remote sensing

MDPI

Article

Local Population Mapping Using a Random Forest Model Based on Remote and Social Sensing Data: A Case Study in Zhengzhou, China

Ge Qiu [1,2], Yuhai Bao [1], Xuchao Yang [3,4], Chen Wang [5,6], Tingting Ye [3], Alfred Stein [2] and Peng Jia [2,7,8,*]

1 College of Geographic Science, Inner Mongolia Normal University, Huhhot 010022, China; qiugehhht@hotmail.com (G.Q.); baoyuhai@imnu.edu.cn (Y.B.)
2 Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500 Enschede, The Netherlands; a.stein@utwente.nl
3 Ocean College, Zhejiang University, Zhoushan 316021, China; yangxuchao@zju.edu.cn (X.Y.); tingting.ye@zju.edu.cn (T.Y.)
4 Center for Global Change and Earth Observations, Michigan State University, East Lansing, MI 48824, USA
5 Satellite Application Center for Ecology and Environment, Ministry of Ecology and Environment, Beijing 100094, China; wangc@secmep.cn
6 State Environmental Protection Key Laboratory of Satellite Remote Sensing, Beijing 100094, China
7 Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China
8 International Initiative on Spatial Lifecourse Epidemiology (ISLE), Hong Kong, China
* Correspondence: jiapengff@hotmail.com; Tel.: +31-53489-4038

Received: 1 April 2020; Accepted: 14 May 2020; Published: 19 May 2020

Abstract: High-resolution gridded population data are important for understanding and responding to many socioeconomic and environmental problems. Local estimates of the population allow officials and researchers to make a better local planning (e.g., optimizing public services and facilities). This study used a random forest algorithm, on the basis of remote sensing (i.e., satellite imagery) and social sensing data (i.e., point-of-interest and building footprint), to disaggregate census population data for the five municipal districts of Zhengzhou city, China, onto 100 × 100 m grid cells. We used a statistical tool to detect areas with an abnormal population density; e.g., areas containing many empty houses or houses rented by more people than allowed, and conducted field work to validate our findings. Results showed that some categories of points-of-interest, such as residential communities, parking lots, banks, and government buildings were the most important contributing elements in modeling the spatial distribution of the residential population in Zhengzhou City. The exclusion of areas with an abnormal population density from model training and dasymetric mapping increased the accuracy of population estimates in other areas with a more common population density. We compared our product with three widely used gridded population products: Worldpop, the Gridded Population of the World, and the 1-km Grid Population Dataset of China. The relative accuracy of our modeling approach was higher than that of those three products in the five municipal districts of Zhengzhou. This study demonstrated potential for the combination of remote and social sensing data to more accurately estimate the population density in urban areas, with minimum disturbance from the abnormal population density.

Keywords: population distribution; random forest; remote sensing; social sensing; point-of-interest; building footprint

1. Introduction

Up-to-date, spatially accurate population datasets are fundamental to many aspects of decision making and risk assessment, such as economic development, disaster response, and public health research [1–4]. For instance, efficient economic policies require insights of the population distribution, as population gathering and economic growth reinforce each other in several ways (e.g., economic growth increases jobs and infrastructures, and vice versa) [5]; disaster response requires assessing the potentially affected population in given areas for resource allocation [6,7]; and public health research relies on population data as denominator to calculate disease prevalence [8,9]. The population data used to be represented by choropleth maps, which presented aggregated population counts over administrative units [2,10,11]. However, such maps have relatively limited spatial details, and thus cannot meet the growing demand for high-resolution population estimations, partly due to the Modifiable Areal Unit Problem (MAUP) [12].

Gridded population datasets, created by a number of approaches, including areal weighting interpolation [13], pycnophylactic interpolation [14], and dasymetric mapping [10] can depict population distribution more accurately and can be integrated with other geospatial datasets, including remote sensing data [15,16]. There are some well-known global efforts that generate high-resolution gridded population data using these approaches, including Gridded Population of the World (GPW) [17], Global Rural-Urban Mapping Project (GRUMP) [18], LandScan Global [3], Global Human Settlement Population Grid datasets (GHS-POP) [19], and Worldpop [20]. Most of these products use a combination of various Remote Sensing (RS) data as ancillary data, including Land Use/Land-Cover (LULC), Nighttime Light (NTL), temperature, precipitation, etc., to transform traditional choropleth population maps into a continuous gridded population surface, with aggregated values re-distributed across regular spatial units [2,10,12,21]. The population density is highly correlated with the environmental features extracted from these RS data [22,23]. However, in complex urban environments, some RS data at medium spatial resolution have limitations in reflecting explicit environmental information and have difficulty in extracting socioeconomic features, which also relates to human population distribution [24]. For instance, the coarse spatial resolution and blooming effect makes NTL data lack finer scale information, and therefore leads to an over- or under-estimation of the population [25]. Additionally, LULC data, although widely used in population mapping, can only provide limited weight factors based on the number of LULC classes [26]. These problems could lead to a relatively low accuracy in some areas. Ancillary data with better quality should be used for population modeling, such as building footprint, which generally performs better than land cover and settlement data with a coarse-resolution [27,28]. Moreover, social sensing data reflecting socioeconomic features with (x,y) spatial coordinates, such as the categories of Point of Interest (POI), have been proven to be strongly correlated with the small-area population distribution [29,30]. These data are becoming increasingly available and hence should be utilized for a more precise estimation of the local population. A high dimension (i.e., made up of many variables) is an important feature of these data, and thus advanced methods that can handle such high dimensional data should be adopted accordingly.

Random Forest (RF), a machine learning algorithm that can process high dimensional datasets with a relatively high reliability and low time complicity, has been successfully used in many aspects, including the classification of land cover and urban buildings, as well as the classification of insect defoliation levels for the purpose of disease control and prevention [31]. It is a randomly constructed classifier containing multiple decision trees, which are created through bootstrap samples and a user-defined number of features. The output of RF is the average value or the result of a majority vote of these decision trees based on whether the input data is continuous or nominal [32]. Furthermore, the generalization capability of the model built in RF is strong. Several studies have successfully used RF to estimate the high-resolution population distribution. Worldpop [33] used RF and ancillary data derived from RS and geographic information systems (GIS) to redistribute census data in a 100 × 100 m spatial resolution in Asia, Africa, and South America [20]. Tan, et al. [34] used RF and five categories of ancillary data (i.e., NTL, roads, waterbodies, built-up areas, and Digital Elevation Model (DEM))

to disaggregate census data in the Zhujiang delta area, China. With the increase of both type and amount of ancillary data, machine learning methods such as RF are much more needed to handle these high-dimensional, high-resolution data.

Another important but usually neglected issue in population modeling is that population distribution may be affected by unobserved reasons. In some cases, the fundamental assumption that population density should be similar in similar types of parcels (a type is usually a function of many modelable factors) may be broken by some abnormal and non-modelable reasons, such as vacant houses and houses being rented by more people than allowed. Such issues are difficult to avoid in large-scale population modeling projects, but could and should be avoided in local population modeling activities, as detecting and investigating areas with an abnormal population density is also an important and useful practice in city planning and management. However, to the knowledge of the authors, this issue has not been considered in previous efforts of population modeling.

