land

MDPI

Article

Optimization of Modelling Population Density Estimation Based on Impervious Surfaces

Jinyu Zang [1], Ting Zhang [1,*], Longqian Chen [1], Long Li [1,2], Weiqiang Liu [3], Lina Yuan [4], Yu Zhang [5], Ruiyang Liu [3], Zhiqiang Wang [1], Ziqi Yu [1] and Jia Wang [1]

1 School of Public Policy and Management, China University of Mining and Technology, Daxue Road 1, Xuzhou 221116, China; jinyu.zang@cumt.edu.cn (J.Z.); chenlq@cumt.edu.cn (L.C.); long.li@cumt.edu.cn or long.li@vub.be (L.L.); zq.wang@cumt.edu.cn (Z.W.); ziqi.yu@cumt.edu.cn (Z.Y.); jia.wang@cumt.edu.cn (J.W.)
2 Department of Geography, Earth System Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
3 School of Environmental Science and Spatial Informatics, China University of Mining and Technology, Daxue Road 1, Xuzhou 221116, China; weiqiang.liu@cumt.edu.cn (W.L.); ruiyang.liu@cumt.edu.cn (R.L.)
4 Key Laboratory of Geographic Information Science (Ministry of Education), School of Geographic Sciences, East China Normal University, Shanghai 200241, China; lnyuan@cumt.edu.cn
5 Department of Land Resource Management, School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China; yuzhang@jsnu.edu.cn
* Correspondence: tingzhang@cumt.edu.cn; Tel.: +86-516-8359-1327

Abstract: Population data are key indicators of policymaking, public health, and land use in urban and ecological systems; however, traditional censuses are time-consuming, expensive, and laborious. This study proposes a method of modelling population density estimations based on remote sensing data in Hefei. Four models with impervious surface (IS), night light (NTL), and point of interest (POI) data as independent variables are constructed at the township scale, and the optimal model was applied to pixels to obtain a finer population density distribution. The results show that: (1) impervious surface (IS) data can be effectively extracted by the linear spectral mixture analysis (LSMA) method; (2) there is a high potential of the multi-variable model to estimate the population density, with an adjusted R^2 of 0.832, and mean absolute error (MAE) of 0.420 from 10-fold cross validation recorded; (3) downscaling the predicted population density from the township scale to pixels using the multi-variable stepwise regression model achieves a more refined population density distribution. This study provides a promising method for the rapid and effective prediction of population data in interval years, and data support for urban planning and population management.

Keywords: population estimation; impervious surface; stepwise regression; remote sensing; Hefei

Citation: Zang, J.; Zhang, T.; Chen, L.; Li, L.; Liu, W.; Yuan, L.; Zhang, Y.; Liu, R.; Wang, Z.; Yu, Z.; et al. Optimization of Modelling Population Density Estimation Based on Impervious Surfaces. *Land* **2021**, *10*, 791. https://doi.org/10.3390/land10080791

Academic Editors: Dimitris Skuras, Panayotis Dimopoulos and Ioannis P. Kokkoris

Received: 5 July 2021
Accepted: 27 July 2021
Published: 28 July 2021

1. Introduction

Population data are considered important indices for the development of a country or region. Urban or land planning, policymaking, public emergencies, and other public aspects require detailed population data [1,2]. Currently, most countries obtain population data through the census. However, accessing the population data is difficult because of the dispersion and dynamics of the population distribution [3,4]. In China, the census is the most common method for collecting population data. Although the census is relatively credible, it is time-consuming, costly, and obstructed by difficulties in reality [5]. Nowadays, population mobility makes it difficult to obtain actual residence figures. People are more protective of their privacy. It is very common to refuse to fill in personal information. The situations above increase the difficulty of the census. Moreover, the national census is conducted once every ten years, which may lead to the absence of population data in interval years. Consequently, scientific decision making lacks support.

Unlike the national census, remote sensing data are not subject to time restrictions [6–9]. Owing to the rapid development of remote sensing and computer technologies, some

scholars have applied these to estimate the population [10–12], enabling the possibility of mapping populations at the pixel scale. The commonly used remote sensing images are as follows: high-resolution images, such as IKONOS, QuickBird, and Worldview images, and moderate-resolution images, such as the Landsat series, hyperspectral, and radar images [2,13,14]. Moreover, land-use types and building areas can be extracted from remote sensing images for population estimation [15–17]. Among the abovementioned remote sensing images, Landsat satellites are the most widely used because of their free access and relatively high spatial and temporal resolutions [2,16].

Impervious surfaces (ISs), as information related to the population, can be extracted from remote sensing images. ISs comprise the surface of artificial buildings, which are closely related to human activities [4]. This concept was introduced in 1996, defining an IS as any material that prevents the flow of water into the soil, including both artificial structures and natural substances (such as hard bare rocks) [18]. Currently, construction areas are continuously expanding because of human activities. Thus, ISs were redefined as artificial surfaces, such as roofs, asphalt or cement roads, parking lots, and other waterproof surfaces [19–21]. Existing studies have used IS data to estimate populations with old census data, such as in 2000 and 2010, most of which are at the county scale [3–5,22–25].

There are several methods for extracting IS data from Landsat satellites, including the manual interpretation, classification, and spectral analysis methods [26–29]. The first method requires prior knowledge, and manual interpretation has a low efficiency [3]. The second method aims to extract the IS data at the pixel level to obtain the IS area. Among the spectral analysis methods, linear spectral mixture analysis (LSMA) can extract IS data at the sub-pixel scale, obtaining the proportion of an IS in a pixel [4]. The method has a high efficiency and accuracy. It has been applied in population estimation [3–5].

In addition to ISs, remote sensing technology increasingly contributes to the study of factors affecting population distribution. Recently, night light (NTL) data have been incorporated with other data sources to improve population estimation [9,30,31]. The existing NTL data include the DMSP/OLS NTL, NPP/VIIRS NTL, and Luojia-1 NTL [32–34]. The first two types of NTL data have a low resolution and sensitivity, particularly for low radiation brightness areas [34,35]. Luojia-1 NTL has the highest resolution (130 m) worldwide and can delineate the scope of human activities more accurately; thus, it has been applied for population estimation with good results. Luojia-1 NTL was provided by the Luojia-1 satellite which was developed by Wuhan University and related institutions [31,35,36].

Data with spatial-temporal information for scientific research are emerging in this era of big data, among which the point-of-interest (POI) is a series of points with location and time information [37]. POI data represent a practical space object crawled from map websites, such as the Baidu Map, and map service providers like OSM (See in https://en.m.wikipedia.org/wiki/Baidu_Maps, accessed on 26 July 2021). POI data have been shown to be closely related to the population's distribution [38,39]. Furthermore, the accessibility of fine-grained geographic factor data has been applied by many scholars in the process of population spatialization [7,38,40].

With the wide application of remote sensing and geographic information in population estimation, the population estimation model was established on the relationship between population and dwelling units [17]; (b) land use [2]; (c) built-up areas [4,5,22–25]; (d) spectral features of the image pixel [16]. The current research mainly focuses on built-up areas, in which ISs are typical representatives of urban areas. Since the methods for extracting ISs may lead to the confusion of ground objects [29], it is easy to lead to the estimation error of delineating population distributions with IS data as the sole independent variable [4].

The census is inefficient and time-consuming. Most regions all over the world lack population data in the interval years. The population estimation model established with ISs as a single variable has the problem of inaccurate delineation of population distribution areas. As mentioned above, a new method is needed to not only estimate populations accurately but also help to delineate population distributions. The main goal of the

presented study is to improve the population density estimation based on impervious surfaces. To better delineate the population distribution area, our study integrates POI data and NTL data with IS data to improve the model.

Therefore, in this study, the geo-spatiotemporal big data POI and the latest high-resolution NTL data from Luojia-1 were fused with the IS information extracted from remote sensing images to try to establish the population estimation model. With IS data as the main independent variable, the population distribution of Hefei was explored using the stepwise regression method. The optimal population estimation model was then applied to the pixel scale. Finally, a finer population distribution map was obtained.

