macau, and Taiwan provinces are missing. Besides this, when RE < −0.5, the township is considered seriously underestimated. A township with −0.5 < RE < −0.25 is considered to be a slight underestimation. Furthermore, −0.25 < RE < 0.25 is an accurate estimate. When 0.25 < RE < 0.5, it is slightly overestimated. A township with RE > 0.5 is considered to be seriously overestimated.

In order to analyze the accuracy of the Popi data, we calculated the RE values of the townships and divided them into five levels (Figure 7). Overall, few townships are significantly overestimated and underestimated in these two products. Among them, the number of townships that are accurately estimated with the Popi product is 19,300 (67.7%), while the number of townships that are accurately estimated with the Worldpop product is 16,237 (56.9%). There are 1051 (3.6%) and 4996 (17.5%) townships in the Popi products that are slightly overestimated and slightly underestimated, respectively. The number of townships that are slightly overestimated in the Worldpop products is 2038 (7.1%), and the number of townships that are slightly underestimated is 5847 (20.5%). Meanwhile, the number of townships seriously overestimated and seriously underestimated in the Popi products are 1572 (5.5%)

and 1586 (5.6%), respectively. The Worldpop product has 2092 (7.3%) seriously overestimated townships, and 2291 (8.1%) seriously underestimated townships. Above all, the sum of the proportions of the severely overestimated or severely underestimated townships is about 10%, indicating that both products have a good accuracy.

Furthermore, the underestimation ratio of the two products exceeds 20%, and is slightly higher than the overestimation ratio (about 10%). For example, there are many underestimations in the central and eastern provinces (such as Jiangsu, Anhui, Sichuan, and Zhejiang), where the total population of the province exceeds 50 million, with a higher density. In contrast, the western provinces (such as Xinjiang, Tibet, Qinghai, and Inner Mongolia) with small populations and low densities have many obvious overestimations for the two products. The township area of the western region is larger, but the township number is small, while the township area of the central and eastern regions is smaller, and the number is larger. Therefore, more townships are underestimated.

In order to analyze the accuracy of the different regions, we calculated the R^2 between the township census and the estimated township population in 30 provinces. As seen from Figure 8, the R^2 value of the Popi product in 22 provinces is higher than that of the Worldpop product, and the R^2 value of the Worldpop product in eight provinces is higher than that of the Popi product.

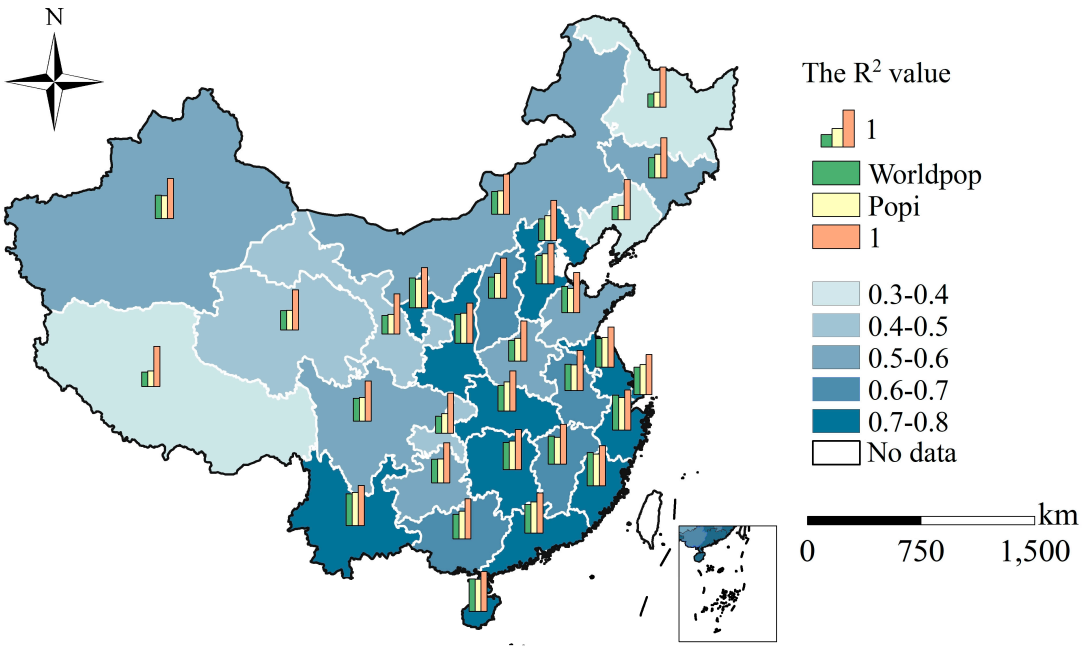

Figure 8. The goodness of fit (R^2) between the township census and the estimated township population of Worldpop and Popi products in 30 provinces (the R^2 values for the Tianjin, Hong Kong, Macau, and Taiwan provinces are missing).