In this study, we employed RF as a semi-automated dasymetric modeling approach, combining a variety of ancillary data to redistribute census data from the county level (also named "district level" in urban regions) to 100 m grid cells in Zhengzhou, China. We assessed the performance of the building footprint and the POI data in population modeling in our study area, and used a statistical tool to detect areas with an abnormal population density, which was validated by our field work. Such areas were then excluded from model training and dasymetric mapping to increase the accuracy of population estimates in other areas with a more common population density. The accuracy of our modeling approach was assessed by comparing it with three major gridded population products that cover Zhengzhou: Worldpop, the Gridded Population of the World (GPW), and the 1-km Grid Population Dataset of China (GPC) [35]. Zhengzhou is geographically at a pivotal location in China, and, as of 2018, has been ranked 2nd among all Chinese cities regarding the population density in urban areas. The city has also recently been officially named the 8th central city; after Beijing, Shanghai, Guangzhou, Chongqing, Tianjin, Wuhan, and Chengdu. Therefore, there is an urgent demand for establishing a benchmark of high-resolution population distribution, so population allocation could be better considered in the planning of this rapidly growing city. This study would significantly contribute to both methodology and practice of urban planning.

2. Methods

2.1. Study Area

Zhengzhou is the capital of Henan province, a key economic region in the central plains and a crucial transportation junction in China. Extending from about 112°42′ to 114°14′E and 34°16′ to 34°58′N, it has a total area of 7446 km^2 and administers five municipal districts: Zhongyuan, Erqi, Guancheng, Jinshui, and Huiji. They further consist of 66 township-level administrative units, hereafter referred to as *Jiedao* (Figure 1).

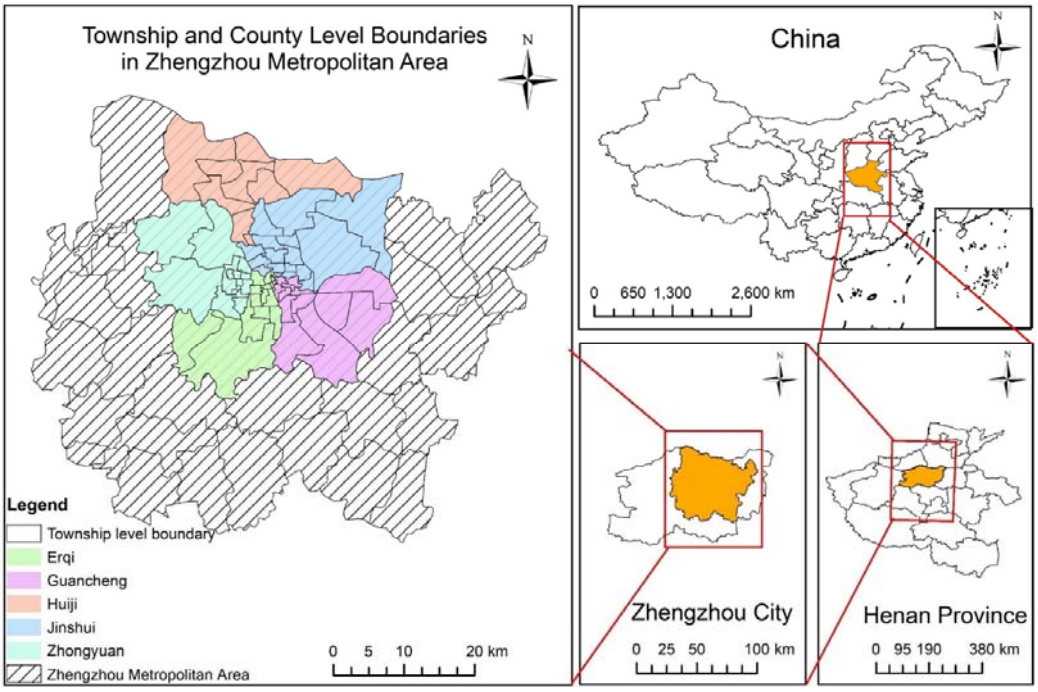

Figure 1. Map of the Zhengzhou metropolitan area in the context of Zhengzhou, Henan and China.

2.2. Datasets

The census of 2010 at county/district and township levels (equivalent to level 3 and 4 of the Global Administrative Unit Layer) was obtained from the China National Bureau of Statistics [36] as the population reference, for the purpose of population redistribution and accuracy assessment, respectively. The county/district- and town-level administrative boundaries were obtained from the Henan Administration of Surveying Mapping and Geoinformation. The relative accuracy of our gridded population product was compared to the relative accuracy of three population datasets: Worldpop (mainland China dataset), the GPW version 4, and the GPC. Specifically, WorldPop (https://www.worldpop.org/) used county level census data for population redistribution and provides gridded population products with the finest spatial resolution (i.e., 100 m) in China. GPW version 4 is presently the most widely used gridded population product, which employed township level census data as population input [17]. The GPC [35] held more local perspectives introduced by the Chinese Academy of Sciences and used county level census data for dasymetric mapping. The spatial resolution of both GPW version 4 and GPC are 1km.

The ancillary data used for population modeling included seven satellite-derived raster datasets (i.e., consisting of grid cells). The LULC data showed strong relationships with the population distribution, particularly those types indicating human settlements [1,37,38]. In this study, the BaseVue 2013 (http://www.mdaus.com/products/land-cover-products) 30 m land cover dataset, which contains 14 types of land cover, was used as ancillary data. The Net Primary Productivity (NPP) is a key driving force of human distribution [39]; we included the 2010 MODIS 17 A3 annual NPP as our NPP data [40]. Furthermore, the precipitation, temperature, slope angle, and elevation of an area also affect human settlement patterns to some extent, which may directly relate to the population distribution [41]. In this case, we employed digital elevation data and its derived slope data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM), as well as the WorldClim/BioClim 1970–2000 mean annual precipitation (BIO12) and mean annual temperature (BIO1) data [42]. Moreover, before the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS) imagery was produced in 1992, the relationship between quantitative population and NTL had been revealed by several researchers [25,43]. NTL is a straighter way to indicate population distribution. We employed the Suomi National Polar-Orbiting

Partnership Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) 500 m resolution NTL data as our ancillary data, which has a higher spatial resolution than DMSP-OLS's 1 km [44].

We also used vector data (i.e., consisting of geometric shapes such as points, lines, and polygons) as ancillary data, including POI, roads and buildings data. The POIs are places of various categories with location information, and some of them are related to human activities [29]. Road networks were obtained from the AutoNavi Map Service, which is one of the largest web map service providers in China [45]. Building footprint and height are very useful for an accurate population estimation, especially in urban regions with a dense population [2,46,47]. In Zhengzhou metropolitan area, these were acquired from AutoNavi Map Services (https://www.amap.com/). Ancillary data used in this study is shown in Table 1.

Table 1. List of ancillary data used in this study.

Data	Description	Year	Source
	Raster data (remote sensing data)		
BaseVue 2013 landcover data	A 30 m spatial resolution global land cover dataset containing 14 types of land cover, including water, wetland, general agricultural land, paddy agricultural land, urban, etc.	2013	MDA Information Systems LLC., USA
MOD17A3 NPP data	A 1 km MODIS annual product that provides an accurate measure of the net primary productivity of terrestrial vegetation	2010	National Aeronautics and Space Administration (NASA), USA
VIIRS 2012 night-time lights data	500 m resolution lights at night that exclude fires and other ephemeral lights	2012	National Oceanic and Atmospheric Administration (NOAA), USA
ASTER GDEM Version 2 data	A global effort that provides 30 m resolution elevation information	-	United States Geological Survey (USGS), USA
WorldClim Version2 temperature data	A global dataset that measured mean temperatures from 1970 to 2000	1970–2000	The Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL), USA
WorldClim Version2 precipitation data	A global dataset that measured mean precipitation from 1970 to 2000	1970–2000	The Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL), USA
	Vector data (social sensing data)		
Boundary maps	Township and county Level Administrative Boundaries	2010	Henan Administration of Surveying Mapping and Geoinformation, China
Road networks	Including railway, national road, provincial road, county road, and township road	2018	AutoNavi Software Co., Ltd., China
Point of interest	20 categories including: residential communities, banks, parking lots, etc.	2010	Baidu Inc., China
Building footprint	Building footprints with height information	2018	AutoNavi Software Co., Ltd., China

