Article

Estimating and Interpreting Fine-Scale Gridded Population Using Random Forest Regression and Multisource Data

Yun Zhou ^{1,2} , Mingguo Ma ^{1,2,*} , Kaifang Shi ^{1,2} and Zhenyu Peng ³

¹ Chongqing Jinpo Mountain Field Scientific Observation and Research Station for Karst Ecosystem, Ministry of Education, School of Geographical Sciences, Southwest University, Chongqing 400715, China; hellozhouyun@email.swu.edu.cn (Y.Z.); skffyy@swu.edu.cn (K.S.)

² Chongqing Engineering Research Center for Remote Sensing Big Data Application, Southwest University, Chongqing 400715, China

³ Chongqing Engineering Research Center for Big Data Application in Spatial Planning, Chongqing Planning & Design Institute, Chongqing 401120, China; zgb@cqghy.com

* Correspondence: mmg@swu.edu.cn; Tel.: +86-23-6825-3912

Received: 14 April 2020; Accepted: 29 May 2020; Published: 3 June 2020

Abstract: Gridded population results at a fine resolution are important for optimizing the allocation of resources and researching population migration. For example, the data are crucial for epidemic control and natural disaster relief. In this study, the random forest model was applied to multisource data to estimate the population distribution in impervious areas at a 30 m spatial resolution in Chongqing, Southwest China. The community population data from the Chinese government were used to validate the estimation accuracy. Compared with the other regression techniques, the random forest regression method produced more accurate results ($R^2 = 0.7469$, $RMSE = 2785.04$ and $p < 0.01$). The points of interest (POIs) data played a more important role in the population estimation than the nighttime light images and natural topographical data, particularly in urban settings. Our results support the wide application of our method in mapping densely populated cities in China and other countries with similar characteristics.

Keywords: population mapping; points of interest; random forest; urban area; Chongqing

1. Introduction

Gridded population data can reflect the explicit size and detailed distribution of the population. They are indispensable for scientific research and policy making, such as in city development planning [1], disaster prevention [2], public health including novel coronavirus disease 2019 (COVID-19) [3] and environmental conservation [4]. In the past decades, many cities have undergone rapid population growth and urban area expansion in the world, particularly in developing countries [5], which makes the increasing population become one of the major challenges for cities. The uncertainties of population information, such as the size, location and migration, have become urgent policy making issues that need to be solved for many governments [6]. Many countries have tried to collect detailed population data through censuses during certain periods of time, though collecting demographical data have high time consumption and high costs and the data may possibly be outdated after it is collected [7,8]. To protect privacy, the high resolution population distribution census data linked with small areas were restricted [9,10]. Consequently, urban population mapping data at a high resolution are often unavailable, which cause difficulties for studies. Nevertheless, thanks to the development of computer science and new models, there have been some studies that have used novel methods to obtain high spatial resolution population data without eroding privacy.

In the 1990 s, many scholars focused on the global population mapping methods and developed many global population datasets [11,12]. In 1995, for the first time, the Gridded Population of the World (GPW) collection disaggregated the global census data to a resolution of 2.5 arc-minutes (~4 km). The latest version (GPWv4) is at fine resolution of 30 arc-seconds (~1 km) using an areal weighting method and land cover datasets [12]. Similarly, the Gridded Rural Urban Mapping Project (GRUMP) succeeded in allocating the global population into 1 km grids considering the location and scale of the city center [13]. LandScan also provided global population data with a spatial resolution of approximately 1 km [14]. The Worldpop took advantage of the random forest model and multisource data, which made the 100 m resolution of global population data possible [15]. These population datasets employed different ancillary datasets with various approaches (Table 1). These datasets featured globally large scales with significance for the data-poor regions, but with concerns regarding accuracy.

Table 1. List of population mapping datasets.

Dataset	Scope	Resolution	Method	Ancillary Data
GPWv4	World	1 km	Areal-weighting	Land cover dataset
GRUMP	World	1 km	Mass-conserving	Night-time lights dataset and buffered settlement centroids
LandScan	World	1 km	Smart interpolation	Roads, slope, land cover, Night-time lights, etc.
Worldpop	World	100 m	Random forest	Settlement location and extent, land cover, satellite nightlights, topography, roads, buildings, etc.
Our research	Complex city	30 m	Random forest	Points of interest, satellite nighttime lights, topography, etc.

There are also some attempts that have used new data and methods [7] to conduct high-resolution population research at the national level [1,16,17]. However, the various urban fabric patterns and city complexity constrain the development of population mapping [8,18,19]. Satellite-derived products, e.g., topography [20], nighttime light images [21], built-up area [22] and land cover [23], were widely used in previous studies. Recently, mobile phones have become more prevalent and location-based services have appeared, such as online food delivery or car sharing, making city life more convenient. In the meantime, the rapid development of mobile location-based services has expanded the data sources for population spatialization [8,24]. Therefore, plenty of geospatial big data obtained from Internet websites became sources for population mapping [9,25,26]. Points of interest (POIs) data are typical geospatial vector points with detailed geographical coordinates provided by GPS and are usually shown on a digital map [17]. There are many kinds of amenities in cities with regards to the categories of the POIs. This social sensing data enriches the traditional approaches of remote sensing data in densely populated urban areas [26,27]. Therefore, integrating multisource data including POIs can improve the accuracy of gridded population data.