From a regional perspective, the southeast regions, such as Guangdong, Fujian, Zhejiang, and Jiangsu, are mainly distributed in plains and hills with low altitudes, which are suitable for humans and have the highest accuracy ($R^2 > 0.7$). Similarly, the northwest, including Shaanxi and Ningxia, with medium- and low-altitude alluvial plains, is also suitable for humans, and has $R^2 > 0.7$. In the central regions, such as Guangxi and Anhui, the landforms are mainly low-altitude mountains and hills, showing that the natural conditions are suitable, and exhibiting a higher accuracy ($0.7 > R^2 > 0.6$).

Next, the western region, including Xinjiang, Sichuan, and Guizhou, has a medium accuracy ($0.6 > R^2 > 0.5$) and a high altitude, and is dominated by mountains and hills, such as the Tianshan Mountains and the Hengduan Mountains. Henan and Shandong, with a total population exceeding 100 million, have a medium accuracy ($0.6 > R^2 > 0.5$). They are followed by the northwest and southwest regions, including Qinghai, Gansu, and Chongqing, where the landform is occupied by

alluvial platforms, mountains, and denuded platforms, etc., illustrating the unsuitability of natural conditions with a lower accuracy ($0.5 > R^2 > 0.4$). The provinces with the lowest accuracy ($R^2 < 0.4$) are in the southwest and northeast, including Tibet, Heilongjiang, and Liaoning. Tibet is dominated by mountains with high altitudes, such as Tanggula Mountain, demonstrating uninhabitable characteristics. Heilongjiang and Liaoning are dominated by mountains and terraces with a low altitude, but the population is less affected by the low temperature.

Above all, the accuracy is affected by the altitude, landform, temperature, and total population, etc. The provinces with a poor accuracy are located in densely-populated or sparsely-populated areas, or areas with poor natural conditions.

In order to analyze the relationship between the population density of the townships and RE, we divided the population density of the townships into six classifications, as shown in Figure 9. Overall, the RE values of the two products are mostly distributed in the range of −0.25 to 0.25. The distribution of the RE values in the Popi data are more concentrated, with fewer extreme points, and the median of the RE values is closer to 0. The distribution of the RE values in the Worldpop dataset is more scattered, with more extreme values.

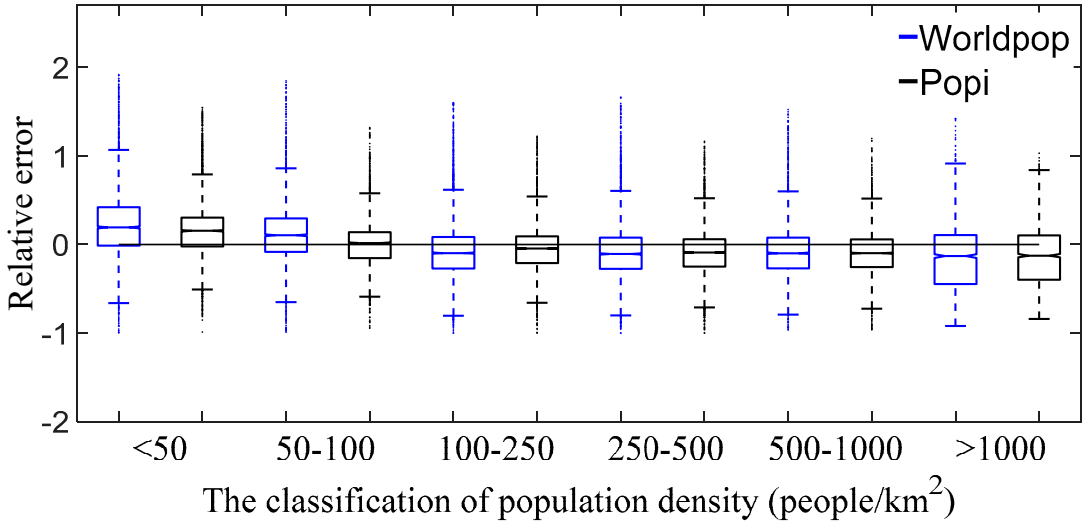

Figure 9. The result of the relative error at the township level (the township censuses of the Tianjin, Hong Kong, Macau, and Taiwan provinces are missing).

There is a certain correlation between the RE value and the population density. We found that there may be an overestimation when the population density is less than 50 people/km^2 and the median of RE in this classification is close to 0.2. There tends to be an underestimation when the population density exceeds 1000 people/km^2 and the median of the RE in this classification is close to −0.15. Moreover, compared with a low population density (less than 50 people/km^2), the median of the RE value with a high population density (exceeds 1000 people/km^2) is closer to 0, with fewer discrete points. Hence, the accuracy of these two products is better in areas with a higher population density than it is in areas with a lower population density. When the population density is in the range of 50–1000 people/km^2, the median value of the RE is close to 0, indicating that both products have good effects.

4.3. The Feature Importance of the Independent Variables

In order to analyze the importance of the independent variables, the feature importance indicator was calculated. The feature importance is the degree of influence on the accuracy of the RF model when an independent variable is replaced by randomly distributed data. If the independent variables do not participate in the model's training, the model's accuracy will decrease. The higher the feature's

importance, the more important the independent variables, and vice versa. The feature importance of the RF model is shown in Figure 10.