2. Study Area

Hefei (30°57′–32°32′ N, 116°41′–117°58′ E; Figure 1) is situated in the middle of Anhui Province in China, covering a total area of 11,408.48 km^2 with an average altitude of 30 m. The terrain comprises plains and low hills. It is characterised by a subtropical monsoon climate with an annual mean daily temperature of 16 °C and a total annual precipitation of 995.2 mm. As the capital, Hefei is the cultural, commercial, financial, and political centre of Anhui Province. Located in the radiation belt of the Yangtze River Delta, it acts as a gateway for the development of the central and western regions. Meanwhile, as a node city of the 'One Belt, One Road' strategy, Hefei has great economic potential [41].

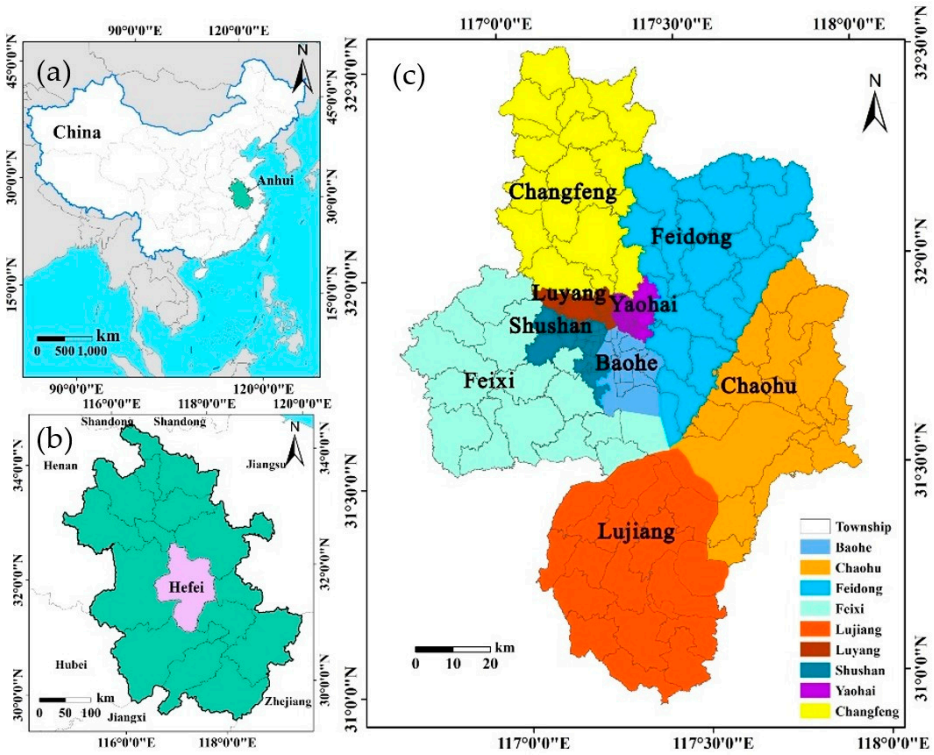

Figure 1. Location of study area; (**a**) Location of Anhui Province in China; (**b**) Location of the study area in Anhui Province; (**c**) The administrative division map of study area, and the minimum level are townships.

The administrative units of the study area are divided into eight county-scale units (covering all counties and districts) and 141 township scale units. Moreover, the population of Hefei increased rapidly from 4.382 million in 2000 to 8.087 million in 2018 (Bulletin of the Hefei sample survey on population changes, http://tjj.hefei.gov.cn/tjyw/tjgb/11546561.html, accessed on 19 January 2021). Owing to the rapid population change in Hefei and the difficulty in obtaining township population data for the intervals of census years, the

stepwise regression model for population estimation was applied to provide a reference for mapping the population density distribution.

3. Data and Methods

3.1. Data Collection

The main data used in this study include the satellite data, POI data, NTL data, township vector boundary data in Hefei, and census data, for 2018 (Table 1). Landsat 8 OLS multi-temporal images were selected to extract IS data on April 10, 2018 because of the sparse vegetation in spring but lush plants in summer and a large area of bare soil in winter affecting the experiment. The cloudiness of this image was 0.03, which had negligible impacts on data processing. Radiometric calibration, atmospheric correction, and geo-referencing were conducted from the remote sensing image downloaded for free from the United States Geological Survey (USGS, http://earthexplorer.usgs.gov/, accessed on 15 April 2020). Subsequently, the study area was masked from the image using administrative data from the Hefei administrative boundary.

Table 1. Data sets used in this study.

Data Sources	Description
Landsat imagery	Path 121 and row 38 on 10 April 2018, cloudiness of 0.03
POI data	123,348 points related to population from Baidu Map of Hefei in 2018
NTL data	Night light data from Luojia-1 satellite of Hefei in 2018
Population data at township scale	141 townships of Hefei based on the census in 2018
Administrative data	Boundary vector at township scale

Census data were obtained from the China County Statistical Yearbook published in 2019, and the Hefei Municipal Bureau Statistics. The 141 townships were used for modelling to explore the population density estimation model at a small scale. Moreover, the POI data, including 123,348 POIs in 2018, were crawled from the Baidu Map Services (http://map.baidu.com, accessed on 3 January 2021), which is the most widely used and largest web map service provider in China [7].

The NTL data of Luojia-1 were downloaded from the High-Resolution Earth Observation System of Hubei Data and Applications Network (http://www.hbeos.org.cn/, accessed on 27 January 2021). Finally, all spatial data were georeferenced in the same projection owing to the different source data. The flowchart of modelling the population density estimation is shown in Figure 2.

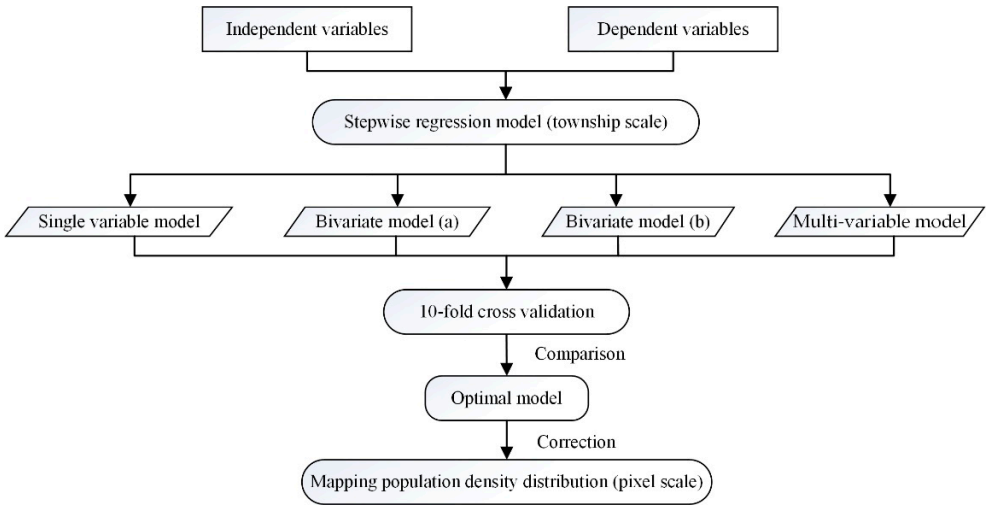

Figure 2. The flowchart of this study.

3.2. Mapping is Distribution and Validating the Results

LSMA can effectively discriminate different objects, such as soil, water, vegetation, and ISs. Therefore, LSMA was utilised in the mapping process to obtain the proportion of an IS within a pixel.

After pre-processing, the image was separated from other land covers using the modified normalised difference water index (MNDWI) [42], aiming to remove the non-IS fractions first. Subsequently, the minimum noise fraction (MNF) was adopted to reduce the redundancy within the image and improve the quality of the endmember selection [29]. Generally, three or four endmembers are selected [43]. As a consequence of confusing pixels in ISs, we selected four endmembers through repeated trials, including high- and low-albedo objects, vegetation, and bare soil, based on the vegetation-impervious surface-soil (V-I-S) model [44]. Finally, the fully constrained LSMA method was used to develop the fractional IS map, during which the high-and low-albedo objects were added together [45]. Thus, a map of the IS distribution in 2018 was obtained.

When measuring the V-I-S model fitness, the average root mean square (RMS) of overall bands was used to evaluate the accuracy of the LSMA (Equation (1)) [44]:

$$RMS = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\varepsilon(\lambda_i)^2} \quad (1)$$

where RMS is the root mean square, m is the number of pixels in this image and $\varepsilon(\lambda_i)$ is the residual error of the pixels after unmixing.