2.3. Data Preparation

The ancillary data were processed to enrich information relevant to population distribution. Since jiedao level census data (transformed into logarithmic population density as response variable) was the finest official population data we could get, the information extracted from ancillary data was summarized at jiedao level and used as independent variables. Different strategies were employed based on the type of data. Areas of different land cover types and the distance to these land cover types can denote human presence from different perspectives. In every jiedao, each type of land cover (e.g., general agricultural land, shrub, urban) was extracted, after which its proportion and the mean value of the Euclidean distance was calculated. For the other RS data (i.e., NPP, NTL, DEM, slope, temperature, and precipitation), the mean value of each dataset in every jiedao was calculated. The distance to buildings, overall building volume (the total of volumes of all buildings in a jiedao), and building volume density (m^3/km^2, overall building volume in each jiedao divided by areas of each jiedao) in each jiedao were attained. Moreover, categories of POIs that may relate to population distribution were first selected based on a literature review [24,29,30]. Twenty categories of POIs, including residential communities, educational locations, hospitals (clinics), parking lots, parks, government buildings, airports, railway stations, bus stations, motor passenger stations, gas stations,

service zones of highways, toll stations, banks, commercial buildings, retails, hotels, restaurants and entertainments, companies, and others (e.g., small business, museum) were selected. Each category of POI was analyzed by Euclidean distance, since the radial distance from the nearest POI had been found to be related to population density [20,29,48]. In order to attain the road density in each jiedao, the road networks were processed by a road density function [49] (1):

$$\mathrm{RSD} = \left(3 \times N_r + 3 \times N_{ne} + 2 \times N_{pe} + N_{cr} + 0.4 \times N_{tr}\right)/A \quad (1)$$

where RSD represents the road network density (km/km^2), and *Nr*, *Nne*, *Npe*, *Ncr*, and *Ntr* stand for the length of railroads, national roads, provincial roads, county roads, and jiedao level roads, respectively, in each jiedao. *A* is the area of each jiedao and the numbers of 3, 3, 2, 1, and 0.4 are conversion ratios based on transport and traffic capacity of the different types of roads. All independent variables generated in this section were classified into three categories (i.e., RS and road variables, POI variables, and building footprint variables). The whole data preparation procedure is shown in a flowchart (i.e., Figure 2) and processed based on ArcGIS 10.2.

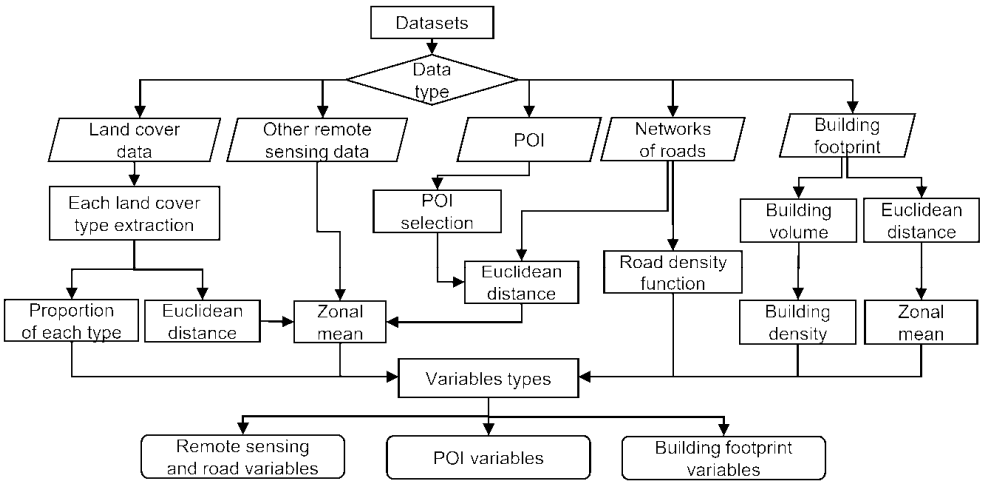

Figure 2. Flowchart of the data preparation process.

2.4. Population Modeling

2.4.1. Random Forest Model

The RF is an ensemble, nonparametric classifier based on a set of Classification and Regression Trees (CARTs) [32]. CARTs in RF are created on the basis of bootstrap samples and a user-defined number of features, which indicate that each CART is independently constructed by drawing samples with replacement from training datasets, and each node of a CART is split using the largest information gain among a subset that is randomly chosen from variables [50]. After the construction of CARTs, a simple majority vote is conducted for classification (arithmetic mean for regression). Because of the bootstrap, the samples involved in the construction of CARTs normally account for two thirds of the training dataset; the remaining one third is called Out-Of-Bag (OOB) data, which can be used for an unbiased internal cross-validation that estimates the performance of the constructed model. The OOB data can also involve in a measurement of the importance score of each variable in RF [32]. By comparing the OOB accuracy of the model before and after the permutation of the variable, an indicator called Mean Decrease Accuracy (MDA) can be used for variable importance evaluation. Important variables cause a higher OOB accuracy decline than unimportant variables [31]. Another indicator of variable importance is the Mean Decrease Impurity (MDI), which measures the effectiveness of variables at reducing uncertainty when creating CARTs in RF [51]. However, since MDI produces more biases while evaluating continuous variables, we decided to use MDI as an aiding

indicator and MDA as the main indicator to rank the importance of variables [52]. The advantages of RF have been explored in several investigations. For example, compared to simple decision trees, it is less affected by noises and can avoid overfitting, because of the way of sampling and the voting of multiple decision trees [53]. In comparison with multi-linear regression, RF is less susceptible to multicollinearity, which makes it more efficient in terms of feature selection [50,54].

To be able to incorporate more training units in the RF algorithm and to include sufficient information indicating population distribution in suburban regions, 28 jiedaos outside the five municipal districts and within the scope of Zhengzhou metropolitan area were included in the study area. Together with the 66 jiedaos in the five municipal districts, a total of 94 jiedaos were used as RF input. While building the RF model, it is important to choose variables that can improve the model accuracy before taking further steps. In this study, we simply selected variables with positive MDA scores. Binary regression was used to understand the direction of the association between each important variable and the population density. After that, two significant parameters named *mtry* and *ntree* were adjusted. *mtry* represents the usage of numbers of variables during a decision tree splitting, and *ntree* is the number of decision trees to be generated in the model. In this study, *mtry* was first decided according to the lowest OOB error. Furthering that, *ntree* was confirmed by the lowest mean squared error and selected *mtry*. After model establishment, variable values within each 100 m grid were used to predict the population density in each grid. This gridded population density layer was then employed as a weighting layer for population redistribution. To highlight the usefulness of POI and building footprint data in population modeling, we additionally employed three combinations of variables (i.e., RS and road variables; RS, and road and POI variables; RS, and road and building footprint variables) as RF input to generate three other weighting layers through the procedure described above. Each of these weighting layers was then used to redistribute census data and therefore a comparison was made.

2.4.2. Dasymetric Mapping

The weighting layers generated by RF were devoted to disaggregating census data. In this stage, we redistributed county level census data into 100 m resolution raster surfaces by Function (2):

$$P_p = C_r \times PD_p / PD_r \quad (2)$$

where P_p is the estimated people count per pixel, PD_p is the people count per pixel produced by RF, r represents each county, C_r is the census population count in counties, and PD_r is the people count in counties summed by PD_p.