Areal interpolation [11,13,28,29] and dasymetric mapping [17,28,30] have been widely applied for the spatial decomposition of census data. To create the weight layer, some linear regression models have been proposed [5,31]. In particular, the random forest, one machine learning method, has been incorporated into models to produce the weighting layer of population data. For the first time, scholars obtained 100 m grid population mapping results in three undeveloped countries [20]. Due to the ease of training and interpretation, the random forest algorithm has received increasing attention in the population mapping research [7,8,20]. Despite some progress, the internal mechanism between the multisource data and the population estimation is still unclear [32]. Furthermore, in terms of the novel machine learning method, further research is required to cope with the impacts of auxiliary variables on population predictions [17].

This study applied the random forest regression model to produce the 30 m resolution gridded population data of Chongqing, Southwest China. The resolution was finer than those in many previous studies. The main aims of the study were (1) to spatially disaggregate subdistrict level census data into 30 m impervious grid cells and (2) to analyze the relation between geospatial features and the population.

2. Materials and Methods

2.1. Study Area

The research was conducted in Chongqing, China (Figure 1). Located in the southwest of the country, Chongqing is the regional economic center for the upper reaches of the Yangtze River region, attracting a large sized migrant population. Since it was selected as one of directly controlled municipalities under China's State Council in 1997 [33], Chongqing has witnessed population and economic booms in the last two decades [34]. Nine central urban districts (Yuzhong, Yubei, Jiangbei, Shapingba, Nan'an, Dadukou, Jiulongpo, Banan and Beibei) in Chongqing were selected for this research (Figure 1). On one hand, the nine districts have more than 9.25 million permanent residents by the end of 2018, accounting for approximately 29.82% of the total resident population in Chongqing. On the other hand, the total area of the nine districts is approximately 5472.68 km², which is only 6.64% of the whole Chongqing area. Consequently, the population density is very high at 1690 persons/km².

Figure 1. Locations of the nine central districts and Chongqing in China.

The central urban areas in Chongqing are famous for their mountainous and watery landscapes [35]. The average elevation is over 391 meters, with approximately 22.9% of the land covered by hills and mountains. Chongqing is also at the confluence of the Yangtze River and Jialing River. Its topographic features are extremely complicated. The region is in the subtropical monsoon climate with large seasonal variations in precipitation and temperature [36]. The relatively complex natural environment results in different levels of economic development and a heterogeneous population distribution, making Chongqing a typical place to research gridded populations.

2.2. Datasets

Six datasets were collected from government agencies, international organizations and research institutions and applied in this study (Table 2). All the data were at the high spatial resolution in the most recent year of 2018, except for the ASTRE GDEM data that were from 2010. The administrative units of the research area include 9 districts and 162 subdistricts. The high-quality and recent 162 subdistricts' census data were the main data source for the disaggregation. The community level population survey data were used for the accuracy assessment. Given that the population distribution data showed a long-tail, the natural logarithm transformation was applied for the population density to get a normal and even distribution [20].

Table 2. List of the datasets used in this study.

Dataset	Format	Time	Source
Census data at the subdistrict and community levels	Table	2018	Chongqing Bureau of Public Security, China
Administrative Boundary	Vector (polygon)	2018	Chongqing Planning and Design Institute, China
Points of interest	Vector (point)	2018	Baidu, Inc., China
Luoja1-01 images	Raster (130 m)	2018	Wuhan University, China
Water	Vector (polygon)	2018	Chongqing Planning and Design Institute, China
Impervious surface	Vector (polygon)	2018	Chongqing Planning and Design Institute, China
ASTER GDEM	Raster (30 m)	2010	American NASA Jet Propulsion Lab and Japan's Ministry of Economy, Trade and Industry

The rest of the five auxiliary spatial datasets, including the POIs, Luoja1-01 image, water, impervious surface and ASTER GDEM data, were used to guide the fine-resolution population mapping. The POI data contained different kinds of information, such as latitudes, longitudes, addresses, classifications and others, which can be regarded as tags. By using Baidu Map's application programming interface under each category, the POI data were retrieved from Baidu, Inc. (<http://map.baidu.com>), which is one of the most popular web map service providers in China. The categories of the POI data were identified based on the Chinese semantic phrase according to Baidu Incand published articles [8,17,26]. Finally, a total of 813253 effective POI records were classified into 14 categories (Table 3) [8]. It is worth noting that single residence POI presented a residential community, including hundreds of apartments. In terms of car POIs, they included the points of sale, repair and service for cars. Baidu, Inc. divided them into three categories for business reasons while they were combined in this study.

Kernel density estimation (KDE) has the capability to convert each separated point into a smooth and continuous density surface. Most notably, the bandwidth of the KDE has a significant effect on the result [29]. KDE has been widely used to calculate the density of features in a surrounding neighborhood while considering the distance decay [37]. In this study, the parameter was adjusted from 100 m to 3000 m to get the best estimation after fitting the model. All the categories of the POI records were converted into raster layers by KDE, and 1000 m is chosen as the proper bandwidth. Due to the various distributions of the different POI categories, it is allowed to change; however, considering the principles of reproducibility and comparability, all the POIs use the same KDE bandwidth [17].

Table 3. Main categories of points of interest (POIs) retrieved for central Chongqing.

