International Journal of
Geo-Information

MDPI

Article

Improved Population Mapping for China Using the 3D Building, Nighttime Light, Points-of-Interest, and Land Use/Cover Data within a Multiscale Geographically Weighted Regression Model

Zhen Lei [1,2], Shulei Zhou [1,2], Penggen Cheng [1,2,*] and Yijie Xie [1,2]

[1] School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China; 2020120292@ecut.edu.cn (Z.L.); 2021110337@ecut.edu.cn (S.Z.); 2021110155@ecut.edu.cn (Y.X.)

[2] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China

* Correspondence: pgcheng@ecut.edu.cn

Abstract: Large-scale gridded population product datasets have become crucial sources of information for sustainable development initiatives. However, mainstream modeling approaches (e.g., dasymetric mapping based on Multiple Linear Regression or Random Forest Regression) do not consider the heterogeneity and multiscale characteristics of the spatial relationships between influencing factors and populations, which may seriously degrade the accuracy of the prediction results in some areas. This issue may be even more severe in large-scale gridded population products. Furthermore, the lack of detailed 3D human settlement data likewise poses a significant challenge to the accuracy of these data products. The emergence of the unprecedented Global Human Settlement Layer (GHSL) data package offers a possible solution to this long-standing challenge. Therefore, this study proposes a new Gridded Population Mapping (GPM) method that utilizes the Multiscale Geographically Weighted Regression (MGWR) model in conjunction with GHSL-3D Building, POI, nighttime light, and land use/cover datasets to disaggregate population data for third-level administrative units (districts and counties) in mainland China into 100 m grid cells. Compared to the WorldPop product, the new population map reduces the mean absolute error at the fourth-level administrative units (townships and streets) by 35%, 51%, and 13% in three test regions. The proposed mapping approach is poised to become a crucial reference for generating next-generation global demographic maps.

Keywords: population mapping; China; building; MGWR

check for updates

Citation: Lei, Z.; Zhou, S.; Cheng, P.; Xie, Y. Improved Population Mapping for China Using the 3D Building, Nighttime Light, Points-of-Interest, and Land Use/Cover Data within a Multiscale Geographically Weighted Regression Model. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 335. https://doi.org/10.3390/ijgi13090335

Academic Editors: Wolfgang Kainz and Godwin Yeboah

Received: 20 July 2024
Revised: 9 September 2024
Accepted: 13 September 2024
Published: 19 September 2024

1. Introduction

Population data are indispensable for various sustainable development applications, including disaster assessment, urban planning, and public health management [1–6]. While census data serve as the primary source of population data, their coarse resolution limits the revelation of spatial heterogeneity within census units, hindering their application in research related to global social and environmental issues [7]. To address this limitation, several large-scale gridded population datasets have been produced, such as GPWv4, HRSL, LandScan, and WorldPop [8].

The datasets above are generated through the top-down population estimation methods, where census data are disaggregated into unified grid cells based on population distribution weighting layers [9]. Over the past three decades, various modeling approaches have been developed to calculate the weighting layers, including areal weighting [10,11], negative exponential [12,13], kernel density [14–16], and dasymetric mapping models [17]. With the rapid development of AI technology, intelligent dasymetric mapping has gradually become the dominant approach in Gridded Population Mapping (GPM) studies. This approach leverages algorithms to model the unknown prior relationships between auxiliary variables and the population to obtain the weighting layer [18–21]. A notable example is the Random Forest (RF) model used to generate the WorldPop product [17,22,23]. Additionally, Multiple Linear Regression (MLR) is commonly used in some GPM studies [5,19]. While both MLR and RF have shown relatively good performance in urban-scale GPM studies, they may not produce accurate results in large-scale study areas (e.g., China). This is primarily due to the significant regional variations in population distribution patterns across such expansive areas. Specifically, the relationship between population and auxiliary (explanatory) variables is spatially heterogeneous (non-stationary) and multi-scale. Using a single global model (e.g., MLR or RF) can lead to interregional heterogeneity being masked by 'average' estimates for the study area as a whole, potentially leading to inaccurate predictions of population distribution in localized regions.

New parameter estimation methods offer promising solutions. For instance, MLR assumes the relationship between the dependent and explanatory variables is spatially stationary [24]. To address this limitation, Geographically Weighted Regression (GWR) was developed by Fotheringham in 1996. GWR improves MLR by employing non-parametric local weighted regression for curve fitting and smoothing applications [25]. Unlike MLR, GWR considers the non-stationarity of the spatial relationship between the dependent variable and explanatory variables, making it more effective for analyzing factors related to spatial locations [4–6,19]. However, both MLR and GWR are limited in revealing spatial scale differences in the relationships between explanatory variables and the dependent variable. Specifically, the influence of different explanatory variables may be similar within a specific range but differ significantly beyond that range. To address this issue, Fotheringham proposed Multiscale Geographically Weighted Regression (MGWR) in 2017 [26]. Yu et al. [27] further supplemented and improved the statistical inference of MGWR, making this method more widely applicable to research. Compared to GWR, MGWR assigns specific bandwidths to each explanatory variable, allowing for the establishment of spatial relationship models closer to reality.