In Function (2), district/county level instead of jiedao level census data was redistributed because of the accuracy assessment requirements. The accuracy assessment requires a finer level census and boundary data than the ones used as inputs to dasymetrically disaggregate the administrative unit-based population counts. However, in China, village and community (a finer level than jiedao) boundaries change frequently due to rapid urbanization, and there are few chances to access precise and real-time village/community boundary data. Therefore, we chose county level census data for the population redistribution and employed jiedao level census data for the accuracy assessment. Additionally, due to this reason, we only redistributed census data for the five municipal districts, rather than the whole Zhengzhou metropolitan area, since some counties outside of the five municipal districts were partly out of the Zhengzhou metropolitan area, which could lead to an incomplete population reference for these counties.

2.4.3. Abnormal Detection

A fundamental assumption of any dasymetric population modeling is that the population density should be similar in similar areas, which could be characterized by the used ancillary data [2,10]. However, this assumption may not always hold true in the real world, as there may be some reasons

that could invalidate this assumption in certain areas, e.g., vacant houses, and houses being rented by more people than allowed, which cannot be detected and modelled by the used ancillary data. Including such areas with an extremely abnormal population density at the model training stage could obscurely and unpredictably lead to a distortion of the predicted population density (i.e., the weighting layer) in most of the areas with a more common population density. In addition, since areas with an extremely abnormal population density could not be well modelled by the used ancillary data, an over- or under-estimated population density in these areas could lead to a wrongly redistributed population across all areas, and hence unfairly affect the accuracy of the population products, without providing any benefit to local planning and management sectors. Therefore, detecting and excluding such areas with an extremely abnormal population density from model training and dasymetric mapping should be able to not only provide potentially useful information to relevant sectors, but also mitigate over- or under-estimation issues in most of the areas with a normal population density.

Boxplot as a statistical tool is used to visually summarize groups of data, and it could also detect abnormal values among a group of data through their quartiles [55]. The values that are larger than the upper quartile by at least k times the interquartile range (i.e., range of the upper and lower quartile) or smaller than the lower quartile by at least k times the interquartile range, are considered abnormal values [56]. According to Tukey [57], the parameter k can be set as 1.5 and 3 to illustrate abnormal values and extremely abnormal values, respectively. In this study, the values of the differences between the actual and predicted population in each jiedao were mapped in a boxplot. Jiedaos with extremely high values (both positive and negative) were considered areas with an abnormal population density, which were wrongly populated by our ancillary data. Considering the poor quality of ancillary data could also raise the differences between the actual and predicted population in some areas. We finally chose k as 3 to minimize this disturbance and only considered areas with values with an extremely high difference as areas with an abnormal population density. The relationship of this stage in our study is shown in a flowchart (i.e., Figure 3).

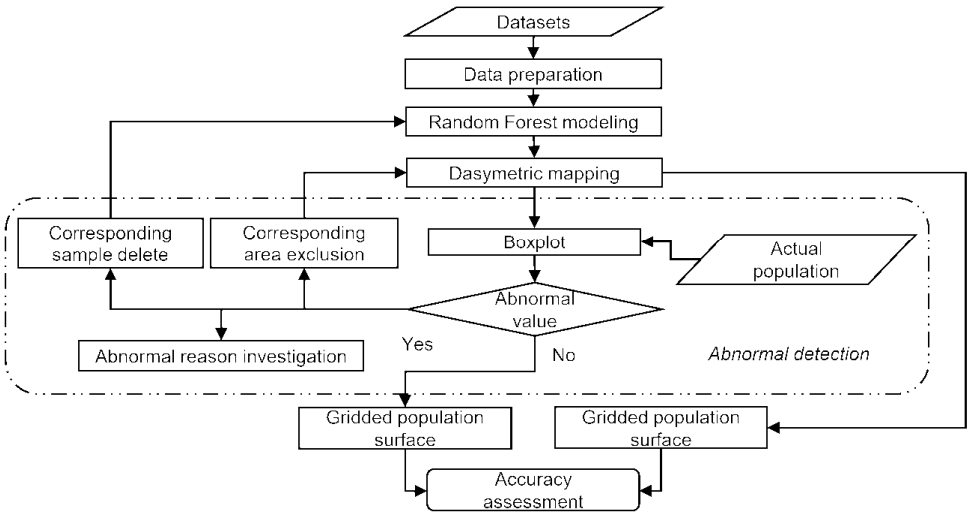

Figure 3. Relationship between abnormal detection and our approach of population mapping.

2.4.4. Accuracy Assessment

Worldpop, GPW, and GPC were employed to compare with our product. Furthermore, gridded population datasets generated before removing the areas with an abnormal population density from model training and dasymetric mapping were used to assess the improved accuracy of the estimated population in other areas, with the more common population density remaining after the exclusion of areas with an abnormal population density. Gridded population datasets generated by the three combinations of variables were used to examine the performance of POI and building footprint data in the population modeling. The population grids of each dataset were aggregated to jiedao

level, after which the difference between each dataset and the census data was calculated. The Root Mean Square Error (RMSE) (3) and the Mean Absolute Error (MAE) (4) were used to quantify the difference [58,59].

$$\mathrm{RMSE} = \sqrt{\frac{\sum\left(P_i^{estimated} - P_i^{observed}\right)^2}{N}} \quad (3)$$

$$\mathrm{MAE} = \frac{1}{N}\sum\left|P_i^{estimated} - P_i^{observed}\right| \quad (4)$$

In Functions (3) and (4); N represents the number of jiedaos (one level lower than county/district) included in the accuracy assessment, $P_i^{estimated}$ is the estimated population in a jiedao i, and $P_i^{observed}$ is the population of the census data in jiedao i.

3. Results

3.1. Abnormal Detection

For each gridded population dataset generated by different combinations of variables, a boxplot was drawn to evaluate the differences between the estimated and actual population numbers in each jiedao. Results showed that one jiedao named Jichenglu was an abnormal unit with an extremely high positive difference value (Figure 4). In addition, since this jiedao was also detected as an area with an abnormal population density without using our building footprint data, temporally mismatched building footprint data were not the main reason for such a severe mis-prediction.

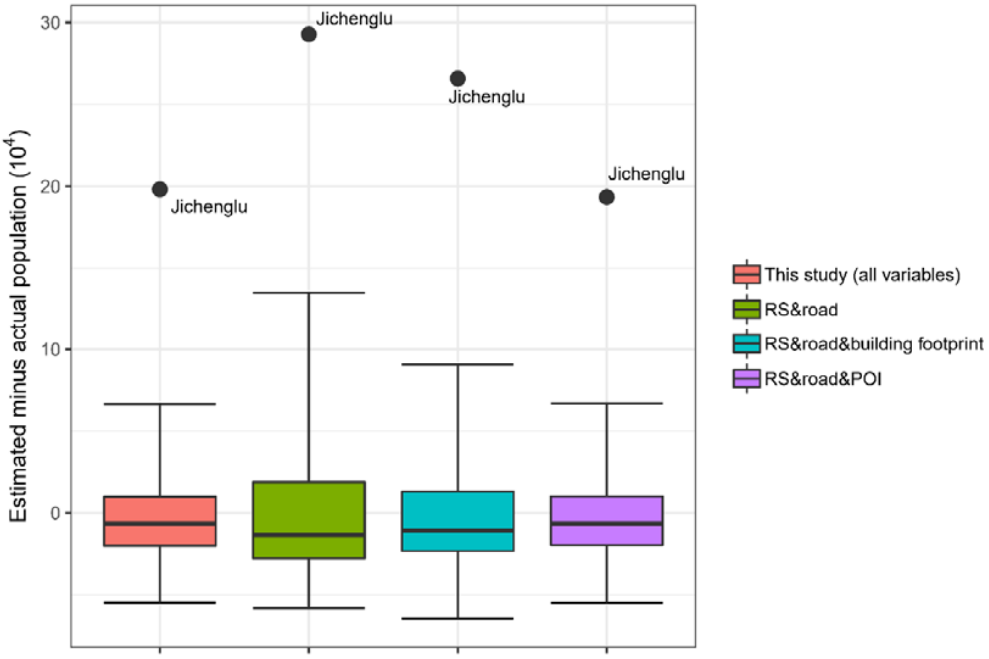

Figure 4. Boxplots of the difference between the estimated and actual population in each jiedao for gridded population datasets generated by different combinations of variables (i.e., remote sensing and road variables; remote sensing, and road and building footprint variables; remote sensing, and road and point of interest variables; all variables).