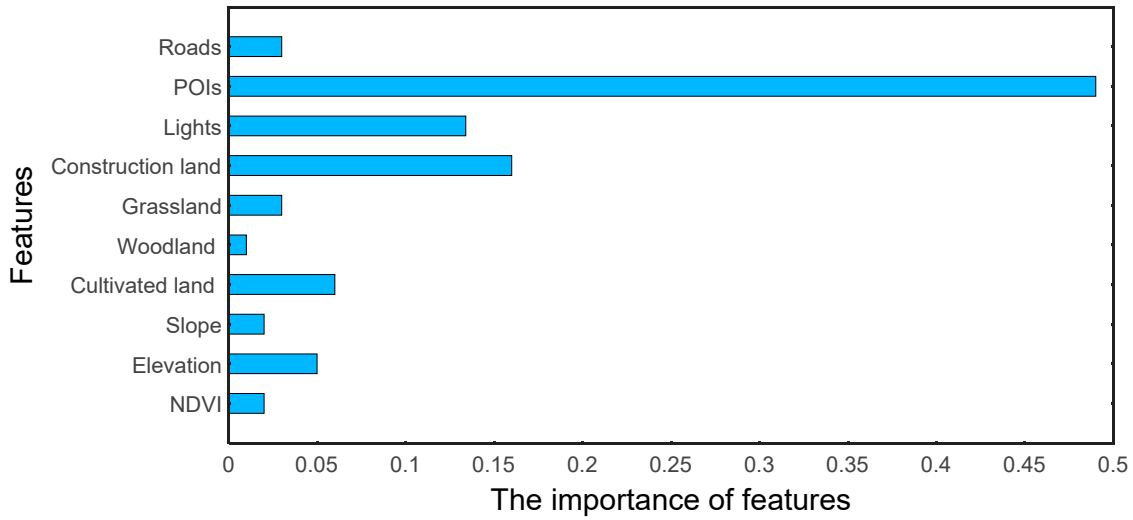

Figure 10. The feature importance of the independent variables.

Compared with the natural factors, such as the NDVI, elevation, slope, and LULC data, the socio-economic factors, including the POI data and NTL, are more important. Among them, the feature importance of the POI data is 0.49. If the POI data does not participate in the model training, the accuracy of the model will drop by approximately half, which indicates that the POI data are the most important. This is followed by the NTL data, where the feature importance of the NPP/VIIRS is 0.14. If neither the POI nor NTL data participate in the RF model training, the model's accuracy decreases by about 63%. Therefore, the method of combining the POI and NPP/VIIRS data is more conducive to the estimation of the population density.

The feature importance value of construction land is 0.16, while the values for cultivated land, forest land, and grassland are all below 0.1. The feature importance of construction land is significantly higher than that of the last three factors. The construction land includes urban construction land, rural construction land, industrial land, and mining land, while the population is usually distributed in built-up areas, rural areas, and workplaces. Only a small number of people are engaged in agriculture or forestry, and are therefore distributed in cultivated land or woodland, etc. Therefore, cultivated land, forest land, and grassland are less important in the spatialization of the population.

The population distribution is limited by natural conditions (NDVI, elevation, and slope) and road factors. Although the feature importance of these factors is less than 0.1, the model's accuracy will decrease when these factors are removed. For example, in Guizhou, as one of the main karst regions in China, Dong et al. [8] confirmed that the altitude, slope, and aspect have a great influence on the population distribution. The feature importance confirms that there is a certain relationship between these natural factors and the population.

5. Discussion

5.1. The Differences between POIs and NPP/VIIRS in Mapping the Population Density

As shown in Figure 10, the feature importance of the POI data is significantly higher than that of the NTL data. In order to analyze the cause of this phenomenon, we divided the NPP/VIIRS data from 2015 into a lighted area and an unlighted area, according to whether the DN value was greater than 0. The lighted area was shown to be only 2,163,658 km^2 (22.5%). Zhuo et al. [15] divided the DMSP/OLS data values within mainland China from 1998 into two types: with and without lights. The area with light only accounts for 503,400 km^2 (5.3%), while the area without light accounts for 8,040,000 km^2

(84.7%). There is also a population distribution in the areas without lights. The NPP/VIIRS and DMSP/OLS data have similar problems.

One limitation of NPP/VIIRS is that the overpass time is later than 12:00 PM, and most of the lights are off at that time. The urban functions differed in their temporal light dynamics [35,36]. For example, Li et al. [36] found that an outdoor sports field and an administrative building lost 97.28% and 4.56% of their measured brightness between 8:08 PM–4:05 AM, respectively, while the entire study area in Wuhan, China, lost 61.86% of its total brightness. Furthermore, the period between 9:06 PM and 10:05 PM was the period with the greatest amount of light loss in the study area.

Furthermore, the lighted areas are mostly distributed in urban areas, and are less covered in rural areas, while the POI data covers urban and rural areas. In particular, the type of village in the POI data can effectively simulate the spatial distribution of the rural population. In 2015, there were 3,567,162 records of villages, which can make up for the lack of NPP/VIIRS in rural areas.