To further assess the unmixing accuracy, detailed ground reference data (the actual IS data) are required for comparison with the experimental data [46]. In this study, a total of 120 validated samples were randomly selected from satellite images in Hefei, each containing 3 × 3 pixels (90 × 90 m). Subsequently, all the validation samples were overlapped on the high-resolution image in 2018 from Google Earth Pro, which has been proven suitable for obtaining ground reference data [46]. Finally, the linear fit between the true and experimental data was evaluated using the coefficient of determination (R^2) and the root mean square error (RMSE). As the study is intended to estimate the population density using the IS data as a variable, the proportion of an IS in a pixel was calculated, and the mean data were used as a proxy for each township.

3.3. Modelling Population Density Estimation Using Stepwise Linear Regression

3.3.1. Data Preparation for Modelling

At present, the commonly used scales are 250, 500, and 1000 m in the study of population spatialisation. The selection of scale is mainly based on the area of the target and model stability. The 250 m scale and below is suitable for research in villages and communities, 500 m for that in counties and cities, and 1000 m for that in large-scale regional studies involving cities and provinces. According to existing research, the number of samples exceeding 50,000 affects the stability of the regression model and the accuracy of the model [13,40]. Therefore, the 500m scale was selected for modelling.

In this study, each 500 m grid was defined as a pixel. The pixel was built using ArcGIS 10.3 software. Using the 'Create Fishnet' function, the grid files of 500 m pixels were generated within the Hefei administrative boundary. The following experiments were conducted using ArcGIS 10.3 software.

The IS data were resampled to a resolution of 500 m. We obtained population-related POI data from Baidu Map in 2018 using Python to crawl the eight categories, including restaurants, shopping, hospitals and clinic facilities, education facilities, entertainment and retail, public service facilities, companies, and residential areas [7]. To avoid multi-collinearity caused by multiple variables, the POI data of eight categories were merged into an image.

The commonly used point-based methods include the analysis based on Euclidean distance and point density. The Euclidean distance considers that the plane space is

homogeneous, ignoring the relationship of facilities and service functions between cities, which is suitable for studying scattered points. Point density analysis is based on the aggregation of point distribution, and these points often interact with each other in space. We tried to aggregate the POIs into the grids to obtain the POI density. However, there was no distribution of POIs in the grids of sparse population areas, leading to the null data in samples.

The kernel density analysis is a point density method based on the first law of geography [7] and is suitable for the estimation of continuous geographic phenomena. POIs are distributed in the range of human activities and interact with urban facilities and transportation, and are consistent with aggregation distribution and continuous geographic phenomena.

Kernel density analysis was then conducted to analyse the point density spatially and discern the hotspots [47–49]. The crucial step was selecting the appropriate bandwidths [48]. We tested bandwidths varying from 1000 to 6000 m at an interval of 100 m, during which the correlation between the POI density generated by different bandwidths and the actual population density was analysed. Finally, it was found that when the bandwidth was 3500 m, the correlation between POI data and population density was the strongest. Therefore, the bandwidth of 3500 m performed well in identifying the hotspots, and we obtained the raster map of POI data at a 500 m pixel scale. The NTL image was clipped and pre-processed after downloading and subsequently resampled to pixels.

The vector file of the township boundary was superimposed with the raster image to obtain the township information on each pixel. To achieve consistency in the dimensions of different variables, the maximum and minimum standardisation methods were adopted to standardise the values of the three independent variables from 0 to 1 [21]. Logarithmic transformation was used to recalculate the population density data to eliminate the negative effects of the excessive population density [7]. Considering each township as a sample, the average value of each variable in each township was obtained as the independent variable value of a sample.

3.3.2. Model Concept and Validation

ISs are known to be closely related to population data [4]. However, the result of using IS data as the sole independent variable for population density estimations always leads to a coarse prediction, as a single variable cannot satisfactorily describe a populated zone [22,23]. Therefore, POI and NTL data were then added to the model. Assuming that the IS data has no significant multicollinearity with the other two data sources, the population density estimation model can be constructed (Equation (2)) [45]:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \quad (2)$$

where Y is the dependent variable (population density), X_i is the independent variable (IS, POI, and NTL data), β_0 is the intercept, β_i is the regression coefficient of the ith independent variable, ε is the error of the models, and n represents the number of independent variables.

To explore the relationship between IS data, NTL data, and POI data, Pearson correlation and partial correlation analyses were conducted. The former aims to measure the intensity of the monotonic relationship between variables [50], and the latter can analyse the correlation between two variables by keeping the other variables constant and eliminating their effects [51]. The correlation analysis provided the basis for modelling. Furthermore, to quantitatively characterise the multicollinearity, it was checked according to a rule of thumb stating that a variance inflation factor (VIF) value above 10 rules out the variable because of the high multicollinearity [52].

The models were constructed by randomly selecting samples from the townships. Hence, a stratified sampling procedure was conducted. To better assess the accuracy of the model, the 10-fold cross validation was applied to evaluate the performance of all models [7]. Approximately 90% of the townships (divided into nine groups) were randomly

selected from all 141 samples, and the rest (one group) were used to repeat 5 trials for the 10-fold cross validation. The census data of nine groups were used for training samples and one group for accessing the accuracy of the results. As the validated model can be applied for the population estimation, it is vital to test the model fit. In this study, the model fit between the predicted and true population density at the township scale was examined using the residual-based adjusted R^2, relative mean square error (RMSE), and the mean average error (MAE) [52].

As the independent variable data used in the modelling at the township scale were the average independent variable values of all pixels in the township, the optimal model can directly be applied to the pixel [4]. The IS data of each pixel were extracted by LSMA and then resampled, NTL data were obtained by directly processing the Luojia-1 NTL data image, and the POI data were obtained by calculating the kernel density. The population density data and independent variables of each township were obtained by connecting the township name with the 'Identity' function of ArcGIS 10.3 software.

The estimation models are based on stepwise linear regression. Despite its limitations, stepwise regression is still very popular in recent studies [53–55], because it can effectively identify the best variables in many related or unrelated variables to build a good prediction model [56,57]. Finally, the IS, NTL, and POI data were merged into the pixels using their IDs for predicting the population density in a pixel with the coefficients of the optimal model. To validate the population density estimation model at the 500 m pixel scale, three administrative units were selected as proxies for high-, medium-, and low-density areas, respectively.

4. Results

4.1. Analysis of is Distribution and Assessment

As the main independent variable of the model, the accuracy of the ISs directly affects the initial regression result. After calculating the MNDWI index, the water was masked. Four endmember types were distinguished based on the terrain properties, and finally, the high- and low-albedo endmembers were superimposed to obtain the distribution map of the IS proportion (Figure 3). The central part of Figure 3 shows the four main districts of Hefei, wherein the deeper colour indicates a higher IS value. Almost every township has a deep-coloured area, representing a population gathering and large IS distribution. The RMS was tested at 0.094 within a reasonable scope [58].

The distribution of the validation points is shown in Figure 4. After calculating the proportion of each endmember type in the external square of each point, the proportion of the IS in the square was compared with the IS abundance value extracted by LSMA. Subsequently, the experimental and true data were compared by calculating the linear fitting between them. The linear regression was up to the standard, with $R^2 = 0.788$ and the RMSE = 0.129 (Figure 5). Notably, a strong correlation was observed between experimental and true data.

4.2. Stepwise Regression Models for Population Density Estimation

4.2.1. Correlation Analysis between Variables

Pearson correlation analysis was conducted between independent variables, and the corresponding coefficients were calculated (Table 2). The IS data were positively correlated with the NTL data and POI data, with correlation coefficients of 0.729 and 0.689, respectively, and with all being less than 0.9, which is the threshold for correlation analysis in population estimation [59]. The partial correlation coefficient of each variable (Table 3) is smaller than the correlation coefficient.