3.2. Accuracy Assessment

The estimated population numbers were used to compare with the census population counts at jiedao level to assess the performance of POI and building footprint data. The results showed that whether we exclude the area with an abnormal population density from the study area, the population datasets generated with POI and building footprint data achieved a better RSME and MAE than those

datasets only produced by RS and road data (Figure 5). Higher correlations between the estimated and census population were also found in datasets generated with POI and building footprint data. Temporally mismatched building footprint data performed relatively worse than POI data. Moreover, after excluding the area with an abnormal population density from both model training and dasymetric mapping, the accuracy of the estimated population in most areas with a more common population density was improved. As for the two datasets generated by all variables; the one that excluded the area with an abnormal population density achieved a better accuracy (RMSE = 24,956.93, MAE = 19,420.04) than the other (RMSE = 27,267.04, MAE = 21,352.75). Similar results were also found in datasets generated by other combinations of variables. The result of this study, which were produced after removing the area with an abnormal population density from the study area, was compared with three other widely used products. Results showed that the accuracy of our study was better than that of the others in both RMSE and MAE (Table 2).

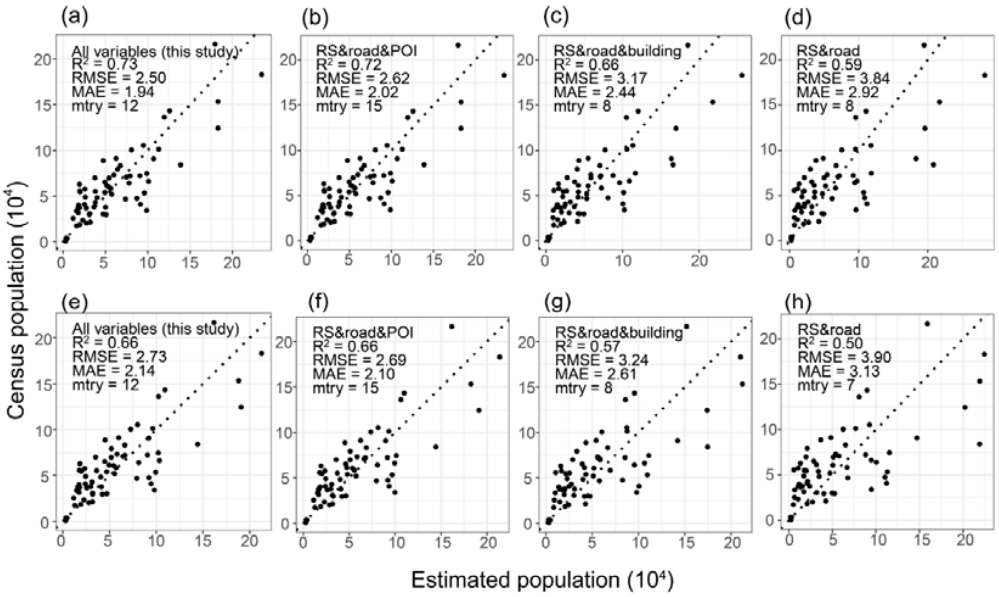

Figure 5. Scatterplots between the census population and the estimated population generated by (**a**) all variables, (**b**) remote sensing, and road and point of interest variables, (**c**) remote sensing, and road and building footprint variables, (**d**) remote sensing and road variables, (**e**) all variables with exclusion of abnormal units, (**f**) remote sensing, and road and point of interest variables with exclusion of abnormal units, (**g**) remote sensing, and road and building footprint variables with exclusion of abnormal units, (**h**) remote sensing and road variables with exclusion of abnormal units, at jiedao level in areas with a normal population density in Zhengzhou in 2010. *mtry* represents the number of variables randomly sampled as candidates at each split of decision trees in random forest, and the number of decision trees to grow in random forest were set as 500 for all datasets.

Table 2. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) of each dataset.

	RMSE	MAE
This study	24,956.93	19,420.04
Worldpop	31,543.66	22,687.94
GPW	33,791.59	26,132.49
GPC	35,800.90	29,074.32

GPW—Gridded Population of World version 4; GPC—1-km Grid Population Dataset of China.

Four residential communities in the jiedao with an abnormal population density were selected as samples for our field survey of abnormal population density validation (Figure 6). The field survey was conducted on 16 May, 2017, and was based on the interviews with residents and the staff of property management in each community. According to the property management records and the interviews,

we found that few people dwelled in these communities in 2010, since the constructions of these residential communities were just completed in 2008, 2009, and 2010, and people were gradually moved in after decoration. As such, the population density in these communities was extremely low compared to communities in other jiedaos in 2010. Furthering that, according to the literature review, the jiedao with an abnormal population density was a key development area planned by the government and experienced fast urbanization during the period from 1995 to 2010 [60]. Multiple commercial residential buildings and corresponding facilities with a low occupation rate were built [61,62].

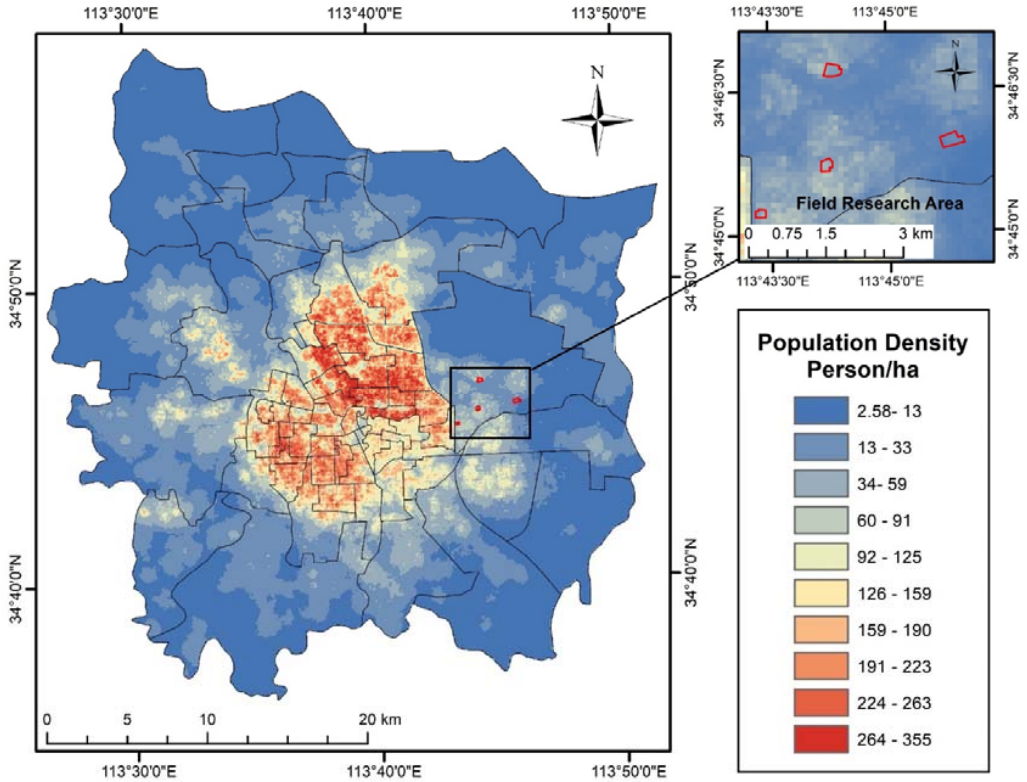

Figure 6. Overview of the gridded population distribution in Zhengzhou (the sampled communities are zoomed in).