Auxiliary datasets, such as those on land use/land cover, topography, roads, and rivers, are often used in large-scale GPM studies. However, these datasets primarily reflect the potential of human settlements rather than directly indicating whether a specific location is inhabited [8]. Although mobile phone location data can provide real-time insights into population distribution [28,29], its limited accessibility poses challenges for large-scale GPM applications. Compared to data such as land use/cover and topography, human settlements directly indicate the site is inhabited, enabling a more accurate and detailed depiction of the population distribution range. In addition, several openly available global or near-global human settlement datasets have been developed, including Microsoft and Google building footprints, HRSL, the World Settlement Footprint (WSF), and the Global Human Settlement Layer (GHSL). The accuracy of gridded population datasets can be improved by using relatively complete human settlement data as ancillary data. Studies have demonstrated that using these datasets can enhance the internal quantitative and qualitative accuracy of population distribution models by 10% to 15% (depending on different indicators) [21,30–32].

Although human settlement data have been applied in GPM, the datasets used in existing large-scale studies usually lack vertical (height or number of floors) and type (residential/non-residential) information [21,33,34]. Currently, mainstream large-scale gridded population products, including HRSL, LandScan, and WorldPop, rely on 2D and non-functional human settlement auxiliary layers [8]. Thomson et al. reported severe underestimation (averaging over 80%) in slum areas of Kenya and Nigeria due to the absence of detailed information about human settlements, such as usage and height, in the products above (and others) [35]. Multiple studies have shown that using building data with vertical information and categorization can significantly improve the accuracy of gridded population outputs [36–40]. This improvement is mainly attributed to considering the vertical distribution across building floors and the exclusion of non-residential buildings. In the past, the lack of openly available large-scale 3D residential/non-residential building datasets has strongly limited their application in continental or global-scale GPM. The emergence of the new GHSL data package offers a potential solution to overcome the challenges above, offering high-resolution global human settlement information (hereinafter referred to as GHSL-3D Building): building footprint, building type (residential/non-residential), and building height [41].

Considering the above discussion, we proposed a new large-scale GPM method to generate a map of nighttime population distribution in mainland China (excluding Taiwan, Hong Kong, Macau, and some surrounding islands due to data limitations). This map corresponds to the concept of the resident population. We also assessed the accuracy of this method across provinces and municipalities with varying population densities and levels of economic development. This GPM method utilized 3D residential building data (from the newly released GHSL data package 2023), POI, nighttime light data, and land use/cover data within the MGWR model. The contributions of this paper are as follows:

(1) Three-dimensional residential building data were used in GPM for the entire mainland China, considering the effect of building height on the population distribution during the model training and imposing strict limits on the range of population distribution.

(2) Population distribution across mainland China was modeled based on MGWR, considering the nonstationarity and multiscale nature of the spatial relationship between population and auxiliary variables. This approach addresses regional differences in population distribution patterns.

To the best of our knowledge, this is the first time the MGWR model has been applied in the context of GPM and the first instance of employing 3D residential building data for national-level GPM in China. Previous studies have shown that WorldPop has a general accuracy advantage over other gridded population data products [8]. Due to the improvements in the model and auxiliary data, the method presented in this paper is expected to yield results with higher accuracy than the WorldPop dataset, providing a crucial reference for generating next-generation global demographic maps.

This paper is organized as follows: Section 2 describes the sources of research data and the preprocessing steps. Section 3 details the methodology. The results and discussion are presented in Section 4. Section 5 concludes the paper and outlines directions for future research.

2. Data and Preprocessing

Table 1 presents this paper's primary data for modeling and accuracy evaluation. The following describes the sources and preprocessing process for these data.

Table 1. Main research data.

Dataset	Format	Source
Population data	Table	Chinese Bureau of Statistics 2018 national sample survey resident population data
Administrative boundary data	Polygon	National Catalogue Service for Geographic Information, China
Building data	Raster	Global Human Settlement Layer (GHSL)
Nighttime light data	Raster	Earth Observation Group, Colorado School of Mines
Land use/cover data	Raster	Chinese Academy of Sciences Resource and Environment Science Data Center
POI data	Table	Amap Service, China

2.1. Population Data

In our research, the resident population data in the study area, excluding Taiwan, Hong Kong, Macau, and some surrounding islands, were obtained from the National Bureau of Statistics 2018 national sample survey resident population data. These data were collected based on the third-level administrative units, encompassing districts and counties, resulting in 2850 units.

The fourth-level resident population (i.e., the resident population at the level of the fourth administrative units) data for Shanghai, Jiangsu, Jiangxi, and Gansu provinces used for the accuracy test were mainly from the 2018 China Statistical Yearbook (Township).