Category	Quantity	Percentage	Examples
Car	24,350	3.22%	Garage, car lot
Company	67,930	8.99%	Enterprise, factory
Education	28,622	3.79%	School, training agency
Finance	11,681	1.55%	Bank, insurance company
Government	16,816	2.22%	Police station, weather bureau
Hospital	25,871	3.42%	Pharmacy, clinic
Hotel	21,182	2.80%	Inn, lodge
Life service	123,309	16.31%	Post office, laundry
Public service	7081	0.94%	Newsstands, washroom
Residence	18,513	2.45%	Apartment, house
Restaurant	183,273	24.24%	Hotpot restaurant, tea house
Retail	203,418	26.91%	Supermarket, shopping center
Scenic spot	3459	0.46%	Monument, forest park
Sport	20,506	2.71%	Stadium, gymnasium

A new generation night-time light (NTL) remote sensing satellite, Luojia1-01, was launched in June 2018. The spatial resolution of Luojia1-01 NTL imagery is 130 m at the nadir point, which is the finest spatial resolution at present. Luojia1-01 has a better capability to simulate socioeconomic parameters than the Suomi National Polar-orbiting Partnership—Visible Infrared Imaging Radiometer (NPP-VIIRS) NTL data [38–40]. After spatial adjustment, radiation calibration was conducted for the Luojia1-01 images on the data website (<http://59.175.109.173:8888>). The final NTL layer was obtained using the following Equation (1):

$$L = \ln(DN^{3/2} + 1), \quad (1)$$

where L is the natural logarithm transformation value of the amplified radiant brightness; and DN is the digital number, which is the gray value of the initial image.

In terms of the natural variables, the Digital Elevation Model (DEM) gives the elevation of study area. The DEM was extracted from the Advanced Space-borne Thermal Emission and Reflection Radiometer Global Digital Elevation Model (ASTER GDEM) version 2 at a spatial resolution of 30 m. A surface analysis tool was used to calculate the slope and topographic relief. Water and impervious area data were obtained from a field investigation by the Chongqing Planning and Design Institute. The water based on vector map data were converted to raster format by calculating the Euclidean distance. The impervious layer was transformed to raster as the mask of the population surface.

All spatial covariates were projected to WGS_1984_UTM_Zone_48 and resampled to raster layers with the spatial resolution of 30 m. The zonal mean values of each continuous dataset within the subdistrict administrative boundary were calculated. The preprocessing methods were implemented using ArcMap 10.3 (ESRI, Inc., Redlands, CA, USA).

2.3. Methods

The multisource geospatial data were input into the dasymetric model due to the spatial heterogeneity of population distributions. To produce the spatial weighting layer, 10-fold cross validation was adopted to build a random forest regression model. Figure 2 represented the procedures for gridded population mapping, including the data processing, model building and result analysis.

Figure 2. Workflow of disaggregating the census population into 30-m-grid cells.

The random forest (RF) is an ensemble learning methodology, which was first systematically proposed by Breiman [41]. Among the essential ingredients, both bagging and the Classification and Regression Trees (CART) criteria play critical roles [41,42]. This machine learning technique is made of a combination of randomly generated decision trees, which can be used for classification and regression and achieve a significant improvement [43]. A decision tree is a nonlinear model to calculate outcomes using many linear boundaries by asking a series of queries about the available feature until reaching the final decision [44]. Therefore, the random forest model based on the spatial segmentation tree is not sensitive to the scale. For the regression, the final output is the mean prediction of the discrete trees that correspond to the input variables. Because the response variable was consecutive, the random forest regression model was adopted to fit the data in this study.

In the process of machine learning, it is necessary to optimize the super parameters and select a group of optimal super parameters for the learners to improve the performance and effect of learning [44,45]. For example, the max number of features and the max depth of trees are the two super parameters in random forest model. Grid search is conducted to try every possibility through loop traversal in all candidate parameter selections. The best parameters are used in the later analysis. To improve the predictive accuracy and control overfitting, the maximum depth and features of the decision tree were applied to stop excessive splitting. Moreover, the assessment accuracy is sensitive to the number of decision trees. If the number of trees is too small, it is easy to underfit the model. Conversely, more decision trees are likely to increase the calculation load. At the proper number, the model reaches a near steady state.

To avoid the impact of the initial data partition on the results, cross validation was performed to reduce the occasionality. In general, grid search needs to be combined with cross validation to ensure the validity and accuracy of the test. Therefore, cross validation (CV) was used to obtain these 3 fitting parameters [46]. When the number of samples was small, K-fold CV was helpful [47]. The K-1-fold was used as the training data and the outcome was validated by the remaining fold. In this research, there were only 162 subdistricts involved in the model and therefore 10-fold CV was adopted. The model estimation was completed using the Python 3.5.2 statistical environment and the Scikit-learn 0.21.3 random forest package.

To predict the weight of the population density, the natural logarithm of the population density and all the zonal mean values of the features in each subdistrict administrative unit were used to fit the random forest model. The impervious mask was used to abstract the final weighting layer. The dasymetric population distribution map for central Chongqing was produced using the following Equation (2):

$$POP_{pixel} = \frac{POP_{sub-district} \times W_{pixel}}{W_{sum}}, \quad (2)$$

where POP_{pixel} is the predicted population for each pixel, $POP_{sub-district}$ is the census population of every subdistrict, W_{pixel} is the population distribution weight of each grid and W_{sum} is the weighted sum within the subdistrict boundary.