3.3. Variable Importance

The thirty most important variables were ranked by MDA and MDI. The direction of the association between each variable and the population density was indicated by the coefficient from binary regression (Figure 7). According to the rank of MDA, distance to the POIs of residential communities, parking lots, government buildings, and distance to roads contributed the most to the population estimation in Zhengzhou, China. A negative association between each of these variables and the population density was found. Building density and overall building volume also contributed to the model accuracy, which were positively associated with population density. Eight of the ten most important variables were originally derived from social sensing data. The proportion of general agricultural land and DEM were the only two RS based variables in the top ten list. Road density, measured by the road density function, was not contributing enough to be included in the list of the thirty most important variables.

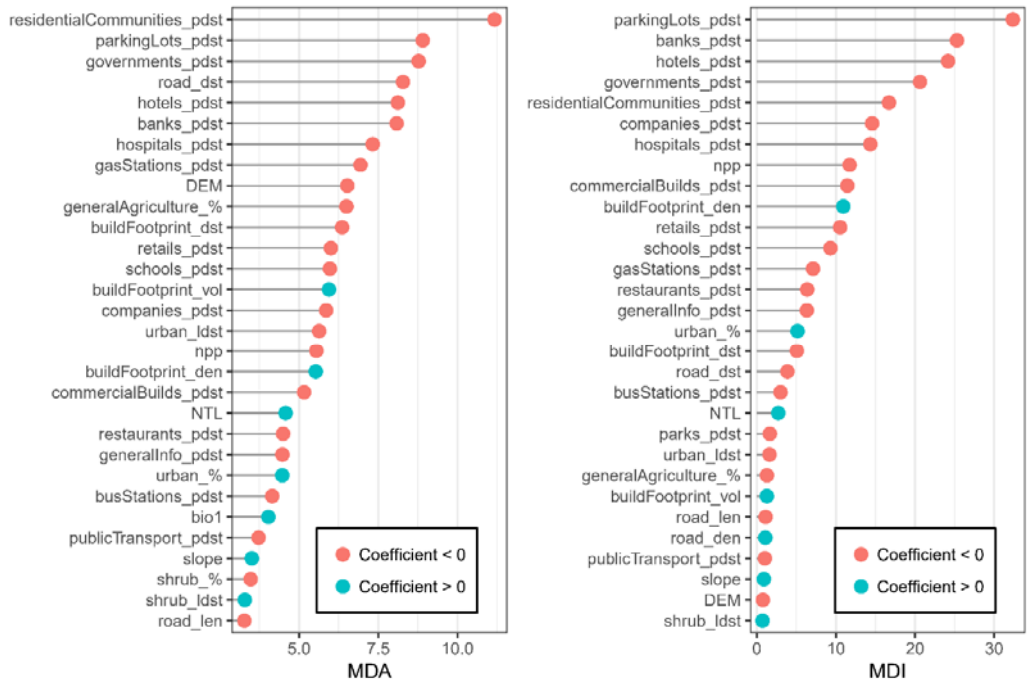

Figure 7. The thirty most important variables ranked according to their contribution to the random forest modeling, measured by Mean Decrease Accuracy (MDA) and Mean Decrease Impurity (MDI), with direction of the association between each ranked variable and the population density. Variable names: xx_pdst—distance to a category of points of interest; xx_ldst—distance to a landcover type; xx_%—proportion of a landcover type; xx_den—density of ?; buildFootprint_vol—volume of all buildings in a jiedao; xx_len—length of ?; DEM—Digital Elevation Model; NTL—Nighttime Lights; npp—Net Primary Productivity.

4. Discussion

The study provided an example of small-area population mapping using an RF algorithm, on the basis of RS and social sensing data in Zhengzhou, which is a typical metropolitan area in China. The variables that were important to the population distribution and their associations with population density were demonstrated. A boxplot of the differences between estimated and actual population counts in each jiedao was used to detect jiedaos with an abnormally high/low population density, which was validated in the field work. After excluding that jiedao, the final population product in the other areas that resulted from this study outperformed three widely used population products (i.e., Worldpop, GPW, and GPC).

The distance to the surrounding residential communities, parking lots, government buildings, roads, and hotels were found to be the top five influential factors of population distribution, which were all negatively associated with population density. The proportions of general agricultural land were also useful to characterize urban and suburban areas with a different population density. With some of the RS data contributed to our study, detailed social sensing data, such as categories of POIs and building footprint data, can provide socioeconomic factors and finer human settlement information that are relevant to human presence, and hence accurately depict the local population distribution [29,30]. The negative association between road length and population density in our study area was mainly because jiedaos in urban areas with a higher population density are normally smaller than jiedaos in suburban regions with a lower population density; and larger jiedaos tend to have longer total length of roads. In addition, the positive association between slope and population density was primarily due to the relatively poor quality of DEM data. This is because the buildings in urban areas are influencing the accuracy of DEM data [63]; tall buildings in urban areas lead to higher values of slope. Moreover, the population distribution in some areas may be difficult to be explained by normal

reasons. Some abnormal reasons, such as vacant houses and houses being rented by more people than allowed, violate the fundamental assumption of dasymetric population modeling that the population density should be similar in similar areas, which could be characterized by the used ancillary data. This violation can obscurely and unpredictably lead to an over- or under-estimation of the population, and hence unfairly affect the accuracy of the population products without providing any benefit to local planning and management sectors. By using a boxplot, one jiedao in our study area was found to be abnormally populated. It was supposed to have a high population density on the basis of the values of the important variables in this jiedao (e.g., close to residential communities, parking lots, etc.). However, because of the high housing vacancy rate, the actual population density in this jiedao was relatively low and extremely different from our prediction [61,62]. Such detection is useful for improving both model performance and accuracy by removing such abnormal units from modeling, but providing information to relevant sectors to potentially solve some other problems may have caused the abnormal population distribution.

The spatial population distribution can be somewhat reflected by many indicators, such as NTL [25], LULC [64], building volume [46,65], POI [30], slope angle, elevation of an area [20], etc. However, the varying performances of these data is a common problem in terms of predicting population distribution in different geographic and socioeconomic conditions. Therefore, choosing an optimal set of ancillary data is of prime importance. By employing RF, we could identify the variables that could improve the model accuracy with multicollinearity overcome. Additionally, the quality of ancillary data is important to population modeling. Our study employed POIs and building footprint data from the two largest map service companies in China (i.e., Baidu and AutoNavi), which provided more complete and reliable data for our study area than the volunteered OpenStreetMap used in WorldPop [66].