2.2. Administrative Boundary Data

The data concerning the third-level administrative boundaries of China were acquired from the official website of the National Catalogue Service For Geographic Information (China) (https://www.webmap.cn/main.do?method=index) (accessed on 22 July 2023). Additionally, the administrative boundary data for Beijing, Shanghai, Jiangsu, Jiangxi, and Gansu at the fourth level were sourced from the National Platform for Common Geospatial Information Services (China) (https:/www.tianditu.gov.cn/, accessed on 12 September 2024). We linked the population data with the administrative boundary data based on the administrative code and name of the respective administrative units. The distribution of the resident population across the third-level administrative units in the study area is illustrated in Figure 3b. The third-level administrative boundary data were used as an input layer for "mask" and "processing extent" in the geoprocessing tool (ArcGIS Pro software 3.0) to ensure that the boundaries of the various types of data were consistent.

2.3. GHSL-3D Building Data

The global building dataset for 2018 was acquired from the official website of the Global Human Settlement Layer (GHSL) (https://ghsl.jrc.ec.europa.eu/download.php, accessed on 12 September 2024). These datasets are categorized into three types: total building footprint data, non-residential building footprint data, and building height data. The two building footprint datasets have a spatial resolution of 10 m, where each pixel value represents a building area ranging from 0 to 100. The building height data have a spatial resolution of 100 m. In this dataset, each pixel represents the mean net height of all buildings at that location. With reference to studies [42–47] related to building height data, we assessed the accuracy of the GHSL building height dataset in Section S1 of the Supplementary Materials. The results show that the dataset has relatively good accuracy.

Initially, these datasets were in Lambert projection. However, to suit our study area, we transformed the datasets into the Albers projection using the projection and mask extraction tools in ArcGIS Pro 3.0. The resampling method employed was the nearest neighbor. All the capture cells were set to the building height data for this paper.

2.4. Nighttime Light Data

Nighttime light data is widely used in large-scale GPM [18,48]. We obtained the 2018 global VIIRS nightlight data (VNL V2.1 annual version) from the Earth Observation Group (EOG) website of the Colorado School of Mines (https://eogdata.mines.edu/products/vnl/, accessed on 12 September 2024). This dataset has been carefully processed to exclude the influences of cloud cover and background light. The original spatial resolution of the data is 15 arc seconds, approximately equivalent to 500 m at the equator. To suit our study area, we utilized the projection and mask extraction tools available in ArcGIS Pro 3.0. Through this process, we obtained the nighttime light data at a spatial resolution of 100 m. Referring to Gaughan et al. [22], the nearest neighbor method was used to resample nighttime light data to avoid changing pixel values.

2.5. Land Use/Cover Data

The 2018 land use/cover raster data were obtained from the official website of the Chinese Academy of Sciences Resource and Environmental Science Data Center (https://www.resdc.cn/, accessed on 12 September 2024). The dataset primarily relies on Landsat satellite remote sensing imagery, which was manually interpreted. It follows a two-level classification system: Level 1 includes six land classes, namely, cultivated land, forest land, grassland, water area, built-up land, and unused land; Level 2 consists of 25 land classes based on the Level 1 classification system. The original spatial resolution of the data is

30 m, and it is projected onto the Albers projection using the Krasovsky ellipsoid. We converted the data into the Albers projection based on the WGS-84 ellipsoid to suit our study area. This conversion resulted in a spatial resolution of 100 m for the raster data. To minimize accuracy loss during resampling, we utilized the majority resampling method. Furthermore, we reclassified the processed land use/cover data by merging all land classes except for urban land, rural residential land (referred to as rural land), and industrial and mining land (referred to as industrial land) into a single class named 'remaining land'.

2.6. POI Data

POI data can represent various human activities in their location and neighborhood (e.g., companies, restaurants, and financial services) that correlate with population density to varying degrees [18,49]. Therefore, POI data is are often used in GPM studies [8,50]. The POI data used in this study was were collected in 2017 and obtained from Amap (https://ditu.amap.com/, accessed on 12 September 2024), a leading provider of digital maps, navigation, and location-based services in China. The raw text data was were carefully cleaned and transformed into vector points using latitude and longitude information. After this processing, the data was were projected for further analysis.

In our research, we utilized 13 types of POI data, including shopping services, government organizations and social groups, health care services, lifestyle services, car maintenance, catering services, sports and leisure services, financial and insurance services, companies and enterprises, car services, education and cultural services, car sales, and motorcycle services. These 13 types collectively amounted to 38,154,240 records, forming the basis for our analyses and investigations.

3. Method

The steps of processing research data, fitting the MGWR model to predict the population at the grid level, and evaluating accuracy are shown in Figure 1.