Additionally, there are several differences between the NTL and POI data in the urban built-up areas and the rural areas. We utilized urban construction land and rural construction land in LULC data as mask data, and the NPP/VIIRS data and POI density layer were divided into built-up areas and rural areas by masks, respectively. The four layers (NPP/VIIRS (built-up area), POIs (built-up area), NPP/VIIRS (county), and POIs (county)) were aggregated at the city level, respectively. We took the city census as the dependent variable and the four aggregated values as the independent variable in order to calculate the correlation coefficient and the R^2 of the linear regression model between the independent and dependent variables, respectively (Figure 11).

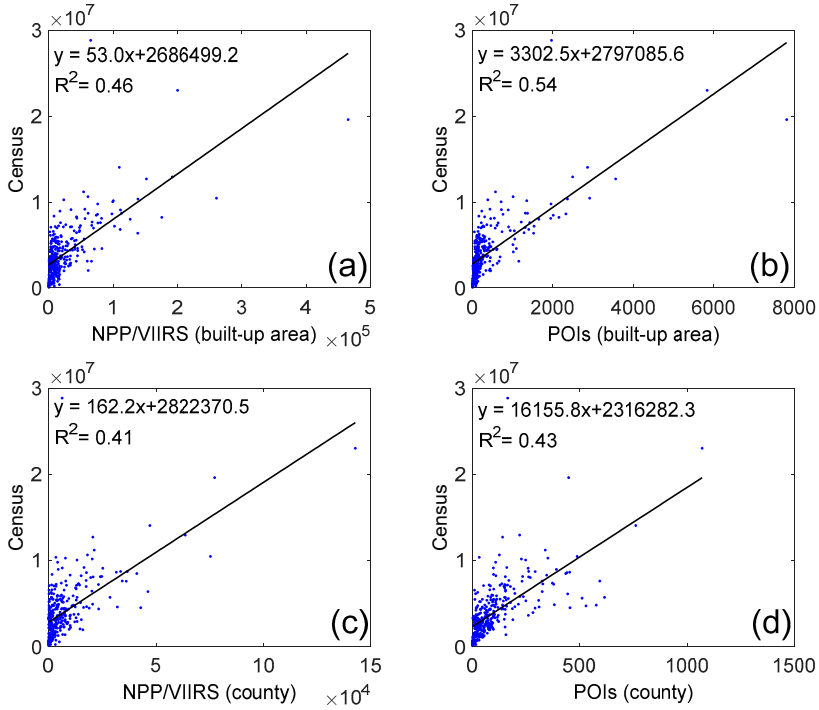

Figure 11. The census data with respect to the National Polar-orbiting Operational Environmental Satellite System Preparatory Project/Visible Infrared Imaging Radiometer (NPP/VIIRS) and point of interest (POI) at the city level (using urban construction land and rural construction land in the 2015 LULC data as masks, the built-up areas and rural areas of 340 cities in mainland China were extracted, respectively). (**a**) The census data with respect to the NPP/VIIRS data in built-up area; (**b**) the census data with respect to the POIs data in built-up area; (**c**) the census data with respect to the NPP/VIIRS data in county; (**d**) the census data with respect to the POIs data in county.

The NPP/VIIRS and POI data have a positive correlation with the city census in urban built-up areas and rural areas. Two independent variables perform better in urban built-up areas than in rural

areas, while the POIs perform better than NPP/VIIRS in the same area. As shown in Figure 11, the city census has the highest correlation with the POIs within the built-up areas; the correlation coefficient is 0.73, and the R^2 is 0.54. Then, the correlation coefficient between the city census and the NPP/VIIRS value within the built-up area is 0.67. Apart from that, the correlation coefficient between the city census and the POIs in the rural area is 0.65, and the R^2 is 0.43. Finally, the correlation coefficient between the city census and the NPP/VIIRS values in the rural area is 0.64, and the R^2 is 0.41.

5.2. The Correlation between POIs and Censuses

Previous studies have proved that there is a correlation between POI and the population. For example, Li et al. [24] and Yang et al. [37] used POI data to estimate the population density, confirming that POI data can be used as a population spatialization modeling factor. Similarly, the feature importance of the POI calculated in this paper is 0.49, and Figure 12a shows that there is a positive correlation between POI data and the total population at the city level. Furthermore, Figure 12l shows that the correlation coefficient between the population density and POI density is 0.83, indicating a positive correlation between the population and POI. This fully illustrates the availability of POI data in the establishment of the model.