There were several limitations to our study. First, over-fitting may be an issue that affected the predictability of our model for future projection of the population. The limited number of jiedaos (i.e., a small sample size) could only provide a relatively small amount of information in our study area, which could restrict the predictability of our model in a larger area with more complex patterns of population distribution. This needs to be further confirmed in future studies with a larger number of areal units included (i.e., a large sample size), so traditional regression methods could be used to compare with machine learning methods. Additionally, due to the small sample size; the direction, and especially the significance of the associations between independent variables and population density, were just preliminarily explored by binary regression, instead of being estimated in multivariate regression. Second, the accuracy of the detection of areal units with an abnormal population density may be affected by its size. A larger areal unit tends to contain more errors (i.e., differences between estimated and actual population counts) during population modeling, which might be wrongly detected as an outlier by the boxplot. Therefore, results of abnormal detection could be further investigated and validated by other abnormal detection methods, such as the k-nearest neighbors algorithm, especially when field validation is not possible in large or inaccessible study areas. Moreover, since this study only detected one jiedao as an area with an abnormal population density, we were unable to build a model specifically for this area and therefore unable to make a better prediction for this particular area. With a larger study area, more abnormal population density units could be found and therefore modelled accurately. Third, the range of data values at the training stage (at the township level) was smaller than that at the prediction stage (at the 100 m grid cells), as extreme values were averaged in courser units. This may lead to an over- or under-estimation of the population density values [67]. It would be more suitable, whenever possible, to conduct model training at finer scales, such as at village or community levels.

5. Conclusions

The high-resolution population product developed in this study, on the basis of the RF method, and RS and social sensing data, outperformed three global population products (i.e., Worldpop,

the Gridded Population of the World version 4, and the 1 km Grid Population Dataset of China). Building footprint data, combined with height information and categories of POI data, can increase the accuracy of the population estimates. Certain types of POI data, including residential communities, parking lots, government buildings, hotels, banks, hospitals, and gas stations were more useful predictors than building footprint data to model population distribution at a 100 m grid cell level. Our approach of population modeling that exclude areas with an abnormal population density (i.e., areas with an extremely different population density than other similar areas) from model training and dasymetric mapping, increased the accuracy of population estimates in most of the areas with a more common population density. Such an exclusion can be applied to larger areas with a more complex population distribution, in order to model population density with a minimum disturbance by abnormal population densities. This holds promise for many applications, including public health, environmental protection, regional planning and development, and policy making.

Author Contributions: Conceptualization, P.J.; methodology, G.Q. and P.J.; formal analysis, G.Q. and P.J.; resources, P.J., Y.B. and C.W.; data curation, G.Q.; writing—original draft preparation, G.Q., P.J., X.Y. and T.Y.; writing—review and editing, P.J., A.S., Y.B., T.Y. and C.W.; visualization, G.Q. and T.Y.; investigation, G.Q. and P.J.; supervision, P.J. and Y.B.; project administration, P.J.; funding acquisition, P.J. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Program of the National Natural Science Foundation of China (Grant No. 61631011); and the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2013BAK05B01).

Acknowledgments: P.J., director of the International Initiative on Spatial Lifecourse Epidemiology (ISLE), thanks the Netherlands Organization for Scientific Research, the Royal Netherlands Academy of Arts and Sciences, the Chinese Centre for Disease Control and Prevention, and the West China School of Public Health and West China Fourth Hospital in Sichuan University for funding ISLE and supporting its research activities. The authors acknowledge the four anonymous reviewers and Editor for their constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Azar, D.; Engstrom, R.; Graesser, J.; Comenetz, J. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sens. Environ.* **2013**, *130*, 219–232. [CrossRef]
2. Jia, P.; Qiu, Y.; Gaughan, A.E. A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. *Appl. Geogr.* **2014**, *50*, 99–107. [CrossRef]
3. 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.
4. Jia, P.; Anderson, J.D.; Leitner, M.; Rheingans, R. High-resolution spatial distribution and estimation of access to improved sanitation in Kenya. *PLoS ONE* **2016**, *11*, e0158490. [CrossRef]
5. Elvidge, C.D.; Baugh, K.E.; Kihn, E.A.; Kroehl, H.W.; Davis, E.R.; Davis, C.W. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *Int. J. Remote Sens.* **1997**, *18*, 1373–1379. [CrossRef]
6. Zhang, N.; Huang, H.; Su, B.; Zhang, H. Population evacuation analysis: Considering dynamic population vulnerability distribution and disaster information dissemination. *Nat. Hazards* **2013**, *69*, 1629–1646. [CrossRef]
7. Wilson, R.; Erbachschoenberg, E.Z.; Albert, M.; Power, D.; Tudge, S.; Gonzalez, M.; Guthrie, S.; Chamberlain, H.; Brooks, C.; Hughes, C. Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake. *PLoS Curr.* **2016**, *8*. [CrossRef]
8. Jia, P.; Shi, X.Y.; Xierali, I.M. Teaming up census and patient data to delineate fine-scale hospital service areas and identify geographic disparities in hospital accessibility. *Environ. Monit. Assess.* **2019**, *191*, 303. [CrossRef]
9. Jia, P.; Wang, F.H.; Xierali, I.M. Differential effects of distance decay on hospital inpatient visits among subpopulations in Florida, USA. *Environ. Monit. Assess.* **2019**, *191*, 381. [CrossRef]

10. Mennis, J. Generating Surface Models of Population Using Dasymetric Mapping. *Prof. Geogr.* **2003**, *55*, 31–42.
11. Yi, G.; Hui, W.; Wang, P. Population Spatial Processing for Chinese Coastal Zones Based on Census and Multiple Night Light Data. *Resour. Sci.* **2013**, *35*, 2517–2523.
12. Martin, D. Directions in population GIS. *Geogr. Compass.* **2011**, *5*, 655–665. [CrossRef]
13. Tobler, W.; Deichmann, U.; Gottsegen, J.; Maloy, K. World population in a grid of spherical quadrilaterals. *Int. J. Popul. Geogr.* **1997**, *3*, 203–225. [CrossRef]
14. Tobler, W.R. Smooth Pycnophylactic Interpolation for Geographical Regions. *J. Am. Stat. Assoc.* **1979**, *74*, 519–530. [CrossRef]
15. Langford, M.; Harvey, J.T. The Use of Remotely Sensed Data for Spatial Disaggregation of Published Census Population Counts. In Proceedings of the IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, DFUA 2001, Rome, Italy, 8–9 November 2001; pp. 260–264.
16. Zhou, C.; Yang, O.U.; Ting, M.A. Progresses of Geographical Grid Systems Researches. *Prog. Geogr.* **2009**, *28*, 657–662.
17. Balk, D.; Yetman, G. *The Global Distribution of Population: Evaluating the Gains in Resolution Refinement*; Center for International Earth Science Information Network (CIESIN), Columbia University: New York, NY, USA, 2004.
18. Balk, D.L.; Deichmann, U.; Yetman, G.; Pozzi, F.; Hay, S.I.; Nelson, A. Determining Global Population Distribution: Methods, Applications and Data. *Adv. Parasitol.* **2006**, *62*, 119–156.
19. Freire, S.; Doxsey-Whitfield, E.; MacManus, K.; Mills, J.; Pesaresi, M. Development of new open and free multi-temporal global population grids at 250 m resolution. *Population* **2000**, *250*.
20. 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]
21. Leyk, S.; Gaughan, A.E.; Adamo, S.B.; de Sherbinin, A.; Balk, D.; Freire, S.; Rose, A.; Stevens, F.R.; Blankespoor, B.; Frye, C.; et al. The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth Syst. Sci. Data* **2019**, *11*, 1385–1409. [CrossRef]
22. Linard, C.; Gilbert, M.; Tatem, A.J.J.G. Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* **2011**, *76*, 525–538. [CrossRef]
23. Cohen, J.E.; Small, C. Hypsographic demography: The distribution of human population by altitude. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14009–14014. [CrossRef] [PubMed]
24. Ye, T.T.; Zhao, N.Z.; Yang, X.C.; Ouyang, Z.T.; Liu, X.P.; Chen, Q.; Hu, K.J.; Yue, W.Z.; Qi, J.G.; Li, Z.S.; et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Sci. Total Environ.* **2019**, *658*, 936–946. [CrossRef] [PubMed]
25. Sutton, P.; Roberts, D.; Elvidge, C.; Baugh, K. Census from Heaven: An estimate of the global human population using night-time satellite imagery. *In. J. Remote Sens.* **2001**, *22*, 3061–3076. [CrossRef]
26. Briggs, D.J.; Gulliver, J.; Fecht, D.; Vienneau, D.M. Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sens. Environ.* **2007**, *108*, 451–466. [CrossRef]
27. Alahmadi, M.; Atkinson, P.M.; Martin, D. A Comparison of Small-Area Population Estimation Techniques Using Built-Area and Height Data, Riyadh, Saudi Arabia. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 1959–1969. [CrossRef]
28. Roni, R.; Jia, P. An Optimal Population Modeling Approach Using Geographically Weighted Regression Based on High-Resolution Remote Sensing Data: A Case Study in Dhaka City, Bangladesh. *Remote Sens.* **2020**, *12*, 1184. [CrossRef]
29. 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]
30. Yang, X.C.; Ye, T.T.; Zhao, N.Z.; Chen, Q.; Yue, W.Z.; Qi, J.G.; Zeng, B.; Jia, P. Population Mapping with Multisensor Remote Sensing Images and Point-Of-Interest Data. *Remote Sens.* **2019**, *11*, 574. [CrossRef]
31. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
32. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Tatem, A.J.; Gaughan, A.E.; Stevens, F.R.; Patel, N.N.; Jia, P.; Pandey, A.; Linard, C. Quantifying the effects of using detailed spatial demographic data on health metrics: A systematic analysis for the AfriPop, AsiaPop, and AmeriPop projects. *Lancet* **2013**, *381*, S142. [CrossRef]