Given that the latest census data from 2018 were used, the current datasets were unsuited to compare with this research since that most of datasets were based upon the 2010 census. Therefore, the community level population data from an authoritative investigation in the year of 2018 were used. Due to the population data confidentiality on the community scale, the assessment dataset only consisted of 200 communities in central Chongqing, which were randomly chosen by the Chongqing Bureau of Public Security. Three regression models were compared, including the multiple linear regression (MLR), the geographically weighted regression (GWR) and the random forest regression (RFR).

MLR models assume that the sample data are independent of each other while the population data have the geospatial dependence. Many attempts were made to make MLR models suitable for spatial data [48–50]. Brunson et al. (1996) first proposed the GWR [48]. Considering the spatial location, the regression coefficients of different auxiliary variables are not evaluated by using all sample data, but rather by using a certain range of the local subset of the data [49,50]. The GWR was used for demographic mapping [51] and the identification of households in potential economic distress [52].

Because different communities have distinct population densities, it was assumed that this model had different performances in communities with various population densities. The accuracy of these communities with different population densities were tested using the coefficient of determination (R^2), the root mean squared error (RMSE) and the mean absolute error (MAE), which are well-known accuracy criteria in the field of population estimation. Additionally, the Akaike information criterion (AIC) can estimate the relative amount of information lost by a statistical model. The smaller the value is, the better the model's fit [53,54]. In general, when the number of samples is small, the corrected Akaike information criterion (AICc) is more helpful than the AIC since it avoids potential overfitting [55]. Due to small samples of 162 subdistricts' populations, the AICc was adopted as the model fit criterion.

Feature importance plays an important role in the explanations of the random forests model and it measures the increase in the prediction error of the model after the values of the variable were randomly permuted [56]. Considering that shuffling feature values increases the model error, the feature is important [57]. Using the Python Scikit learn 0.21.3 package, the feature importance of the random forest regressions was calculated as the impurity value reduction using the mean squared error [58] in the following Equation (3):

$$\text{Impurity value} = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 \quad (3)$$

where y_i is the size of the population in each subdistrict, N is the number of subdistricts and μ is the mean value of all subdistricts.

The partial dependence plot (PDP) is one of the smartest ways to extract insights from complicated models [17,56,59]. The same as the function of the coefficients in the linear model, the PDP can be used to interpret machine learning models. After fitting the model on real data, the features were changed to make a series of predictions by averaging the impacts of all other variables. The PDP shows the

change of the average estimation of the population density with the varying features. Using the Python Pdpbox package, the PDP was formed using the following Equation (4):

$$\text{Change} = \frac{1}{N} \sum_{i=1}^N F(x_{i,k}, \dots, x_{i-1,k}, x, x_{j+1,k}, \dots, x_{p,k}) \quad (4)$$

where N is the number of observations of the response variables; $x_{i,k}$ denotes the covariates for features $i = 1, 2, \dots, p$ and samples $k = 1, 2, \dots, N$; and $F(x)$ is the mathematical function of the random forest regression model.

3. Results

3.1. Population Estimation

Due to only 162 subdistricts being used to fit the model, 10-fold CV was chosen. 9-folds of samples were adopted into the model to obtain the proper parameters and the accuracy was assessed using the remaining samples. 10-fold cross validation and grid search were conducted to find the proper parameters in the RF (Figure 3). When the max depth of the decision tree was 4 and the max number of features was 12, the average correlation coefficient (R^2) of the CV score reached the maximum value of 0.8861.

Figure 3. The ten-fold cross validation for the proper parameters in the random forest.

The suitable number of estimators was tested in the random forest and it ranged from 1 to 300. When the number of estimators reached 50, the accuracy of the model was 88% (Figure 4). After that, the accuracy was stable with small fluctuations. Therefore, 50 trees were sufficient in the random forest regression model. The max depth of the decision trees and the max number of features were 4 and 12, respectively, when the final tuning parameters were adopted.

Figure 4. Relationship between model accuracy and the number of estimators.

After fitting the random forests regression model, the final distribution weight of each raster was produced (Figure 5a). The weights where only impervious surface cover existed were used to disaggregate the census population to each pixel at the subdistrict level (Figure 5b).

Figure 5. (a) Distribution weights of the population at a 30 m spatial resolution and (b) census population at the subdistrict level in central Chongqing.

In the final population estimated layer (Figure 6), the population presented a lumpy or banded distribution, concentrating around Yuzhong in central Chongqing. From the high-resolution photomicrograph disaggregation results (Figure 6a,b), more details were able to be captured.

Figure 6. Predicted population per grid cell in central Chongqing in 2018. (a) The photomicrograph of Zhongliang Mountain. (b) The photomicrograph of Yuzhong Peninsula.

3.2. Accuracy Validation

When comparing the results obtained by the three different models, GWR had a smaller Akaike information criterion than the MLR and the RFR model had the highest R^2 value (0.7469) (Table 4). The RMSE and MAE of the RFR were also smaller (2785.04 and 1889.70, respectively). Therefore, the RFR had better overall accuracy than the other two linear regression models. The RFR model did improve the estimation.

Table 4. Comparative reliability assessment results.