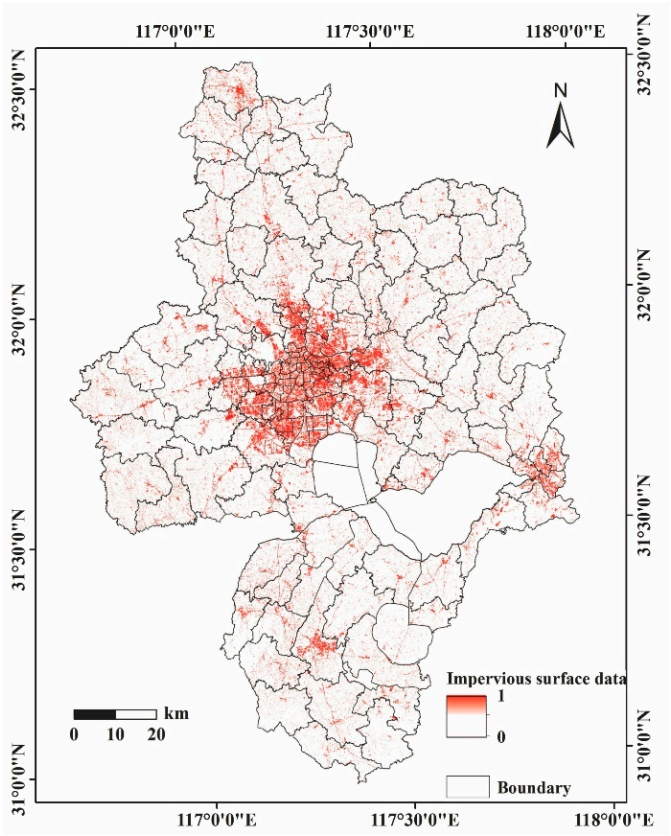

Figure 3. IS Distribution in Hefei. IS fraction values greater than 0.5 are calculated as impervious surfaces [27]. The deeper the colour, the greater is the value.

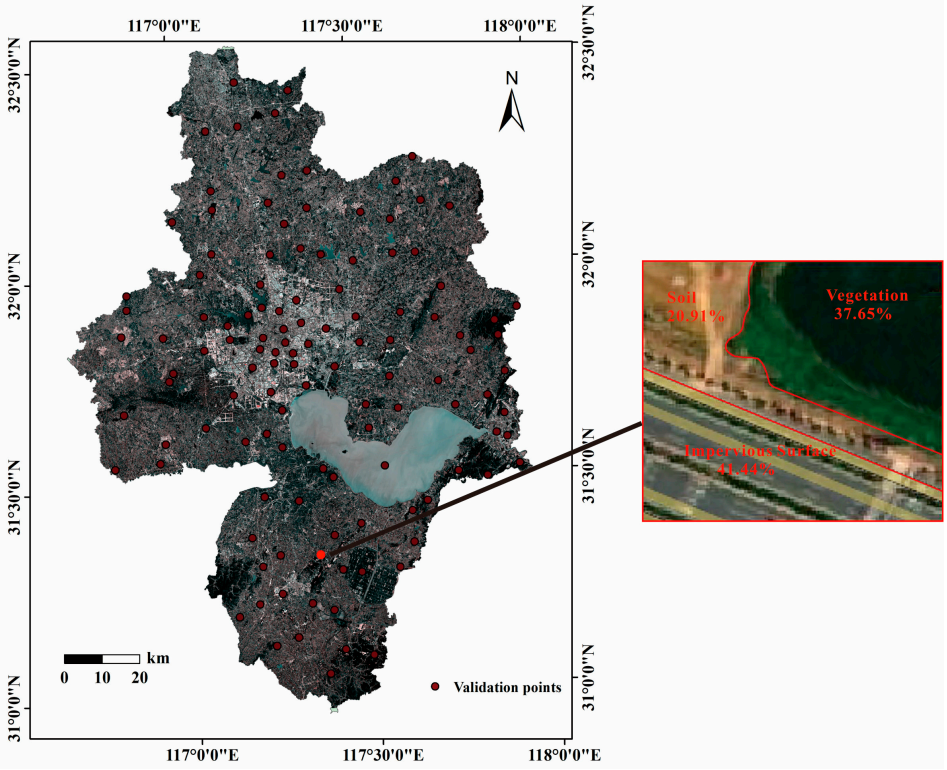

Figure 4. The proportion of each endmember in the image of validation points.

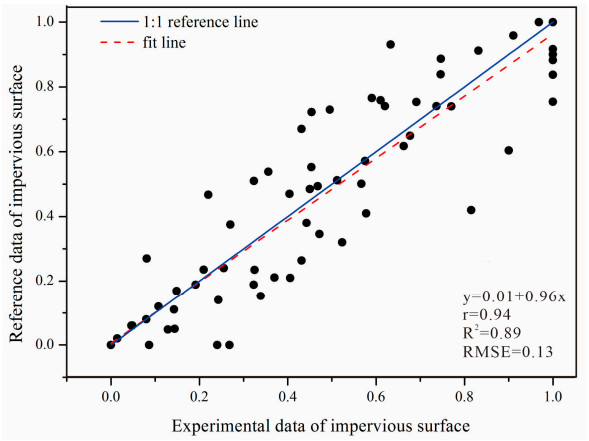

Figure 5. Result of the linear fitting of reference and experimental data of the IS.

Table 2. Pearson correlation coefficients between all variables in the township scale (***—at the 1% level, **—at the 5% level).

	IS Data	POI Data	NTL Data
IS data	1	0.729 ***	0.689 ***
POI data		1	0.673 **
NTL data			1

Table 3. Partial correlation coefficients between all variables in the township scale (***—at the 1% level).

	IS Data	POI Data	NTL Data
IS data		0.495 ***	0.391 ***
POI data	0.495 ***		0.345 ***
NTL data	0.391 ***	0.345 ***	

To further explore the correlation between the three independent variables, the variance inflation factor (VIF) values were calculated when the population density was the dependent variable [57] (Equation (3)):

$$VIF = \frac{1}{1 - R_i^2} \quad (3)$$

where R_i is the correlation coefficient of the ith independent variables.

The VIF values were all below 3 (Table 4). Based on the above analysis, the correlation between independent variables was weak and the multicollinearity was low, which provided a basis for constructing the multi-variable regression model [57]. Note that the bivariate model (a) included IS data and NTL data, and the bivariate model (b) included IS data and POI data.

Table 4. The coefficients (on a log scale) of different types of models of variables and VIF (***—at the 1% level).

	Single Variable Model	Bivariate Model (a)	Bivariate Model (b)	Multi-Variable Model
IS data	7.922 ***	6.700 ***	5.582 ***	4.567 ***
POI data			2.515 ***	1.497 ***
NTL data		2.317 ***		2.932 ***
Constant	4.983 ***	5.125 ***	5.325 ***	5.402 ***
Max VIF	1.000	2.438	2.916	2.989

4.2.2. Comparison and Validation of Models

In this study, the error between the predicted and actual population densities was calculated by mean absolute error (MAE) (Equation (4)):

$$MAE = \frac{\sum_{a=1,b=1}^{n} |Pop_a - Pop_b|}{n} \quad (4)$$

where MAE is the relative error of the population density estimation, Pop_a is the actual data of population density on a log scale, Pop_b is the prediction data of the population density on a log scale, and n is the number of townships.

The 10-fold cross validation results show that the single variable regression model for population density estimation was built with IS data as a single independent variable to test its prominence, which performs well, with R^2 = 0.689 and the RMSE = 0.922 (Table 5). Alternatively, the addition of POI data changed the model fit from 0.689 to 0.834 and the RMSE from 0.922 to 0.686, indicating the influence of POI data on the population density. The model fit performed best after adding all three independent variables in the regression, and the RMSE continued to decrease slightly.

Table 5. The validation result (on a log scale) for models.

	Training Group		Validation Group		
Types of Models	**Adj.R^2**	**RMSE**	**Adj.R^2**	**RMSE**	**MAE**
Single variable model	0.687	0.940	0.689	0.922	0.687
Bivariate model (a)	0.711	0.847	0.715	0.910	0.661
Bivariate model (b)	0.834	0.685	0.834	0.686	0.514
Multi-variable model	0.856	0.633	0.852	0.632	0.460

Table 5 shows the results of the mean value of the adjusted R^2, RMSE, and MAE. It indicates that the value of MAE decreased with the introduction of NTL data and POI data, and decreased most obviously after the introduction of POI data. In the bivariate models, model (b) with IS data and POI data as independent variables performed better. The MAE of the multi-variable model after the introduction of NTL data and POI data was the lowest, and the model fit was the best, which is regarded as the optimal model.