3.1. Processing of Research Data

3.1.1. Processing of Building Data

In the initial stage of our study, we performed a difference calculation between the total building footprint data and the non-residential building footprint data to obtain the residential building footprint data. Referring to the official manual of the GHSL data package (https://data.europa.eu/doi/10.2760/19817, accessed on 12 September 2024), the ratio of population density between fully non-residential building units (pixels with a non-residential building area of 100 m^2) and fully residential building units (pixels with a residential building area of 100 m^2) is 0.04915 [41]. Assuming that the population located in buildings is evenly distributed according to this ratio, we calculated that the population distributed in non-residential buildings accounted for only 3.31‰ of the study area's total population. Given the general accuracy of GPM and the processing time of the data, we assumed that the population is only distributed in residential buildings for this study. As a result, we aggregated the residential building footprint data into 100 m pixels to match the building height data. The final step was multiplying the residential building footprint data by the height data, thus obtaining the volume data of residential buildings in China.

To recognize significant population density variations among residential buildings on different land classes, such as higher density in residential buildings on urban land compared to rural land, we used preprocessed land use/cover data for residential building classification. We selected four land classes for population distribution: urban, rural, industrial, and remaining land, as there are notable differences in population density among these categories [5]. Through an overlay analysis of the land use/cover data and residential building volume data, we derived the volume data for residential buildings in urban, rural, industrial, and remaining land. Additionally, considering that residential buildings in illuminated areas generally have higher population density than unilluminated areas [5], the four residential building datasets were again overlaid with the preprocessed nighttime light data to classify the four residential buildings into illuminated and unilluminated areas. All overlay analysis operations of raster data are realized through the "Con()" function of the raster calculator tool of the ArcGIS Pro software. In conclusion, we obtained a comprehensive dataset of eight types of residential buildings in China, including the volume (in cubic meters) of residential buildings on illuminated urban, rural, industrial, and remaining land, as well as the volume (in cubic meters) of residential buildings on unilluminated urban, rural, industrial, and remaining land.

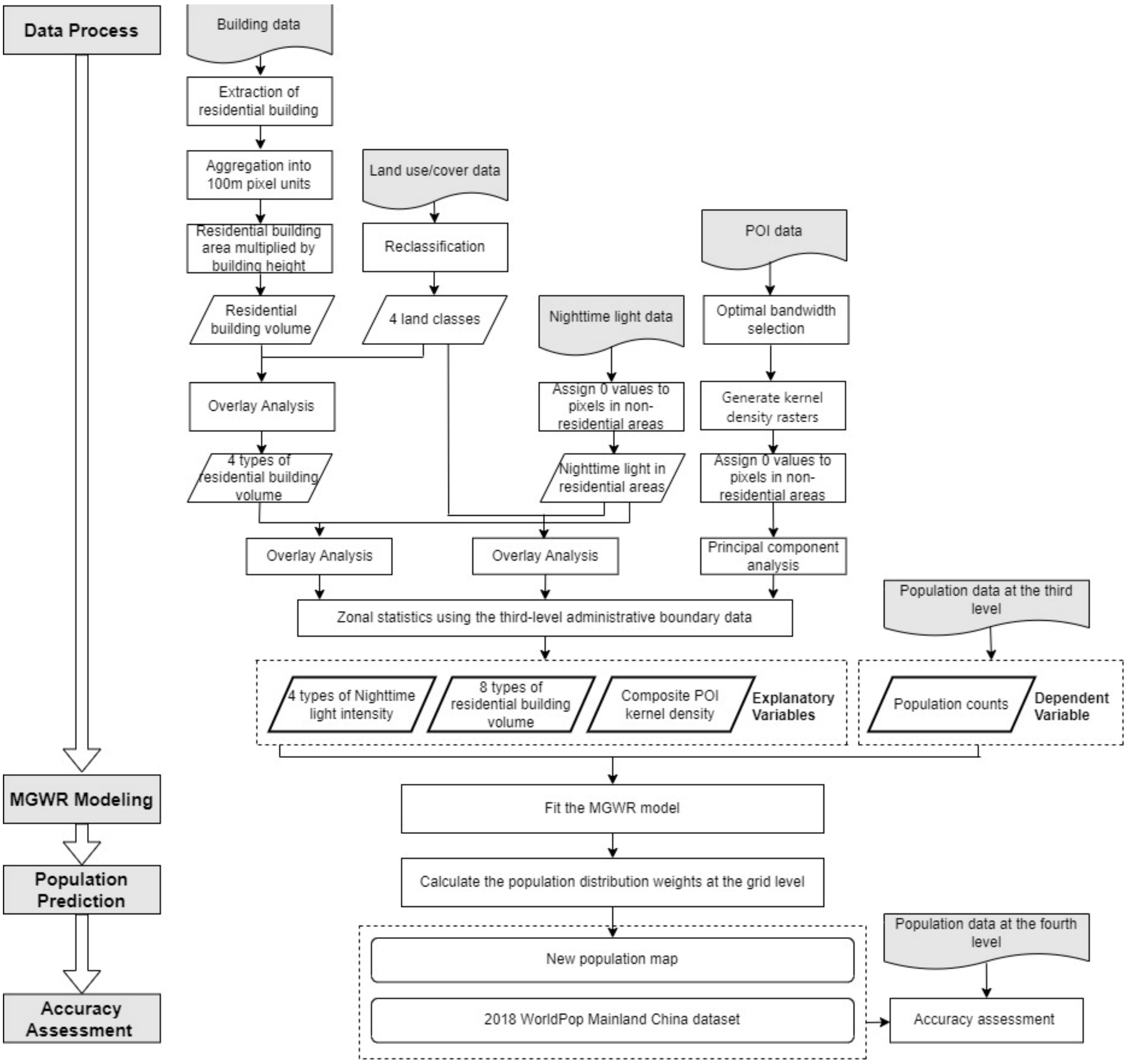
Data Process
MGWR Modeling
Population Prediction
Accuracy Assessment
Building data
Extraction of residential building
Aggregation into 100m pixel units
Land use/cover data
Residential building area multiplied by building height
Reclassification
POI data
Residential building volume
4 land classes
Nighttime light data
Optimal bandwidth selection
Overlay Analysis
Assign 0 values to pixels in non-residential areas
Generate kernel density rasters
4 types of residential building volume
Nighttime light in residential areas
Assign 0 values to pixels in non-residential areas
Overlay Analysis
Overlay Analysis
Principal component analysis
Zonal statistics using the third-level administrative boundary data
Population data at the third level
4 types of Nighttime light intensity
8 types of residential building volume
Composite POI kernel density
Explanatory Variables
Population counts
Dependent Variable
Fit the MGWR model
Calculate the population distribution weights at the grid level
New population map
2018 WorldPop Mainland China dataset
Population data at the fourth level
Accuracy assessment