Methodology	Akaike Criterion	Validation R^2	RMSE	MAE
MLR	366.78	0.7199	2944.25	2056.81
GWR	357.29	0.7221	2939.09	2042.35
RFR	—	0.7469	2785.04	1889.70

To estimate the prediction precision in communities with different population densities, all the communities were divided into three groups according to the density of the community's population. In Figure 7, the green plus markers indicated the communities with low population densities below 150 persons per hectare. It is worth mentioning that the coefficient of determination (R^2) of this group was 0.8498, demonstrating the high accuracy of the results. The R^2 of the communities with medium population densities from 150 to 300 persons per hectare (the blue square markers) was 0.6729 and the R^2 of the communities with population densities above 300 persons per hectare (the red triangle markers) was 0.4903. For all the communities, the overall accuracy was acceptable ($R^2 = 0.7469$).

Figure 7. Scatterplot of the predicted and census populations of the communities.

4. Discussion

4.1. Interpretation of the Gridded Population

As marked in the green box in Figure 8, the densities of life service, restaurant and residence POIs were the top 3 most important features. Their importance values exceeded 0.15, reflecting that the model had learned to take advantage of those POI data. The density of life service POIs, such as post offices and barbershops, was the most important feature influencing the population mapping and had the highest feature importance score. Restaurants and residences were also of vital significance for everyday life and so they were closely related to population distribution. The feature importance values below 0.01 were shown in the red box. When the density of government POIs was high, it was usually the administrative office land in the city, corresponding to small populations. When the densities of car, transport and scenic spot POIs were high, they were industrial land, transportation facilities land and green land, respectively. Surprisingly, the brightness of the NTL corresponded poorly to the population distribution. The natural factors, such as the elevation, the slope and the distance to the river, were also of little influence, probably due to the large disturbances of anthropogenic activities on the natural environment.

Figure 8. Feature importance for the random forest regression model.

Figure 9 shows the complex influence of features on the predicted weight of the population distribution. Since the kernel density of POIs strongly affected predictions strongly, the top 3 features were analyzed, including the kernel density of the life service, restaurant and residence POIs. Then, another three variables were chosen, including the brightness of the NTL, the elevation and the distance to the river. The larger densities of these POIs indicated denser populations

of these three kinds of POIs (life service, residence and restaurant) (Figure 9a–c, respectively). However, this pattern disappeared for the brightness of the NTL, the elevation and the distance to the river (Figure 9d–f, respectively).

Figure 9. Partial dependency plots (PDP) for the features: (a) the kernel density of life service points of interest (POIs), (b) the kernel density of residence POIs, (c) the kernel density of restaurant POIs, (d) the brightness of night time light (NTL), (e) elevation and (f) distance to the river in the random forest regression model to estimate the population redistribution weight. Dark blue bar at the bottom indicates the deciles (11.1% quantiles) of the observed distribution of the continuous prediction features. The blue shaded area indicates the level of confidence.

Regarding life service POIs, when the density increased from 0.06 to 7.25, the change of the population weight increased from 0 to 0.6, respectively (Figure 9a). Due to numerous kinds of life services, the city attracted many people who wanted to work and live. The more life services facilities that were constructed, the more convenient city life was. After 7.25, the change remained constant, indicating the limited impact of extra life service facilities after the threshold.

Regarding the density of residence POIs, when the density was below 14.62, the change was near to zero. First, the population density was relatively low in the suburbs with scattered residences. Second, people pursued healthy and comfortable living environments and the low density residential development provided a good trade-off. When the density of the points exceeded 14.62, the change increased with some stable values. The higher density of residence POIs implied more people lived around that point.

Regarding the high density of restaurant POIs, the model predicted a generally high population size. For one thing, there was no change when the density was below 7.4, indicating small numbers of restaurants in places with few people. For another, diverse businesses, including restaurants, were located close to the central business district (CBD), attracting many people to live near the CBD. Thus, when the density exceeded 7.4, the change rose rapidly.

Regarding the brightness of the NTL graph (Figure 9d), in agreement with previous studies [21], the light was so dim that little energy could be captured by the sensors in the sparsely populated countryside. Therefore, the difference was not notable when the brightness of the NTL was below 3.95. Afterwards, the high brightness of the NTL increased the probability of population aggregation, and it was saturated at 5.24. First, the NTL was closely linked to human activity. Second, the high brightness of the NTL reflected more details of transportation and industrial activities than those of the population distribution. Therefore, the change of the confidence was from -0.02 to 0.04 , reflecting the dual characters.

The local natural environment had a very limited impact on the population distribution, thanks to the progress of technology. Thus, the predicted population size usually remained constant (Figure 9e,f). Interestingly, there was a modest rise when the elevation increased from 449.77 to 766.17 m. Meanwhile, the confidence interval was between -0.025 and 0.075 , which implied both positive and negative effects. Because of the limited training data, the prediction accuracy of the random forest model for this variable may be affected.

In terms of the distance to the river, the prediction was nearly constant when the distance changed from 29.18 to 835.71 m. As the distance further increased, the change decreased below 0 and finally went down to -0.4 . Heterogeneous effects may not be revealed since the PDP plots only showed the average marginal effects [46]. For one thing, people lived near to rivers to get beautiful natural views; thus, this feature had a positive association with the prediction. For another, people remained a distance from the river in order to avoid the threat of natural flood disasters. Henceforth, the PDP of the distance to the river was a horizontal line with both positive and negative confidence ranges since the two effects could counteract each other.