34. Tan, M.; Liu, K.; Liu, L.; Zhu, Y.; Wang, D. Spatialization of population in the Pearl River Delta in 30 m grids using random forest model. *Prog. Geogr.* **2017**, *36*, 1304–1312.
35. Fu, J.; Jiang, D.; Huang, Y. 1 km grid population dataset of China (2005, 2010). *Acta Geogr. Sin.* **2014**, *69*, 136–139. [CrossRef]
36. Census Office; Department of Population and Employment Statistics. *China 2010 Population Census Information*; China Statistics Press: Beijing, China, 2012.
37. Lo, C.P. Raster approach to population estimation using high-altitude aerial and space photographs. *Remote Sens. Environ.* **1989**, *27*, 59–71. [CrossRef]
38. Tatem, A.J.; Noor, A.M.; Von Hagen, C.; Di Gregorio, A.; Hay, S.I. High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLoS ONE* **2007**, *2*, e1298. [CrossRef]
39. Luck, G.W. The relationships between net primary productivity, human population density and species conservation. *J. Biogeogr.* **2007**, *34*, 201–212. [CrossRef]
40. Running, S.W.; Nemani, R.R.; Heinsch, F.A.; Zhao, M.S.; Reeves, M.; Hashimoto, H. A continuous satellite-derived measure of global terrestrial primary production. *Bioscience* **2004**, *54*, 547–560. [CrossRef]
41. Walsh, S.J.; Evans, T.P.; Welsh, W.F.; Entwisle, B.; Rindfuss, R.R. Scale-dependent relationships between population and environment in northeastern Thailand. *Photogramm. Eng. Remote Sens.* **1999**, *65*, 97.
42. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Clim.* **2010**, *25*, 1965–1978. [CrossRef]
43. Lo, C. Urban indicators of china from radiance-calibrated digital dmsp-ols nighttime images. *Ann. Assoc. Am. Geogr.* **2002**, *92*, 225–240. [CrossRef]
44. Elvidge, C.D.; Baugh, K.E.; Zhizhin, M.; Hsu, F.-C. Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proc. Asia Pac. Adv. Netw.* **2013**, *35*, 62. [CrossRef]
45. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
46. Wang, S.; Tian, Y.; Zhou, Y.; Liu, W.; Lin, C. Fine-scale population estimation by 3D reconstruction of urban residential buildings. *Sensors* **2016**, *16*, 1755. [CrossRef]
47. Tomás, L.; Fonseca, L.; Almeida, C.; Leonardi, F.; Pereira, M. Urban population estimation based on residential buildings volume using IKONOS-2 images and lidar data. *Int. J. Remote Sens.* **2016**, *37*, 1–28. [CrossRef]
48. Zhang, C.Y.; Qiu, F. A Point-Based Intelligent Approach to Areal Interpolation. *Prof. Geogr.* **2011**, *63*, 262–276. [CrossRef]
49. Bai, Z.; Wang, J.; Yang, Y.; Sun, J. Characterizing spatial patterns of population distribution at township level across the 25 provinces in China. *Acta Geogr. Sin.* **2015**, *70*, 1229–1242.
50. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R. News* **2013**, *2*, 18–22.
51. Hur, J.-H.; Ihm, S.-Y.; Park, Y.-H. A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing. *Wirel. Commun. Mob. Comput.* **2017**, *2017*, 6817627. [CrossRef]
52. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef]
53. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
54. He, L.; Levine, R.A.; Fan, J.; Beemer, J.; Stronach, J. Random forest as a predictive analytics alternative to regression in institutional research. *Pract. Assess. Res. Eval.* **2018**, *23*, 1.
55. Williamson, D.F.; Parker, R.A.; Kendrick, J.S. The box plot: A simple visual method to interpret data. *Ann. Intern. Med.* **1989**, *110*, 916–921. [CrossRef] [PubMed]
56. Frigge, M.; Hoaglin, D.C.; Iglewicz, B. Some implementations of the boxplot. *Am. Stat.* **1989**, *43*, 50–54. [CrossRef]
57. Tukey, J.W. *Exploratory Data Analysis: Limited Preliminary Ed*; Addison-Wesley Publishing Company: Ann Arbor, MI, USA, 1970.
58. Liu, X.; Kyriakidis, P.C.; Goodchild, M.F. Population-density estimation using regression and area-to-point residual kriging. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 431–447. [CrossRef]
59. Langford, M. An evaluation of small area population estimation techniques using open access ancillary data. *Geogr. Anal.* **2013**, *45*, 324–344. [CrossRef]

60. State Council of The People's Republic of China. *Gazette of the State Council of The People's Republic of China*; The State Council of The People's Republic of China, Ed.; General Office of The State Council of The People's Republic of China: Beijing, China, 1998; pp. 1004–3438.
61. Niu, J. Research on the Countermeasures for the Healthy Development of Commercial Housing Market in Zhengzhou City. *China Mark.* **2015**, 176–183. [CrossRef]
62. Guo, S. About Empty City, Vacancy and Housing Vacancy Rate. *City House* **2012**, 37–38.
63. Jacobsen, K.; Passini, R. Analsysis of ASTER GDEM Elevation Models. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: [2010 Canadian Geomatics Conference And Symposium Of Commission I, ISPRS Convergence In Geomatics-Shaping Canada's Competitive Landscape] 38 (2010), Nr. Part 1, Calgary, AB, Canada, 15–18 June 2010.
64. Jia, P.; Gaughan, A.E. Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Appl. Geogr.* **2016**, *66*, 100–108. [CrossRef]
65. Zhang, J.L.; Xu, W.; Qin, L.J.; Tian, Y.G. Spatial Distribution Estimates of the Urban Population Using DSM and DEM Data in China. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 435. [CrossRef]
66. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervas. Comput.* **2008**, *7*, 12–18. [CrossRef]
67. Sinha, P.; Gaughan, A.E.; Stevens, F.R.; Nieves, J.J.; Sorichetta, A.; Tatem, A.J. Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Comput. Environ. Urban Syst.* **2019**, *75*, 132–145. [CrossRef]