Figure 1. Research workflow.

3.1.2. Processing of Nighttime Light Data

The subject of this study is GPM during nighttime. To ensure that our final results closely align with reality (i.e., the population is only distributed within residential buildings), we applied a value of 0 to the nighttime light intensity at locations where the volume of residential buildings was 0.

Subsequently, we employed preprocessed land use/cover data to extract four separate nighttime light data layers by overlay analysis. These layers represent the nighttime light intensities in urban, rural, industrial, and remaining residential areas. By differentiating between these categories, we can capture variations in nighttime light intensity based on the specific land use patterns, which enables us to study population distribution with greater precision and accuracy.

3.1.3. Processing of POI Data

We employed the kernel density analysis tool on the 13 types of POI data, quantitatively expressing the density of each POI type in continuous raster cells. During the kernel density analysis, we utilized administrative boundary data, excluding non-residential areas, to aggregate each type of kernel density raster by summation to the residential building areas at the third-level administrative units. This allowed us to select an appropriate bandwidth. We calculated the Pearson correlation coefficients between the aggregated POI density with different bandwidths and the population counts. The tests were conducted at 400 m intervals, ranging from 400 to 8000 m. After a thorough evaluation, we found the highest correlation coefficients between most categories (8 categories) of POI and population at a bandwidth of 800 m. Consequently, we set the kernel density analysis for all 13 types of POI using this 800 m bandwidth. The pixel size of all kernel density rasters was set at 100 m. A summary of the correlation coefficients at different bandwidths is shown in Table S2 in the Supplementary Materials.

To ensure that the population was exclusively distributed within residential building areas, we assigned zero values to all POI kernel density pixels where the residential building volume was 0. Furthermore, to optimize the model fitting and prediction time, we followed the approach proposed by Yang and Ye et al. [50] to reduce the kernel density raster of all POIs to a composite POI kernel density raster layer using principal component analysis (PCA). The explanatory variables derived from auxiliary data are shown in Table 2.

Table 2. Description of explanatory variables derived from auxiliary data.

Name	Acronyms	Unit	Description
Illuminated urban residential volume	IUV	m^3	The volume of residential buildings in urban land where the nighttime light intensity value is more significant than zero
Unilluminated urban residential volume	UUV	m^3	The volume of residential buildings in urban land where the nighttime light intensity value is equal to zero
The nighttime light intensity of urban residential areas	NTLU	/	The nighttime light intensity of residential building areas in urban land
Illuminated rural residential volume	IRuV	m^3	The volume of residential buildings in rural land where the nighttime light intensity value is more significant than zero
Unilluminated rural residential volume	URuV	m^3	The volume of residential buildings in rural land where the nighttime light intensity value is equal to zero
The nighttime light intensity of rural residential areas	NTLRu	/	The nighttime light intensity of residential building areas in rural land
Illuminated industrial, residential volume	IIV	m^3	The volume of residential buildings in industrial land where the nighttime light intensity value is more significant than zero
Unilluminated industrial, residential volume	UIV	m^3	The volume of residential buildings in industrial land where the nighttime light intensity value is equal to zero
The nighttime light intensity of industrial residential areas	NTLI	/	The nighttime light intensity of residential building areas in industrial land
Illuminated remaining residential volume	IReV	m^3	The volume of residential buildings in the remaining land where the nighttime light intensity value is more significant than zero
Unilluminated remaining residential volume	UReV	m^3	The volume of residential buildings in the remaining land where the nighttime light intensity value is equal to zero
The nighttime light intensity of the remaining residential areas	NTLRe	/	The nighttime light intensity of residential building areas in urban land
Composite POI kernel density value	POI	/	Composite value of POI kernel density for 13 categories