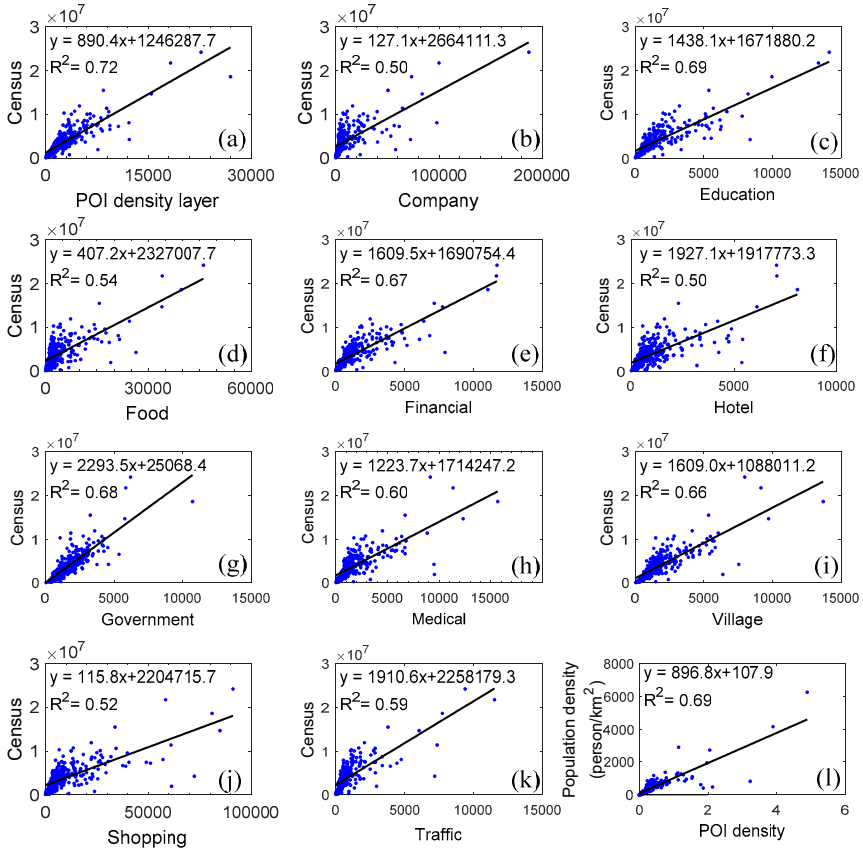

Figure 12. The relationship between the census and different type POI data at the city level. (**a**) POI density layer (the POI density layer is the combination of the 10 individual POI layers using the entropy weight method); (**b**) the company type POI data; (**c**) the education type POI data; (**d**) the food type POI data; (**e**) the financial type POI data; (**f**) the hotel type POI data; (**g**) the government type POI data; (**h**) the medical type POI data; (**i**) the village type POI data; (**j**) the shopping type POI data; (**k**) the traffic type POI data; (**l**) the POI density data.

In addition, there are many types of POI data, but not every type is related to the population distribution. For example, airports and railway stations, etc., are usually far away from urban (higher population density) areas, and have no obvious correlation with population density. The POI

types selected in the article, including education, factories, finance, and villages, are more closely related to the population.

In order to explore the relationship between censuses and multiple POIs, we utilized the city administrative boundary data to aggregate the POI density layer and the 10 types of POI layers, respectively. Then, we took the city census as the dependent variable and the aggregated value at the city level as the independent variable, and the correlation coefficient and the R^2 were calculated, respectively (Figure 12).

There is a positive and strong correlation between the census and the POI density layer [38]. The correlation coefficient between the aggregated value of the POI density layer and the city census is 0.85, and the R^2 of the linear regression model is 0.72, which is significantly higher than the values for the other 10 POIs. Therefore, the method of integrating multiple POI layers into one POI density layer can effectively reduce the calculation, and can form the distribution of the population under the combined action of multiple POI types.

The 10 POI data are positively correlated with the population, and the correlation coefficient is greater than 0.7. The population has the highest correlation with four types of POI: education, government, finance, and villages. Specifically, the correlation coefficients between the city censuses and education, government, finance, and villages are 0.83, 0.82, 0.81, and 0.81, respectively. These four types of POI data are closely related to human life. The correlation coefficients between the city censuses and medical services and transportation are 0.77 and 0.76, respectively, indicating that medical services and transportation are indispensable in people's daily lives. Moreover, the correlation coefficients between the city census and catering, entertainment, hotels, and workplaces are 0.73, 0.72, 0.7, and 0.7, respectively.

5.3. Error Analysis

These two products performed well in most areas of mainland China. However, due to the large differences in the natural resources and socioeconomic conditions in different regions, certain errors remain.

The Popi dataset is better than the Worldpop dataset for the following reasons: both datasets employ the RF model and similar modeling factors, including natural factors (elevation, slope, LULC, etc.) and socioeconomic factors (NTL, roads, etc.). It can be seen from Section 4.3 that the feature importance of POI and NTL is significantly higher than for other natural factors. The POI data applied in Worldpop are derived from the Open Street Map (OSM) dataset. The OSM data are an online map collaboration program. Any registered user can edit the map content. The lack of professional editing may lead to inaccurate positioning or classification, which reduces the credibility of the data [39]. Compared with OSM data, POIs are produced and edited by professionals, providing more accurate location information and attribute information. Especially in urban areas, POI data can accurately describe the relevant information of humanity, reflect the spatial heterogeneity, and improve the accuracy of the population spatial modeling.

In addition, the data sources used by Worldpop have a low resolution and poor timeliness. For example, the NTL data used by Worldpop is DMSP/OLS data, with a resolution of 1 km, which has a pixel saturation effect [16]. The resolutions of the LULC and NTL data used in this paper are 30 and 750 m, respectively, which are significantly better than those given in Worldpop. The LULC and NTL data of Worldpop use 2009 data and 1992–2013 time-series data, respectively [39]. Due to timeliness, there are certain shortcomings in the estimation of the spatial distribution of the population in 2015.