In the optimal model, the group of minimum MAE was selected to explore the relationship between the predicted population density and the actual population density (Figure 6). The model achieved a cross-validated MAE of 0.420 and an adjusted R^2 of 0.832.

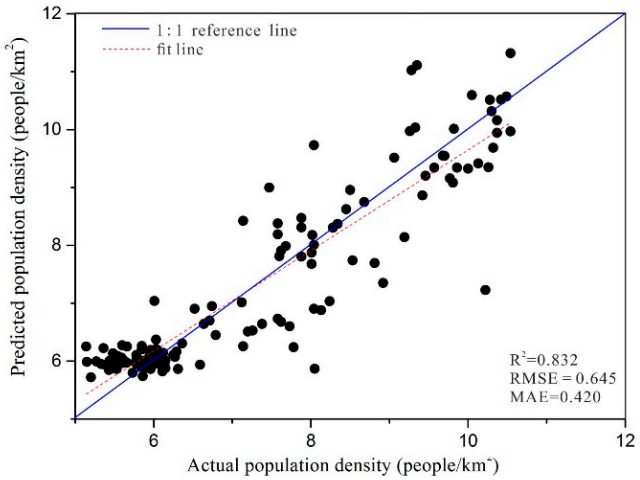

Figure 6. Relationship between true population density and predicted population density of optimal model (on a log scale). The data are from a group of multi-variable models with minimum MAE.

Before population spatialization, it is necessary to modify the population density between the township scale and the pixel scale to reduce the error caused by downscaling [60] (Equation (5)):

$$Pop'_{i,j} = Pop_{b\_i,j} \cdot \frac{Pop_{a\_i}}{\sum_{j=1}^{N} Pop_{b\_i,j} \cdot Area_j} \quad (5)$$

where $Pop'_{i,j}$ is the modified population density of jth pixel of the ith township after exponential transformation, $Pop_{b\_i,j}$, is the prediction data of population density of the jth pixel of the ith township after exponential transformation, Pop_{a_i} is the actual data of population density of ith township, N is the number of townships, $Area_j$ is the area of the jth pixel.

Figure 7a shows the population distribution maps based on census data, and Figure 8 shows those based on predicted data. It can be seen from Figure 7a that the population density within the administrative unit is homogeneous, and distinguishing between densely populated and sparsely populated areas is not possible; however, the township scale is the smallest in the census. Figure 7b shows the population distribution using a multi-variable regression model at a 500 m pixel scale. In the main urban areas with large population density gaps, the population distribution is no longer divided by administrative boundaries, and the value of each pixel represents the population density in the area. Therefore, this map more precisely reflects the population distribution within the administrative unit.

The current study areas for population estimation are concentrated in countries at the province or county scale [23,25], always leading to a coarse population distribution. Our study explored the population density estimation model in Hefei (a developing city), emphasising the importance of population data for some policies, such as the introduction of talent and household registration for migrant workers, which is closely concerned with the population. The model for population estimation can predict the population in the interval years of census data and then be applied to the pixel scale to redistribute the population. Owing to the lack of census data at the pixel scale, it was difficult to assess the accuracy of the population spatialisation quantitatively. Therefore, we compared the predicted data with high-resolution images and maps of administrative units at the township scale with three representative areas: high-, medium-, and low-density areas (Figures 8–10) [60].

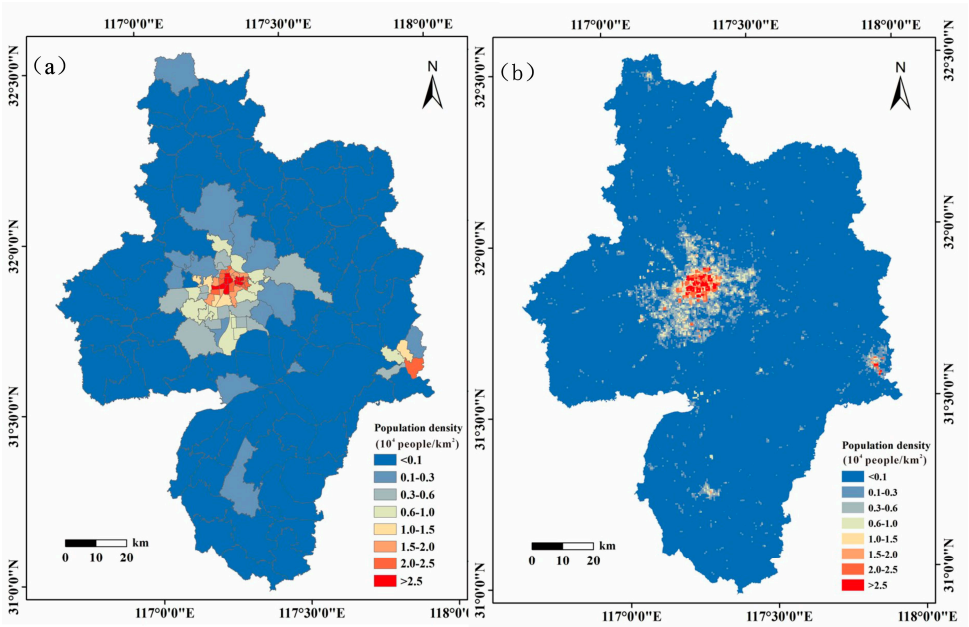

Figure 7. (**a**) Population density (10^4 people/km^2) distribution of Hefei at the township scale; (**b**) Spatial distribution of the population density (10^4 people/km^2) at the 500 m pixel scale.

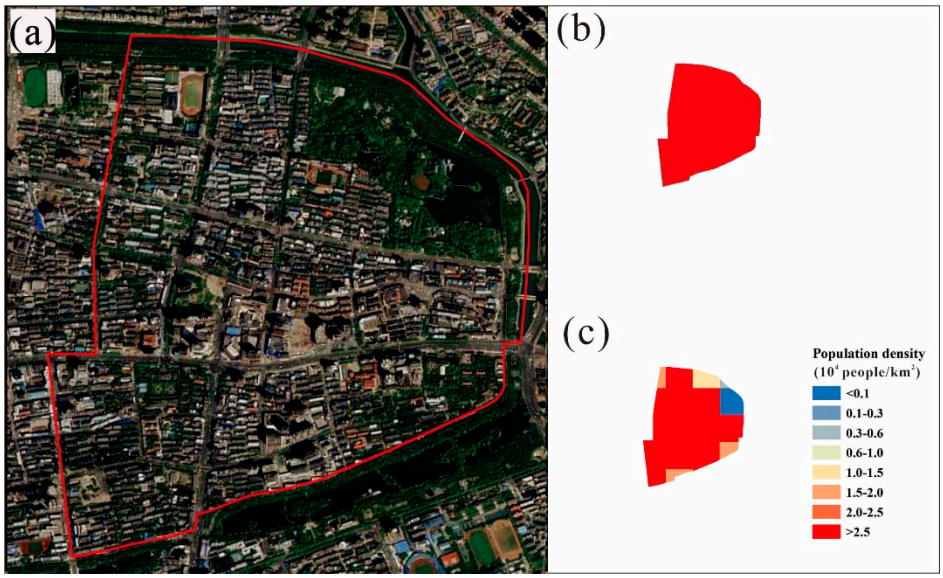

Figure 8. Comparison of (**a**) the high-resolution image, (**b**) the population density at the township scale, and (**c**) that at the 500 m scale in Xiaoyaojin Township as the high-density areas.