3.2. MGWR Model and Population Prediction

3.2.1. Theories Related to MGWR

The formula for the MGWR model is

$$y_i = \sum_{j=1}^{m} \beta_{bwj}(u_i,v_i)x_{i,j} + \varepsilon_i \quad (1)$$

where y_i represents the dependent variable for the ith region, $x_{i,j}$ represents the jth explanatory variable for the ith region, $\beta_{bwj}(u_i,v_i)$ represents the coefficient of the jth explanatory variable for the ith region, m represents the number of explanatory variables, ε_i represents the random error term for the ith region, and bwj represents the bandwidth used for the coefficient of the jth explanatory variable.

Each coefficient of the MGWR model has a different bandwidth, while all coefficients of the GWR model have the same bandwidth. This is the main difference between MGWR and GWR models. The GWR model uses weighted least squares to obtain the coefficients. The coefficients and bandwidths of the MGWR model are obtained through a method known as the Back-Fitting Algorithm (BFA), which was initially used to estimate parameters for Generalized Additive Models (GAM) [51]. Following the logic of GAM, the term $\beta_{bwj}(u_i,v_i)x_{i,j}$ is defined as the jth additive term f_j, leading to the GAM-style MGWR

$$y = \sum_{j}^{m} f_j + \varepsilon \tag{2}$$

where y represents the truth value. ε represents the residual, f_j represents the jth additional term, and m represents the number of additional terms.

Based on the above, the MGWR model is estimated following the process in Figure 2.

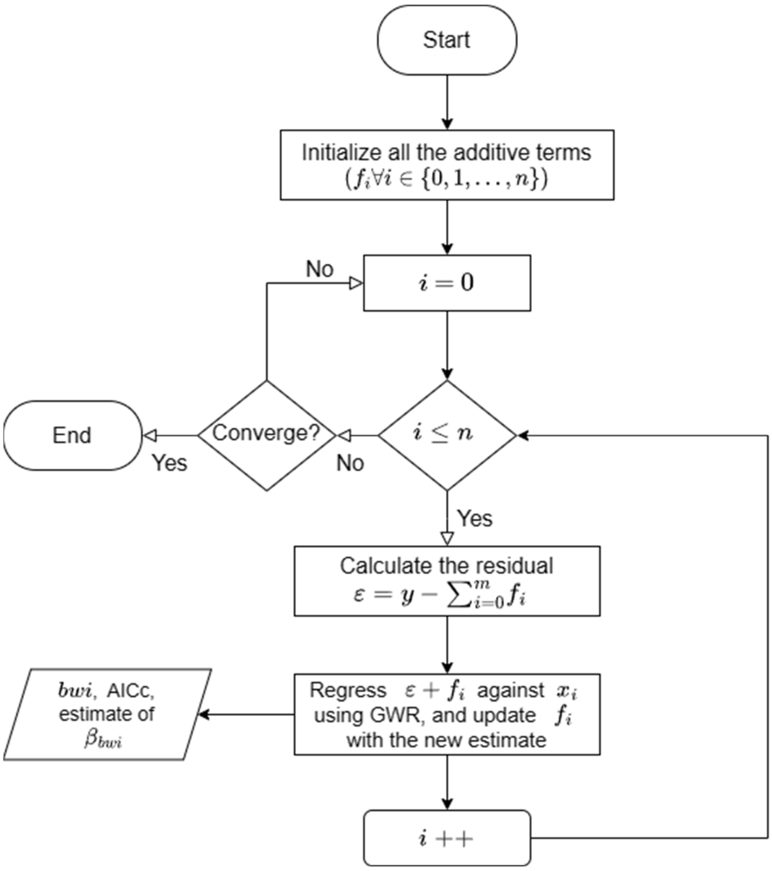
Start
Initialize all the additive terms $(f_i \forall i \in \{0,1,\ldots,n\})$
No
$i = 0$
End
Yes
Converge?
No
$i \leq n$
Yes
Calculate the residual $\varepsilon = y - \sum_{i=0}^{m} f_i$
bwi, AICc, estimate of β_{bwi}
Regress $\varepsilon + f_i$ against x_i using GWR, and update f_i with the new estimate
i ++

Figure 2. The process of estimating MGWR by BFA.