When the population density is higher (>1000 people/km^2) or lower (<50 people/km^2), more errors are likely to occur. Furthermore, the accuracy of these two products in low-density areas is not as high as it is in high-density areas. Areas with a higher population density include developed urban areas, such as Beijing and Shanghai, while areas with a lower population density may be found in rural regions. Gaughan et al. [6] believe that the modeling process does not concentrate people enough in densely-populated urban areas, and the estimates are scattered to less-populated areas. This is inherent in the dasymetric approach used in the mapping population redistribution, but it affects relatively few

total census units. Additionally, the greater the city population, the richer the types and quantities of the POI data [26]. The rural population is relatively scattered, and the number of POIs in rural areas is small and scattered [40]. Next, the NTL in the urban area with a DN value greater than 0 is higher than that in rural areas, and the NTL information is more abundant in the city [15]. Furthermore, the urban landforms are relatively simple, mostly dominated by plains and basins, and the terrain is relatively flat. The rural landforms may be hills or mountains, etc. These complex topographies and landforms affect the results of the population spatialization. For example, Bai et al. [9] found that the Worldpop dataset has considerable errors in hilly areas, such as the Hengduan Mountains. Besides this, due to the sparse distribution of the cultivated land and residential areas in rural areas, it is difficult to achieve an accurate classification of the land cover data in rural areas, which is not conducive to the spatial modeling of the population.

Overall, the quality and appropriateness of the modeling factors will affect the accuracy of population density data. For example, the Twitter data used by our predecessors in population density mapping has achieved better results in Indonesia; however, the use of Twitter is not widespread in China [41]. In contrast, as the accuracy of the location and attributes of the POI data is reliable, and it is optimal to utilize the socioeconomic factors in the quantification of the population density in China. Furthermore, the anisotropic characteristic of the artificial light at night may also impact the NPP/VIIRS data [42]. The Luojia-1 satellite provided a new data source for nighttime light applications, which could be a future direction of study. In addition, the modeling method also affects accuracy of the population density data. For example, Yang et al. [37] used POI data to calculate the population density of Zhejiang Province, China, and applied the linear regression equation to build the model. The linear model needs to assume that there is a linear relationship between the population and the modeling factors. However, the Enhanced Vegetation Index (EVI) used by Yang et al. [37] was not shown to have to have a linear relationship with the population density.

Finally, although the accuracy of the Popi product is slightly higher than that of Worldpop in this research area, there are few shortcomings in this article. For example, the LULC data used in this paper is highly accurate, but is not open source data. Secondly, the Baidu POI data can only be used in China, and not the rest of the world. It is still difficult to promote to the world due to the lack of Baidu POI data. In addition, the coverage of the OSM data used by the Worldpop data in China has not been completed yet [43]. Moreover, OSM, as a volunteered geographic information (VGI)-based dataset, may have incorrect locations or attributes that will affect the accuracy of the population density products. However, Worldpop products also have their own advantages. This product is a global high-precision population density product, and is currently the mainstream population density product.

6. Conclusions

Taking the LULC, NDVI, NPP/VIIRS, POI, road, and DEM data as the independent variables, the RF regression model was applied to disaggregate the 2015 county-level census population, in order to map the population density in a 100 m × 100 m grid in mainland China. The population density map that was produced showed a higher accuracy than the Worldpop dataset. This study demonstrated that the utilization of multi-source data can effectively improve the accuracy of population mapping at a finer scale.

The combination of POIs and NPP/VIIRS data is more conducive to the estimation of the population density. The sum of the feature importance of the socioeconomic factors, including POI and NPP/VIIRS, exceeds 60%, which is significantly higher than the other natural factors. Without the POIs and NPP/VIIRS in the RF model training, the model accuracy will decrease by about 63%.

The accuracy of the model is affected by the altitude, landform, temperature, and total population, etc. The areas with a poor accuracy were mainly located in the densely-populated and sparsely-populated areas. Obvious overestimation and underestimation appeared in areas with a population density of less than 50 people/km^2 and more than 1000 people/km^2, respectively.

The method of population density mapping based on remote sensing and POI data utilized in this paper has a certain practical significance. However, POIs are concentrated in urban areas, and are less concentrated in rural areas (except for rural settlement types). In our future work, we will introduce socioeconomic data, such as building data [25], Tencent location data [44], and Twitter [45], etc. The combination of POIs and NTL with other socioeconomic data sources may improve the accuracy.

Author Contributions: C.H. and Y.W. designed the experiment; Y.W. and M.Z. collected the required data and performed the experiment; Y.W., C.H., M.Z., J.H., Y.Z. and J.G. contributed towards the writing and review. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number XDA19040500).