4.2. Principal Findings and Meaningful Implication

This study produced the latest 30 m gridded population data in central Chongqing using contemporary data publicly available at the time of the analysis in 2018. The results provide timely and important information to policy makers for city planning, natural disaster and disease control and environment protection [3,22,26,60,61]. The state-of-the-art random forest model was used to integrate multiple data and 10-fold cross validation was conducted to produce the weighting layer, which was used for the dasymetric population mapping. As some of the most popular social sensing data, POI data are helpful for getting high accuracy gridded population results (overall $R^2 = 0.7469$, $RMSE = 2785.04$ and $p < 0.01$), especially after the kernel density estimation was used to generate various kinds of POI surfaces [27]. With the help of the feature importance analysis and the PDP, the results indicate that POI data had a more crucial effect on population mapping than the remote

sensing data, like the LuoJia1-01 NTL images, in cities [26]. However, the natural variables had both negative and positive impacts on population spatialization. In terms of a large scale study, the NTL images were major auxiliary data for the estimation [62–64]. Despite the finest resolution of the NTL images being 150 meters [21], the complex urban fabric patterns and the uncertainty of the radiation calibration of LuoJia1-01 constrain the application of the images in conducting accurate population estimations at a fine scale.

4.3. Limitations and Future Research

Similar to many studies, there are some limitations in the current study. In terms of data, our study mainly focused on POIs, and it may reflect the land use as well. In fact, the previous studies demonstrated that POI data could be used for the classification of urban land use [59,65,66]. Meanwhile, after being divided into 14 categories, all the data were fitted for the prediction model without integrating all the POIs into one layer. Consequently, the various kinds of POIs may be correlated with each other, which may cause some feature importance bias. Further studies must deal with the independence of each feature.

Because of the confidentiality of the population data at the fine spatiotemporal scale, this study only chose central Chongqing as a case study instead of applying the model to the whole China. The POI data in the current study were provided by Baidu, Inc. Outside China, POI data could be extracted from the open access website Open Street Map (OSM) (<https://www.openstreetmap.org/#map=6/-0.396/32.893>) [67]. Open Street Map POIs are one of the most popular volunteered geographic information (VGI) data since they are open source and easy to access [17,25]. However, the data quality is uncertain. First, most of the POIs are reported in urban areas and many POIs in rural areas are not collected. Second, the positional and thematic accuracy also need to be improved [68]. To address these issues, we adopted commercial POI data, which were collected using rigorous processes and found to be reliable [8,17].

With regard to the random forest model, the results depended on the prediction of each tree, which added randomness to the calculation. When the model was repeated, the results may vary greatly [20]. In a future study, attention should be paid to averaging the results over repetitions in order to stabilize the measurements, as well as controlling the time of the modeling process, as computing power continues to develop.

The calibration and validation of the population mapping was tough work due to population mobility and inadequate data. The statistical population counts conducted by the government are generally considered to be high quality, although they still have inherent limitations [1]. We are unable to compare our results with China's national census data in 2010, but the local community investigation data from 2018 are available. The comparison indicated various accuracies in different parts of the city because of the changes of natural and social-economic situations. In the communities with relative larger errors, more efforts are needed to apply alternative ways to build new models with proper parameters.

5. Conclusions

This study provided a dasymetric model for downscaling the census population in central Chongqing, Southwest China. This model took advantage of the smart random forest algorithm to generate the predicted population layer at a 30 m resolution. It offered a reference for multisource data fusion and the proper interpretation of the output. The POI data were adopted to overcome the overestimation and underestimation of the remote sensing images. According to the accuracy validation at the community level, the proposed approach achieved acceptable overall accuracy (overall $R^2 = 0.7469$, $RMSE = 2785.04$ and $p < 0.01$). Furthermore, social sensing data and the random forest regression can be used to disaggregate socioeconomic data and the results are helpful for policy makers.

Author Contributions: Conceptualization and methodology, Mingguo Ma, Kaifang Shi and Yun Zhou; software, validation and formal Analysis, Zhenyu Peng and Yun Zhou; writing—original draft preparation, Yun Zhou; writing—review & editing, Mingguo Ma, Kaifang Shi; funding acquisition, Mingguo Ma. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly supported by the National Natural Science Foundation of China project (Grant Number: 41830648, 41771453), the Chongqing R&D Project of the high technology and major industries (Grant Number: [2017] 1231).

Acknowledgments: In this study, multi-resource data were downloaded from different data sources. The NTL data downloaded from the High-Resolution Earth Observation System of the Hubei Data and Application Center. DEM data were downloaded from the US Land Processes Distributed Active Archive. The population and water and impervious area polygon datasets were provided by the Chongqing Planning and Design Institute. The authors express their gratitude for the data sharing of above datasets.