The specific estimation process is described below. Firstly, it is necessary to initialize all the additional terms f_j in Formula (2), that is, to obtain initial estimates of all coefficients β_{bwj} for the MGWR model. There are generally three choices for the initial estimation method: (1) set all coefficients to 0; (2) MLR estimation; (3) GWR estimation. Theoretically, the initial estimation method will only affect the number of iterations and not the final bandwidth result [26]. After the initialization, the initial residuals ε are calculated by

$$\varepsilon = y - \sum_{j=1}^{m} f_j \tag{3}$$

where ε is the residual, y is the truth value, f_j is the estimate of the jth additional term, and m is the number of additional terms. Then, GWR is performed to obtain the new optimal bandwidth $bw1$ and the new coefficient β_{bw1} by taking the sum of the residual ε and the estimate f_1 of the first additional term as the dependent variable and the first variable x_1 as the single explanatory variable. The new coefficient β_{bw1} is used to calculate the new f_1 and new ε, which replace the corresponding old values. Next, GWR is performed to obtain the new optimal bandwidth $bw2$ and the new coefficient β_{bw2} by taking the sum of the residual ε and the estimate f_2 of the second additional term as the dependent variable and the second variable x_2 as the single explanatory variable. This process is repeated until the last variable x_m. The above process is repeated as one step until the final estimate satisfies the convergence condition.

This article uses the classic Proportional Change in the Residual Sum of Squares (PCRSS) as the convergence criterion:

$$SOC_{RSS} = \frac{|RSS_{new} - RSS_{old}|}{RSS_{new}} \tag{4}$$

RSS_{old} represents the residual sum of squares from the previous step, and RSS_{new} represents the residual sum of squares from the current step.

3.2.2. Fitting the Models and Population Prediction

In this study, we aggregated the rasters of explanatory variables (listed in Table 2) using the third-level administrative boundaries as regional elements. We then employed the GWmodelS software (https://www.sciencedirect.com/science/article/pii/S2352711022002096, accessed on 12 September 2024) to fit three regression models: MLR, GWR, and MGWR. GWmodelS is a versatile software that integrates various geographically weighted models, and its graphical user interface allows for quick and easy model construction.

In order to investigate the impact of building height data on population distribution modeling, a comparison experimental group was created. It did not utilize residential building volumes (i.e., no building height information added) but used residential building area, nighttime light intensity, and POI kernel density as explanatory variables. Pixels of the explanatory variables for non-residential building areas are set to 0 to ensure that the population is distributed only in residential buildings. We summarized the explanatory variable rasters for the comparison experimental group in the same way. Then, we built three regression models (MLR, GWR, and MGWR)_for the comparison experimental group to explore the relationship between the explanatory variables and the population.

We used the MGWR model with building height and POI data added to disaggregate the Chinese population data. The standard experimental group's 13 explanatory variable layers (listed in Table 2) were used by the MGWR model to predict population distribution weights on 100 m grid cells. Specifically, the regression coefficients from the MGWR model were first converted to raster layers at 100 m spatial resolution. Then, raster computations were performed with the explanatory variable layers in the order of the model to obtain the population distribution weights layer. This follows the assumption of scale invariance of the relationship between population and explanatory variables. Figure 3 displays the population distribution weights on 100 m grids and mainland China's third-level resident population counts. Finally, following Formula (5), all third-level population counts were disaggregated into 100 m grid cells, resulting in the population distribution map of mainland China with a spatial resolution of 100 m, as shown in Figure 4.

$$\mathrm{POP}_{\mathrm{grid}} = \mathrm{POP}_{\mathrm{ADM}} \times \frac{\mathrm{W}_{\mathrm{grid}}}{\mathrm{W}_{\mathrm{ADM}}} \tag{5}$$

where $\mathrm{POP}_{\mathrm{grid}}$ represents the population counts of a particular grid unit, $\mathrm{POP}_{\mathrm{ADM}}$ represents the resident population counts of the third-level administrative unit corresponding to the grid unit, $\mathrm{W}_{\mathrm{grid}}$ represents the population distribution weight of the grid unit, $\mathrm{W}_{\mathrm{ADM}}$ represents the total population distribution weight of the third-level administrative unit corresponding to the grid unit.