Acknowledgments: The authors thank the anonymous reviewers for the helpful comments that improved this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lu, D.; Tian, H.; Zhou, G.; Ge, H. Regional mapping of human settlements in southeastern China with multisensor remotely sensed data. *Remote Sens. Environ.* **2008**, *112*, 3668–3679. [CrossRef]
2. Ferguson, N.M.; Cummings, D.A.; Cauchemez, S.; Fraser, C.; Riley, S.; Meeyai, A.; Iamsirithaworn, S.; Burke, D.S. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **2005**, *437*, 209–214. [CrossRef] [PubMed]
3. Storeygard, A.; Balk, D.; Levy, M.; Deane, G. The Global Distribution of Infant Mortality: A subnational spatial view. *Popul. Space Place* **2008**, *14*, 209–229. [CrossRef]
4. Balk, D.; Storeygard, A.; Levy, M.; Gaskell, J.; Sharma, M.; Flor, R. Child hunger in the developing world: An analysis of environmental and social correlates. *Food Policy* **2005**, *30*, 584–611. [CrossRef]
5. Wang, Y.; Huang, C.; Feng, Y.; Zhao, M.; Gu, J. Using Earth Observation for Monitoring SDG 11.3.1-Ratio of Land Consumption Rate to Population Growth Rate in Mainland China. *Remote Sens.* **2020**, *12*, 357. [CrossRef]
6. Gaughan, A.E.; Stevens, F.R.; Huang, Z.; Nieves, J.J.; Sorichetta, A.; Lai, S.; Ye, X.; Linard, C.; Hornby, G.M.; Hay, S.I.; et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Sci. Data* **2016**, *3*, 160005. [CrossRef] [PubMed]
7. Openshaw, S. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Sciences*; Wrigley, N., Ed.; Pion: London, UK, 1979; pp. 127–144.
8. Dong, C.; Liu, J.; Zhao, R.; Wang, G. An discussion on correlation of geographical parameter with spatial population distribution. *Remote Sens. Inform.* **2002**, *4*, 61–64.
9. Bai, Z.; Wang, J.; Wang, M.; Gao, M.; Sun, J. Accuracy Assessment of Multi-Source Gridded Population Distribution Datasets in China. *Sustainability* **2018**, *10*, 1363. [CrossRef]
10. Deichmann, U.; Balk, D.; Yetman, G. Transforming Population Data for Interdisciplinary Usages: From Census to Grid. In *Population Health Metrics*; Center for International Earth Science Information Network: Washington, DC, USA, 2001.
11. Balk, D.L.; Deichmann, U.; Yetman, G.; Pozzi, F.; Hay, S.I.; Nelson, A. Determining Global Population Distribution: Methods, Applications and Data. In *Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications*; Academic Press: London, UK, 2006; pp. 119–156.
12. Dobson, J.E.; Bright, E.A.; Coleman, P.R.; Durfee, R.C.; Worley, B.A. LandScan: A Global Population Database for Estimating Populations at Risk. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 849–857.
13. Jiang, D.; Yang, X.; Wang, N.; Liu, H. Study on spatial distribution of population based on remote sensing and GIS. *Adv. Earth Sci.* **2002**, *17*, 734–738.
14. Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE* **2015**, *10*, e0107042. [CrossRef]
15. Zhuo, L.; Ichinose, T.; Zheng, J.; Chen, J.; Shi, P.J.; Li, X. Modelling the population density of China at the pixel level based on DMSP/OLS non-radiance-calibrated night-time light images. *Int. J. Remote Sens.* **2009**, *30*, 1003–1018. [CrossRef]