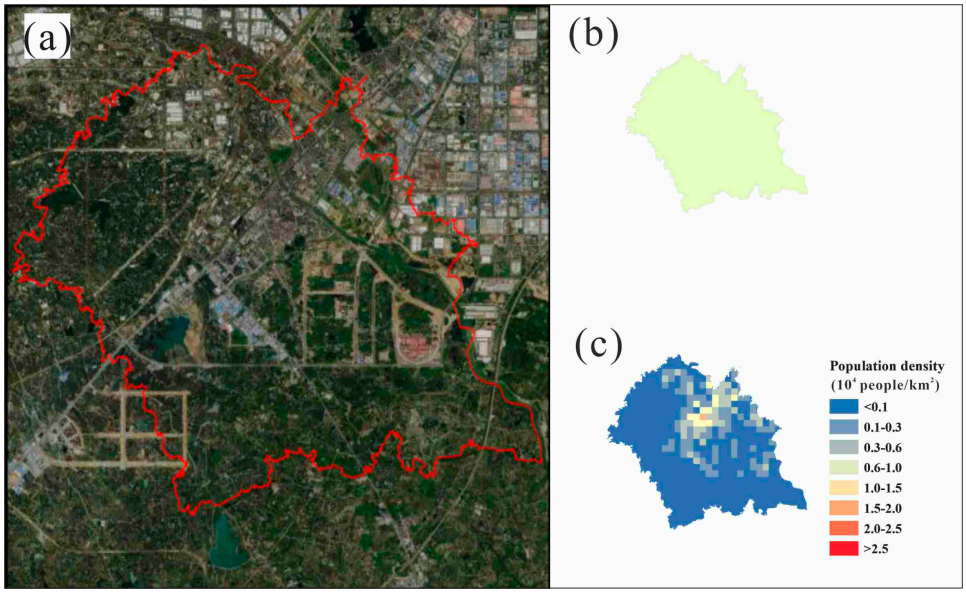

Figure 9. Comparison of (**a**) the high-resolution image, (**b**) the population density at the township scale, and (**c**) that at the 500 m scale in Shangpai Township as the medium-density areas.

Figures 8–10 depict the comparison between Google Earth high-resolution images, census data, and population density at a 500 m pixel scale. Notably, all population distributions at the township scale were homogenous. From the finer-scale maps of the three figures, the pixels at the administrative boundary of the townships are segmented. The value of independent variables in the segmented pixel is determined by the township with the pixels' largest area. However, the two segmented parts are involved in the calculation, resulting in the repeated calculation of the boundary part. This is one error source in the pixel-scale population distribution map.

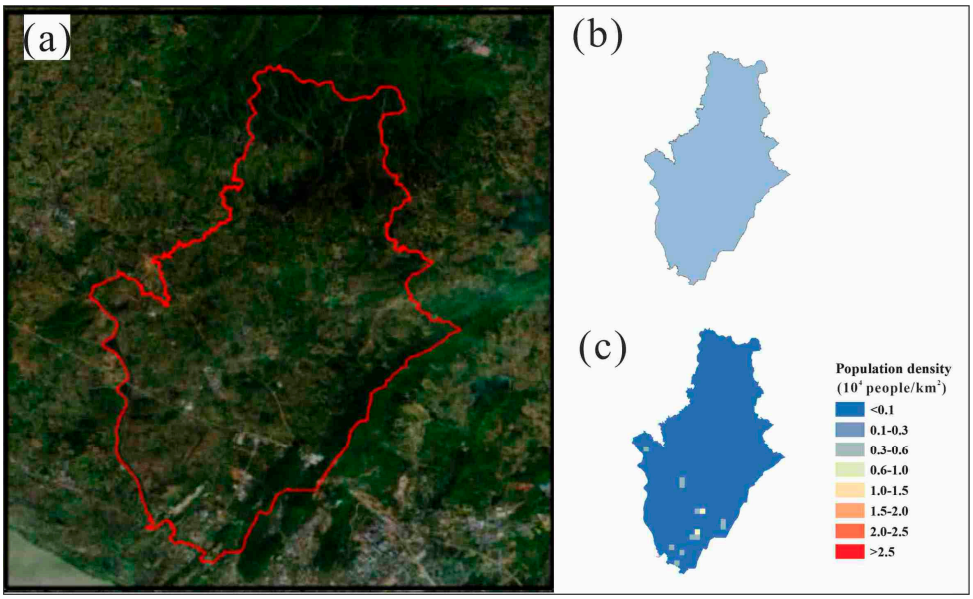

Figure 10. Comparison of (**a**) the high-resolution image, (**b**) the population density at the township scale, and (**c**) that at the 500 m scale in Xiage Township as the low-density areas.

Xiaoyaojin Township is located in the southeast of Luyang District, Hefei, covering an area of 2.92 km^2. The population density here is 23.3 $\times$ 10^4 people/km^2. The north-eastern part of Xiaoyaojin Township is dispersed with vegetation, and other parts are full of buildings, as shown in Figure 8a. The north-eastern part is sparsely populated. The building area has high IS and NTL data values, along with a high density of POI data with a concentrated population distribution. Figure 9b cannot distinguish between the north-eastern part and other parts of the country's population distribution differences. The pixels from the northeast of Figure 8c are dark in colour, indicating a sparse population. The others are yellow and orange, indicating a high population density. Therefore, the pixel map (Figure 8c) can effectively reflect the population distribution in a high-density area.

Located at the junction of the main city of Hefei and Feixi County, Shangpai covers an area of 121 km^2, with a population density of 2228.29 people/km^2. From the high-resolution image (Figure 9a), the central area of Shangpai Township is located in the northern part, and several dwelling units are sprinkled in the south-eastern area. Other areas are mainly forests or croplands. The population density of the Shangpai town, displayed in Figure 9b, does not consider the internal population distribution differences. The yellow pixels shown in Figure 9c correspond to the building area in Figure 9a, which represents the area where the population is concentrated.

Xiage Township belongs to Chaohu City (a county-level city in Hefei), covering an area of 184 km^2, with a population density of 306.27 people/km^2. It is located on the northern bank of Chaohu Lake, with beautiful scenery. The township (Figure 10a) mainly consists of mountainous terrain with vast tracts of forests, and the residential areas are distributed in the southern part, as shown in Figure 10c.

5. Discussion

Testing on extracting the proportion of ISs by LSMA obtained a satisfying result, corresponding with the previous findings [29,61]. The map of the ISs showed that their distribution was related to the population (Figures 3 and 7). However, estimating the population with a sole variable may led to a coarse result. Thus, POI and NTL data were considered as added independent variables. Note that the number of independent variables in the multi-variable model should be selected with caution to avoid being affected by overfitting and model complexity. From the VIF value in Table 4, with the increase of the number of variables, the collinearity between variables also gradually

increased, which indicates that three is a stable number of variables. Low VIF values and partial correlation coefficients implied low multicollinearity among the independent variables for the population density estimation. In the stepwise process, the higher adjusted R^2, lower RMSE, and lower MAE (Table 5) were attributed to the addition of POI and NTL data, corresponding to our assumption. Compared with the population density estimation model based on IS data, the method obtains a smaller error [3].

The population density estimation model has a small number of samples: only 141 township samples were used for modelling. Therefore, the model applies to similar scales, such as municipal study areas or small regions. The accuracy of the model was assessed by a 10-fold cross-validation method, indicating that the model was effective in Hefei in 2018. The verification of the model can be further deepened, such as trying to verify it in other urban areas, or applying the model in historical years, and then using census data to verify it. This will be our future research direction.

Figure 11 depicts the residuals of the optimal model. We have tested the residuals of these predictions of the multi-variable model. The Moran' I was −0.016, and the *p*-value was 0.466. Therefore, the spatial distribution of residuals does not conform to spatial autocorrelation. The township with the largest residual is Yafu township, with a residual error of 2602 people/km^2. Yafu township is located in the southeast of Chaohu (Figure 1). As can be seen in Figure 3, the ISs are distributed here. Therefore, in future research, we will try to establish population estimation models by calculating the height of urban buildings, which might be very useful to improve the estimation accuracy of the model.

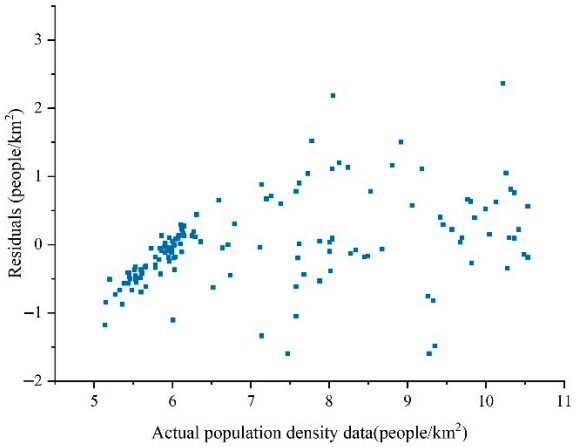

Figure 11. Residuals (on a log scale) between the estimated population density and the corresponding actual population density at the township scale.