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

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. Balk, D.; 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. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Weber, E.; Seaman, V.Y.; Stewart, R.; Bird, T.J.; Tatem, A.J.; McKee, J.; Bhaduri, B.; Moehl, J.J.; Reith, A.E. Census-independent population mapping in northern Nigeria. *Remote Sens. Environ.* **2018**, *204*, 786–798. [\[CrossRef\]](#) [\[PubMed\]](#)
4. O'Neill, B.C.C.; Dalton, M.; Fuchs, R.; Jiang, L.; Pachauri, S.; Zigova, K. Global demographic trends and future carbon emissions. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 17521–17526. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Gaughan, A.E.; Stevens, F.R.; Linard, C.; Jia, P.; Tatem, A.J. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **2013**, *8*, e55882. [\[CrossRef\]](#)
6. Su, H.; Wei, H.; Zhao, J. Density effect and optimum density of the urban population in China. *Urban Stud.* **2016**, *54*, 1760–1777. [\[CrossRef\]](#)
7. Deville, P.; Linard, C.; Martin, S.; Gilbert, M.; Stevens, F.R.; Gaughan, A.E.; Blondel, V.D.; Tatem, A.J. Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15888–15893. [\[CrossRef\]](#)
8. 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. [\[CrossRef\]](#)
9. 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. [\[CrossRef\]](#)
10. Wardrop, N.A.; Jochem, W.C.; Bird, T.J.; Chamberlain, H.R.; Clarke, D.; Kerr, D.; Bengtsson, L.; Juran, S.; Seaman, V.; Tatem, A.J. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3529–3537. [\[CrossRef\]](#)
11. Goodchild, M.F.; Anselin, L.; Deichmann, U. A framework for the areal interpolation of socioeconomic data. *Environ. Plan A Econ. Space* **1993**, *25*, 383–397. [\[CrossRef\]](#)
12. Doxey-Whitfield, E.; MacManus, K.; Adamo, S.B.; Pistolesi, L.; Squires, J.; Borkovska, O.; Baptista, S.R. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. *Pap. Appl. Geogr.* **2015**, *1*, 226–234. [\[CrossRef\]](#)
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. 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.
15. Tatem, A.J. Comment: WorldPop, open data for spatial demography. *Sci. Data* **2017**, *4*, 170004. [\[CrossRef\]](#)
16. Reed, F.J.; Gaughan, A.E.; Stevens, F.R.; Yetman, G.; Sorichetta, A.; Tatem, A.J. Gridded Population Maps Informed by Different Built Settlement Products. *Data* **2018**, *3*, 33. [\[CrossRef\]](#)
17. Ye, T.; Zhao, N.; Yang, X.; Ouyang, Z.; Liu, X.; Chen, Q.; Hu, K.; Yue, W.; Qi, J.; Li, Z.; 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. [\[CrossRef\]](#)

18. 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. [\[CrossRef\]](#)
19. 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. [\[CrossRef\]](#)
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. Gao, N.; Li, F.; Zeng, H.; Van Bilsen, D.; De Jong, M. Can More Accurate Night-Time Remote Sensing Data Simulate a More Detailed Population Distribution? *Sustainability* **2019**, *11*, 4488. [\[CrossRef\]](#)
22. Mossoux, S.; Kervyn, M.; Soulé, H.; Canters, F. Mapping Population Distribution from High Resolution Remotely Sensed Imagery in a Data Poor Setting. *Remote Sens.* **2018**, *10*, 1409. [\[CrossRef\]](#)
23. Zeng, C.; Zhou, Y.; Wang, S.; Yan, F.; Zhao, Q. Population spatialization in China based on night-time imagery and land use data. *Int. J. Remote Sens.* **2011**, *32*, 9599–9620. [\[CrossRef\]](#)
24. 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.* **2019**, *33*, 687–706. [\[CrossRef\]](#)
25. Langford, M. An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data. *Geogr. Anal.* **2013**, *45*, 324–344. [\[CrossRef\]](#)
26. 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. [\[CrossRef\]](#)
27. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [\[CrossRef\]](#)
28. Lin, J.; Cromley, R.G. Evaluating geo-located Twitter data as a control layer for areal interpolation of population. *Appl. Geogr.* **2015**, *58*, 41–47. [\[CrossRef\]](#)
29. Shi, X.; Li, M.; Hunter, O.; Guetti, B.; Andrew, A.; Stommel, E.; Bradley, W.; Karagas, M.R. Estimation of environmental exposure: Interpolation, kernel density estimation or snapshotting. *Ann. GIS* **2019**, *25*, 1–8. [\[CrossRef\]](#)
30. Qiu, F.; Cromley, R. Areal Interpolation and Dasymetric Modeling. *Geogr. Anal.* **2013**, *45*, 213–215. [\[CrossRef\]](#)
31. Azar, D.; Graesser, J.; Engstrom, R.; Comenetz, J.; Leddy, R.M.; Schechtman, N.G.; Andrews, T. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. *Int. J. Remote Sens.* **2010**, *31*, 5635–5655. [\[CrossRef\]](#)
32. Li, J.; Li, J.; Yuan, Y.; Li, G. Spatiotemporal distribution characteristics and mechanism analysis of urban population density: A case of Xi'an, Shaanxi, China. *Cities* **2019**, *86*, 62–70. [\[CrossRef\]](#)
33. Liu, Y.; Fan, P.; Yue, W.; Song, Y. Impacts of land finance on urban sprawl in China: The case of Chongqing. *Land Use Policy* **2018**, *72*, 420–432. [\[CrossRef\]](#)
34. Bao, H.X.H.; Li, L.; Lizieri, C. City profile: Chongqing (1997–2017). *Cities* **2019**, *94*, 161–171. [\[CrossRef\]](#)
35. Cheng, L.; Guan, D.; Zhou, L.; Zhao, Z.; Zhou, J. Urban cooling island effect of main river on a landscape scale in Chongqing, China. *Sustain. Cities Soc.* **2019**, *47*, 101501. [\[CrossRef\]](#)
36. Chen, S.; Yan, Y.; Liu, G.; Fang, D.; Wu, Z.; He, J.; Tang, J. Spatiotemporal characteristics of precipitation diurnal variations in Chongqing with complex terrain. *Theor. Appl. Clim.* **2018**, *137*, 1217–1231. [\[CrossRef\]](#)
37. Silverman, B.W. Kernel Density Estimation Using the Fast Fourier Transform. *Appl. Stat.* **1982**, *31*, 93. [\[CrossRef\]](#)
38. Jiang, W.; He, G.; Long, T.; Guo, H.; Yin, R.; Leng, W.; Liu, H.; Wang, G. Potentiality of Using LuoJia 1-01 Nighttime Light Imagery to Investigate Artificial Light Pollution. *Sensors* **2018**, *18*, 2900. [\[CrossRef\]](#)
39. Ou, J.; Liu, X.; Liu, P.; Liu, X. Evaluation of LuoJia 1-01 nighttime light imagery for impervious surface detection: A comparison with NPP-VIIRS nighttime light data. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *81*, 1–12. [\[CrossRef\]](#)
40. Wang, C.; Chen, Z.; Yang, C.; Li, Q.; Wu, Q.; Wu, J.; Zhang, G.; Yu, B. Analyzing parcel-level relationships between LuoJia 1-01 nighttime light intensity and artificial surface features across Shanghai, China: A comparison with NPP-VIIRS data. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *85*, 101989. [\[CrossRef\]](#)
41. Breiman, L.E.O. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
42. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [\[CrossRef\]](#)