16. Hsu, F.; Baugh, K.; Ghosh, T.; Zhizhin, M.; Elvidge, C. DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration. *Remote Sens.* **2015**, *7*, 1855–1876. [CrossRef]
17. Zhao, J.; Ji, G.; Yue, Y.; Lai, Z.; Chen, Y.; Yang, D.; Yang, X.; Wang, Z. Spatio-temporal dynamics of urban residential CO_2 emissions and their driving forces in China using the integrated two nighttime light datasets. *Appl. Energy* **2019**, *235*, 612–624. [CrossRef]
18. Wang, X.; Sutton, P.C.; Qi, B. Global Mapping of GDP at 1 km^2 Using VIIRS Nighttime Satellite Imagery. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 580. [CrossRef]
19. Chen, X. Nighttime Lights and Population Migration: Revisiting Classic Demographic Perspectives with an Analysis of Recent European Data. *Remote Sens.* **2020**, *12*, 169. [CrossRef]
20. Elvidge, C.D.; Sutton, P.C.; Ghosh, T.; Tuttle, B.T.; Baugh, K.E.; Bhaduri, B.; Bright, E. A global poverty map derived from satellite data. *Comput. Geosci.* **2009**, *35*, 1652–1660. [CrossRef]
21. Yu, B.; Lian, T.; Huang, Y.; Yao, S.; Ye, X.; Chen, Z.; Yang, C.; Wu, J. Integration of nighttime light remote sensing images and taxi GPS tracking data for population surface enhancement. *Int. J. Geogr. Inf. Sci.* **2018**, *33*, 687–706. [CrossRef]
22. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
23. Bakillah, M.; Liang, S.; Mobasheri, A.; Arsanjani, J.J.; Zipf, A. Fine-resolution population mapping using OpenStreetMap points-of-interest. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1940–1963. [CrossRef]
24. Li, K.; Chen, Y.; Li, Y. The Random Forest-Based Method of Fine-Resolution Population Spatialization by Using the International Space Station Nighttime Photography and Social Sensing Data. *Remote Sens.* **2018**, *10*, 1650. [CrossRef]
25. Yao, Y.; Liu, X.; Li, X.; Zhang, J.; Liang, Z.; Mai, K.; Zhang, Y. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1220–1244. [CrossRef]
26. Wang, L.; Fan, H.; Wang, Y. Improving population mapping using Luojia 1-01 nighttime light image and location-based social media data. *Sci. Total Environ.* **2020**, *730*, 139–148. [CrossRef]
27. Sutton, P.C. Modeling population density with night-time satellite imagery and GIS. *Comput. Environ. Urban Syst.* **1997**, *21*, 227–244. [CrossRef]
28. Wang, L.; Wang, S.; Zhou, Y.; Liu, W.; Hou, Y.; Zhu, J.; Wang, F. Mapping population density in China between 1990 and 2010 using remote sensing. *Remote Sens. Environ.* **2018**, *210*, 269–281. [CrossRef]
29. Wang, L.; Fan, H.; Wang, Y. Fine-Resolution Population Mapping from International Space Station Nighttime Photography and Multisource Social Sensing Data Based on Similarity Matching. *Remote Sens.* **2019**, *11*, 1900. [CrossRef]
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Liu, J.; Kuang, W.; Zhang, Z.; Xu, X.; Qin, Y.; Ning, J.; Zhou, W.; Zhang, S.; Li, R.; Yan, C.; et al. Spatiotemporal characteristics, patterns, and causes of land-use changes in China since the late 1980s. *J. Geogr. Sci.* **2014**, *24*, 195–210. [CrossRef]
32. Liu, J.; Liu, M.; Tian, H.; Zhuang, D.; Zhang, Z.; Zhang, W.; Tang, X.; Deng, X. Spatial and temporal patterns of China's cropland during 1990–2000: An analysis based on Landsat TM data. *Remote Sens. Environ.* **2005**, *98*, 442–456. [CrossRef]
33. Shi, K.; Yu, B.; Huang, Y.; Hu, Y.; Yin, B.; Chen, Z.; Chen, L.; Wu, J. 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.* **2014**, *6*, 1705–1724. [CrossRef]
34. Sun, M.; Wang, T.; Xu, X.; Zhang, L.; Li, J.; Shi, Y. Ecological risk assessment of soil cadmium in China's coastal economic development zone: A meta-analysis. *Ecosyst. Health Sustain.* **2020**, *6*, 1733921. [CrossRef]
35. Li, X.; Levin, N.; Xie, J.; Li, D. Monitoring hourly night-time light by an unmanned aerial vehicle and its implications to satellite remote sensing. *Remote Sens. Environ.* **2020**, *247*, 111942. [CrossRef]
36. Dobler, G.; Ghandehari, M.; Koonin, S.E.; Nazari, R.; Patrinos, A.; Sharma, M.S.; Tafvizi, A.; Vo, H.T.; Wurtele, J.S. Dynamics of the urban lightscape. *Inform. Syst.* **2015**, *54*, 115–126. [CrossRef]
37. Yang, X.; Ye, T.; Zhao, N.; Chen, Q.; Yue, W.; Qi, J.; Zeng, B.; Jia, P. Population Mapping with Multisensor Remote Sensing Images and Point-Of-Interest Data. *Remote Sens.* **2019**, *11*, 574. [CrossRef]
38. Chun, J.; Zhang, X.; Huang, J.; Zhang, P. A Gridding Method of Redistributing Population Based on POIs. *Geogr. Geo-Inf. Sci.* **2018**, *34*, 83–89. (In Chinese)

39. Lloyd, C.T.; Sorichetta, A.; Tatem, A.J. High resolution global gridded data for use in population studies. *Sci. Data* **2017**, *4*, 170001. [CrossRef] [PubMed]
40. Zhao Xin, S.Y.; Liu, Y.; Chen, F.; Hu, Y. Population Spatialization Based on Satellite Remote Sensing and POI Data: Guangzhou as an Example. *Trop. Geogr.* **2020**, *40*, 101–109.
41. Patel, N.N.; Stevens, F.R.; Huang, Z.; Gaughan, A.E.; Elyazar, I.; Tatem, A.J. Improving Large Area Population Mapping Using Geotweet Densities. *Trans. GIS* **2016**, *21*, 317–331. [CrossRef]
42. Li, X.; Ma, R.; Zhang, Q.; Li, D.; Liu, S.; He, T.; Zhao, L. Anisotropic characteristic of artificial light at night—Systematic investigation with VIIRS DNB multi-temporal observations. *Remote Sens. Environ.* **2019**, *233*, 111357. [CrossRef]
43. Tian, Y.; Zhou, Q.; Fu, X. An Analysis of the Evolution, Completeness and Spatial Patterns of OpenStreetMap Building Data in China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 35. [CrossRef]
44. Liu, Y.; Zhu, A.X.; Wang, J.; Li, W.; Hu, G.; Hu, Y. Land-use decision support in brownfield redevelopment for urban renewal based on crowdsourced data and a presence-and-background learning (PBL) method. *Land Use Policy* **2019**, *88*, 104188. [CrossRef]
45. Jiang, Y.; Li, Z.; Ye, X. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartogr. Geogr. Inf. Sci.* **2018**, *46*, 228–242. [CrossRef]