In our future research, we will try to combine all independent variables to estimate the population. Additionally, data sources are also important. Geographical and economic data are popular in population estimation studies [62,63], but are difficult to combine with the population because of the spatial gap. Moreover, the lowest scale of economic data is the county (or district) scale, obstructing a deep study on the township scale or smaller. In contrast, the remote sensing data in our study can correspond with the township scale and have obvious advantages over these data types, which are not only easy to access but are updated periodically.

The optimal model developed in this study—incorporating IS data with POI and NTL data in modelling population density and producing a high-resolution map of the population distribution—can effectively and rapidly estimate populations, and has considerable potential in the era of big data. Our future research will try to estimate populations from a smaller scale, such as using sentinel satellites to extract impervious surfaces, in order to obtain a higher resolution of population spatialization results.

6. Conclusions

Based on IS data, this study gradually introduces NTL and POI data and establishes an optimization model for population density estimation. Exploring the correlation and multicollinearity between variables, the results meet the requirements of establishing a population density estimation model. The 10-fold cross validation of the four models shows that the multi-variable model achieves the best prediction effect. The parameters of the optimal model are applied to 500 m pixels, and the population density distribution map at the pixel scale is obtained after correction. Overall, the multi-variable (including IS, NTL, and POI data) model can effectively predict the population density in interval years and for areas lacking census data.

Author Contributions: Conceptualisation, J.Z., T.Z., L.L. and L.C.; methodology, J.Z., W.L. and R.L.; validation, J.Z. and L.Y.; software, J.Z., Z.W. and Z.Y.; formal analysis, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, W.L., L.L., L.C., Y.Z. and J.W.; supervision, project administration, and funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities under grant number 2018ZDPY07.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: The authors would like to thank the Wikipedia (https://en.m.wikipedia.org/wiki/Baidu_Maps, accessed on 26 July 2021), the United States Geological Survey (USGS, http://earthexplorer.usgs.gov/, accessed on 15 April 2020), Hefei Municipal Bureau Statistics, the Baidu Map Services (\protect\unhbox\voidb@x\hbox{http://}map.baidu.com, accessed on 3 January 2021), and the High-Resolution Earth Observation System of Hubei Data and Applications Network (http://www.hbeos.org.cn/, accessed on 27 January 2021), and to the anonymous associate editors and reviewers for their constructive comments and suggestions.

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

References

1. Ciommi, M.; Egidi, G.; Salvia, R.; Cividino, S.; Rontos, K.; Salvati, L. Population dynamics and agglomeration factors: A non-linear threshold estimation of density effects. *Sustainability* **2020**, *12*, 2257. [CrossRef]
2. Wu, S.S.; Qiu, X.; Wang, L. Population estimation methods in GIS and remote sensing: A review. *GIScience Remote Sens.* **2005**, *42*, 80–96. [CrossRef]
3. Zhu, H.; Li, Y.; Liu, Z.; Fu, B. Estimating the population distribution in a county area in China based on impervious surfaces. Photogramm. *Eng. Remote Sens.* **2015**, *81*, 155–163.
4. Li, L.; Lu, D. Mapping population density distribution at multiple scales in Zhejiang Province using Landsat Thematic Mapper and census data. *Int. J. Remote Sens.* **2016**, *37*, 4243–4260. [CrossRef]
5. Lu, D.; Weng, Q.; Li, G. Residential population estimation using a remote sensing derived impervious surface approach. *Int. J. Remote Sens.* **2006**, *27*, 3553–3570. [CrossRef]
6. Zhang, G.; Rui, X.; Poslad, S.; Song, X.; Fan, Y.; Wu, B. A method for the estimation of finely-grained temporal spatial human population density distributions based on cell phone call detail records. *Remote Sens.* **2020**, *12*, 1–28.
7. 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]
8. 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]
9. Yu, S.; Zhang, Z.; Liu, F. Monitoring population evolution in China using time-series DMSP/OLS nightlight imagery. *Remote Sens.* **2018**, *10*, 194. [CrossRef]
10. Jing, C.; Zhou, W.; Qian, Y.; Yan, J. Mapping the urban population in residential neighborhoods by integrating remote sensing and crowdsourcing data. *Remote Sens.* **2020**, *12*, 3235. [CrossRef]
11. Xu, M.; Cao, C.; Jia, P. Mapping Fine-Scale Urban Spatial Population Distribution Based on High-Resolution Stereo Pair. *Remote Sens.* **2020**, *12*, 608. [CrossRef]

12. Luo, P.; Zhang, X.; Cheng, J.; Sun, Q. Modeling population density using a new index derived from multi-sensor image data. *Remote Sens.* **2019**, *11*, 2620.
13. Li, G.; Weng, Q.; Li, G.; Weng, Q. Fine-scale population estimation: How Landsat ETM + imagery can improve population distribution mapping. *Can. J. Remote. Sens.* **2014**, *16*, 8992. [CrossRef]
14. Dong, P.; Ramesh, S.; Nepali, A. Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data. *Int. J. Remote. Sens.* **2010**, *31*, 1161. [CrossRef]
15. Karunarathne, A.; Lee, G. Estimating Hilly Areas Population Using a Dasymetric Mapping Approach: A Case of Sri Lanka ' s Highest Mountain Range. *Int. J. Geo-Inf.* **2019**, *8*, 166. [CrossRef]
16. Hegedus, E. Population Estimation from Landsat Imagery. *Remote Sens. Environ.* **1982**, *272*, 259–272.
17. Lo, C.P. Automated population and dwelling unit estimation from high-resolution satellite images: A GIS approach. *Int. J. Remote. Sens.* **2007**, *1161*, 16–34.
18. Arnold, C.L.; Gibbons, C.J. Impervious Surface Coverage: The Emergence of a Key Environmental Indicator. *J. Am. Plan. Assoc.* **1996**, *62*, 243–258. [CrossRef]
19. Zhou, Z.; Sha, J.; Ji, J. The Study of the Relationship between Urban Heat Island Effect and Impervious Surface and Spatio-temporal Change in Urban Areas of Fuzhou. *J. Fujian Norm. Univ.* **2019**, *35*, 24–32.
20. Wang, X.; Xue, W.; Zhao, J. Sepctral Mixture Analysis and Mapping of Impervious Surface in Central Uran of Xi'an. Shaanxi. *For. Sci. Technol.* **2019**, *47*, 32–38.
21. Cui, Q.; Pan, Y.; Yang, X. Beijing Plain Area of Remote Sensing Images Based on Landsat 8 Impervious Layer Coverage Estimates. *J. Cap. Norm. Univ.* **2015**, *36*, 89–92.
22. Wu, C.; Murray, A.T. Population Estimation Using Landsat Enhanced Thematic Mapper Imagery. *Geogr. Anal.* **2007**, *39*, 26–43. [CrossRef]
23. Joseph, M.; Wang, L.; Wang, F. Using Landsat Imagery and Census Data for Urban Population Density Modeling in Port-au-Prince, Haiti. *GIScience Remote Sens.* **2013**, *49*, 1603. [CrossRef]
24. Xu, H.; Shi, T.; Wang, M.; Fang, C.; Lin, Z. Predicting effect of forthcoming population growth–induced impervious surface increase on regional thermal environment: Xiong'an New Area, North China. *Build. Environ.* **2018**, *136*, 98–106. [CrossRef]
25. Azar, D.; Graesser, J.; Engstrom, R.; Comenetz, J.; Andrews, T. Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. *Remote Sens.* **2010**, *31*, 5635–5655. [CrossRef]
26. Sugg, Z.P.; Finke, T.; Goodrich, D.; Moran, M.S.; Yool, S.R. Mapping Impervious Surfaces Using Object-oriented Classification in a Semiarid Urban Region. *Photogramm Eng. Remote Sens.* **2014**, *80*, 343–352. [CrossRef]
27. Wu, C. Normalized spectral mixture analysis for monitoring urban composition using ETM + imagery. *Remote. Sens. Environ.* **2004**, *93*, 480–492. [CrossRef]
28. Duan, P.; Li, J.; Lu, X.; Feng, C. Estimation of Impervious Surface Distribution by Linear Spectral Mixture Analysis: A Case Study in Nantong. In Proceedings of the 2nd EAI International Conference on Robotic Sensor Networks, Kunming, China, 25–26 August 2018; pp. 41–51.
29. Li, L.; Lu, D.; Kuang, W. Examining urban impervious surface distribution and its dynamic change in Hangzhou metropolis. *Remote Sens.* **2016**, *8*, 265. [CrossRef]
30. Liu, Q.; Sutton, P.C.; Elvidge, C.D. Relationships between Night Imagery and Population Density for Hong Kong. Proc. *Asia-Pacific Adv. Netw.* **2011**, *31*, 79.
31. Zou, Y.; Yan, Q.; Huang, J.; Li, F. Modeling the Population Density of Su-Xi-Chang Region Based on Luojia-1A Night Light Image. *Resour. Environ. Yangtze Basin* **2020**, *29*, 1086–1094.
32. Huang, X.; Yang, J.; Li, J.; Wen, D. Urban functional zone mapping by integrating high spatial resolution night light and daytime multi-view imagery. *ISPRS J. Photogramm Remote Sens.* **2021**, *175*, 403–415. [CrossRef]
33. Zhao, M.; Zhou, Y.; Li, X.; Cheng, W.; Huang, K. Mapping urban dynamics (1992–2018) in Southeast Asia using consistent nighttime light data from DMSP and VIIRS. *Remote Sens. Environ.* **2020**, *248*, 111980. [CrossRef]
34. Levin, N.; Kyba, C.; Zhang, Q.; Miguel, A.; Elvidge, C.D. Remote sensing of night lights: A review and an outlook for the future. *Remote Sens. Environ.* **2020**, *237*, 111443. [CrossRef]
35. Gao, H.; Li, D.; Zhou, L.; Yu, T.; Huang, J. Population spatialization based on multiple night lighting data comparison. *Intelligent City.* **2020**, *6*, 26–27.
36. Zhong, L.; Liu, X. Application potential analysis of LJ1-01 new nighttime light data. *Bull. Surv. Mapp.* **2019**, *7*, 132–137.
37. Yin, J.; Fu, P.; Hamm, N.A.S.; Li, Z.; You, N.; He, Y.; Cheshmehzangi, A.; Dong, J. Decision-Level and Feature-Level Integration of Remote Sensing and Geospatial Big Data for Urban Land Use Mapping. *Remote Sens.* **2021**, *13*, 1579. [CrossRef]
38. Chun, J.; Zhang, X.; Huang, J.; Zhang, P. A Gridding Method of Redistributing Population Based on POIs. *Geogr. Geo-Inf. Sci.* **2018**, *34*, 89–95.
39. Chen, Y.; Ge, Y.; An, R.; Chen, Y. Super-Resolution Mapping of Impervious Surfaces from Remotely Sensed Imagery with Points-of-Interest. *Remote Sens.* **2018**, *10*, 242. [CrossRef]
40. Zhao, Y.; Li, Q.; Zhang, Y.; Du, X. Improving the accuracy of fine-grained population mapping using population-sensitive POIs. *Remote Sens.* **2019**, *11*, 2502. [CrossRef]
41. Feng, X.; Jin, Z. Belt and Road: An analysis based on Hefei. *J. Suihua Univ.* **2019**, *39*, 19–23.