43. Robinson, R.L.M.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792. [[CrossRef](#)] [[PubMed](#)]
44. Criminisi, A.; Shotton, J.; Konukoglu, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends Comput. Graph. Vis.* **2011**, *7*, 81–227. [[CrossRef](#)]
45. Boulesteix, A.-L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [[CrossRef](#)]
46. Oh, Y.J.; Park, H.S.; Min, Y. Understanding location-based service application connectedness: Model development and cross-validation. *Comput. Hum. Behav.* **2019**, *94*, 82–91. [[CrossRef](#)]
47. Gholinejad, S.; Naeini, A.A.; Amiri-Simkooei, A.R. Robust Particle Swarm Optimization of RFMs for High-Resolution Satellite Images Based on K-Fold Cross-Validation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2019**, *12*, 2594–2599. [[CrossRef](#)]
48. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geogr. Anal.* **1996**, *28*, 281–298. [[CrossRef](#)]
49. Brunsdon, C.; Fotheringham, S.; Charlton, M. Geographically Weighted Regression. *J. R. Stat. Soc. Ser. D (Stat.)* **1998**, *47*, 431–443. [[CrossRef](#)]
50. McMillen, D.P. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. *Am. J. Agric. Econ.* **2004**, *86*, 554–556. [[CrossRef](#)]
51. Matthews, S.A.; Yang, T.-C. Mapping the results of local statistics: Using geographically weighed regression. *Demogr. Res.* **2012**, *26*, 151–166. [[CrossRef](#)] [[PubMed](#)]
52. Benassi, F.; Naccarato, A. Households in potential economic distress. A geographically weighted regression model for Italy, 2001–2011. *Spat. Stat.* **2017**, *21*, 362–376. [[CrossRef](#)]
53. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **1974**, *19*, 716–723. [[CrossRef](#)]
54. Oshan, T.M.; Li, Z.; Kang, W.; Wolf, L.J.; Fotheringham, A.S. MGWR: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 269. [[CrossRef](#)]
55. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [[CrossRef](#)]
56. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
57. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *Am. Stat.* **2009**, *63*, 308–319. [[CrossRef](#)]
58. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **2017**, *27*, 659–678. [[CrossRef](#)]
59. Zhao, Q.; Hastie, T. Causal Interpretations of Black-Box Models. *J. Bus. Econ. Stat.* **2019**, 1–10. [[CrossRef](#)]
60. Linard, C.; Alegana, V.A.; Noor, A.; Snow, R.W.; Tatem, A.J. A high resolution spatial population database of Somalia for disease risk mapping. *Int. J. Health Geogr.* **2010**, *9*, 45. [[CrossRef](#)]
61. Yang, H. China must continue the momentum of green law. *Nature* **2014**, *509*, 535. [[CrossRef](#)] [[PubMed](#)]
62. Amaral, S.; Monteiro, A.M.V.; Camara, G.; Quintanilha, J.A. DMSP/OLS night-time light imagery for urban population estimates in the Brazilian Amazon. *Int. J. Remote Sens.* **2006**, *27*, 855–870. [[CrossRef](#)]
63. Townsend, A.C.; Bruce, D. The use of night-time lights satellite imagery as a measure of Australia’s regional electricity consumption and population distribution. *Int. J. Remote Sens.* **2010**, *31*, 4459–4480. [[CrossRef](#)]
64. Li, X.; Zhou, W. Dasymeric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data. *Sci. Total. Environ.* **2018**, *643*, 1248–1256. [[CrossRef](#)]
65. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* **2015**, *53*, 36–46. [[CrossRef](#)]
66. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [[CrossRef](#)]

67. Arsanjani, J.J.; Helbich, M.; Bakillah, M.; Hagenauer, J.; Zipf, A. Toward mapping land-use patterns from volunteered geographic information. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2264–2278. [[CrossRef](#)]
68. Ma, D.; Sandberg, M.; Jiang, B. Characterizing the Heterogeneity of the OpenStreetMap Data and Community. *ISPRS Int. J. Geo Inf.* **2015**, *4*, 535–550. [[CrossRef](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).