42. Xu, H. A Study on Information Extraction of Water Body with the Modified Normalized Difference Water Index (MNDWI). *J. Remote. Sens.* **2005**, *9*, 79–85.
43. Small, C. The Landsat ETM+ spectral mixing space. *Remote Sens. Environ.* **2004**, *93*, 1–17. [CrossRef]
44. Ridd, M.K. Exploring a V-I-S (Vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: Comparative anatomy for citiest. *Int. J. Remote Sens.* **1995**, *16*, 2165–2185. [CrossRef]
45. Li, L.; Canters, F.; Solana, C.; Ma, W.; Chen, L.; Kervyn, M. Discriminating lava flows of different age within Nyamuragira's volcanic field using spectral mixture analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *40*, 1–10. [CrossRef]
46. Cui, Y.; Li, L.; Chen, L.; Zhang, Y.; Cheng, L.; Zhou, X.; Yang, X. Land-use carbon emissions estimation for the Yangtze River Delta Urban Agglomeration using 1994-2016 Landsat image data. *Remote Sens.* **2018**, *10*, 1334. [CrossRef]
47. Lin, Y.P.; Chu, H.J.; Wu, C.F.; Chang, T.K.; Chen, C.Y. Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques. *Int. J. Environ. Res. Public Health* **2011**, *8*, 75–88. [CrossRef] [PubMed]
48. Chainey, S. Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bull. Geogr. Soc. Liege* **2013**, *60*, 7–19.
49. Gatrell, A.C.; Bailey, T.C.; Diggle, P.J.; Rowlingson, B.S. Spatial point pattern analysis and its application in geographical epidemiology. *Trans. Inst. Br. Geogr.* **1996**, *21*, 256–274. [CrossRef]
50. Nahler; Gerhard Pearson correlation coefficient. *Springer Vienna* **2009**, *10*, 132.
51. Finn, J.D. *A General Model for Multivariate Analysis*; Holt Rinehart Winst: New York, NY, USA, 1974; pp. 173–174.
52. Li, L.; Zhou, X.; Chen, L.; Chen, L.; Zhang, Y.; Liu, Y. Estimating urban vegetation biomass from sentinel-2A image data. *Forests* **2020**, *11*, 125. [CrossRef]
53. Abdelhafidi, N.; Bachari, N.E.I.; Abdelhafidi, Z. Estimation of solar radiation using stepwise multiple linear regression with principal component analysis in Algeria. *Meteorol. Atmos. Phys.* **2020**, *133*, 1–12. [CrossRef]
54. Nouman, S. Multiple and stepwise regression of reproduction efficiency on linear type traits in Sahiwal cows. *Int. J. Livest. Prod.* **2013**, *4*, 14–17. [CrossRef]
55. Zhou, Y.; Qureshi, R.; Sacan, A. Analysis of paired miRNA-mRNA microarray expression data using a stepwise multiple linear regression model. *Lect. Notes Comput. Sci.* **2017**, *10330*, 59–70.
56. Li, L.; Bakelants, L.; Solana, C.; Canters, F.; Kervyn, M. Dating lava flows of tropical volcanoes by means of spatial modeling of vegetation recovery. *Earth Surf. Process. Landforms.* **2018**, *43*, 840–856. [CrossRef]
57. Yuan, L.; Li, L.; Zhang, T.; Chen, L.; Liu, W.; Hu, S.; Yang, L. Modeling Soil Moisture from Multisource Data by Stepwise Multilinear Regression: An Application to the Chinese Loess Plateau. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 233. [CrossRef]
58. Wu, C.; Murray, A.T. Estimating impervious surface distribution by spectral mixture analysis. *Remote Sens. Environ.* **2003**, *84*, 493–505. [CrossRef]
59. Zou, Y. Research on Population Spatialization Based on Multi-Source Data. Master's Thesis, China University of Mining and Technology (Jiangsu), Xuzhou, China, 2020.
60. He, M.; Xu, Y.; Li, N. Population Spatialization in Beijing City Based on Machine Learning and Multisource Remote Sensing Data. *Remote Sens.* **2020**, *12*, 1910. [CrossRef]
61. Li, H.; Li, L.; Chen, L.; Zhou, X.; Cui, Y.; Liu, Y.; Liu, W. Mapping and characterizing spatiotemporal dynamics of impervious surfaces using landsat images: A case study of Xuzhou, East China from 1995 to 2018. *Sustainability* **2019**, *11*, 1224. [CrossRef]
62. Xu, Z.; Ouyang, A. The Factors Influencing China's Population Distribution and Spatial Heterogeneity: A Prefectural-Level Analysis using Geographically Weighted Regression. *Appl. Spat. Anal. Policy.* **2018**, *11*, 465–480. [CrossRef]
63. Mi, R.; Gao, X. Factors influencing population distribution in Shaanxi Province using spatial econometric analysis. *Arid Land Geogr.* **2020**, *43*, 491–498.