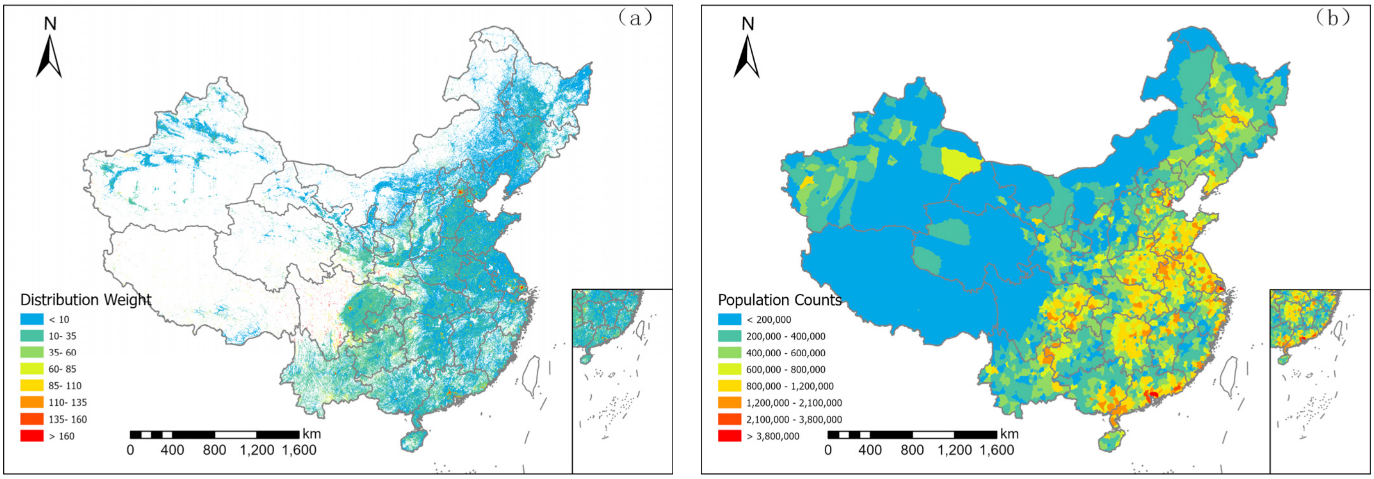

Figure 3. (**a**) Population distribution weight map of mainland China. (**b**) Third-level resident population count map of mainland China.

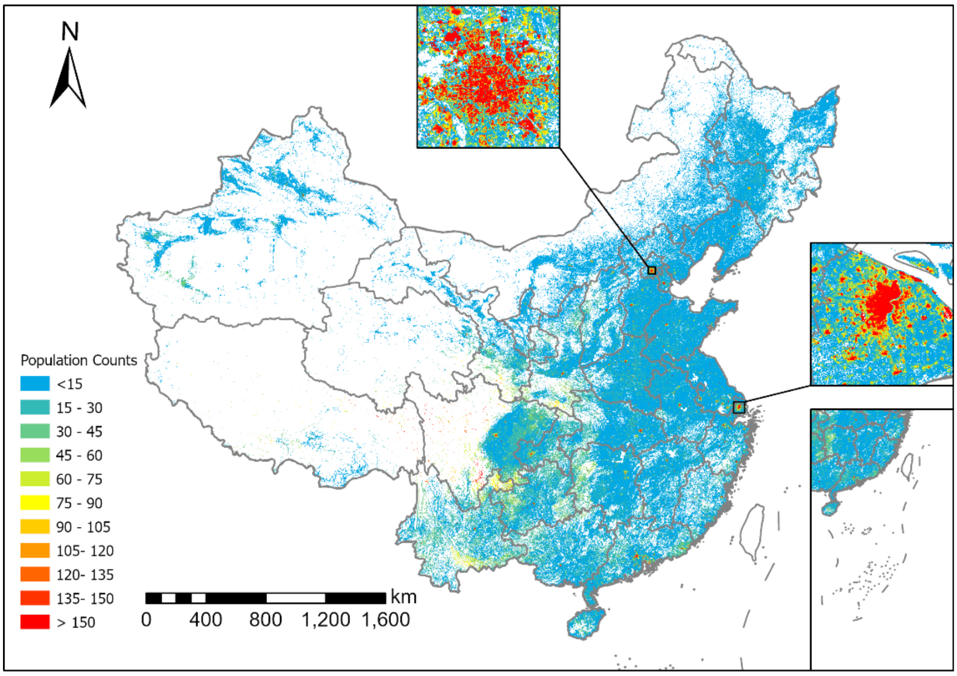

Figure 4. The 100 m spatial resolution nighttime population distribution map of mainland China.

3.3. Evaluation of Accuracy

To comprehensively evaluate the accuracy of the new gridded population map in diverse regions, we tested it using 2018 fourth-level population data from three regions with varying overall population density and economic development levels. Group 1 encompassed areas with high overall population density and economic development, namely Beijing, Shanghai, and Jiangsu Province (Jiangsu Province was included due to the small sample size of Beijing and Shanghai), resulting in 2037 test samples. Group 2 represented regions with moderate overall population density and economic development, namely Jiangxi Province, providing a total of 1780 test samples. Group 3 consisted of regions with low overall population density and economic development, namely the Gansu Province, with 1417 test samples.

We employed two indicators to quantify the estimation errors of the new gridded population map: the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). In addition, we calculated the ratio of MAE and RMSE to the mean value of the fourth-level population counts in each group to mitigate the impact of overall population density differences on the MAE and RMSE indicators. Finally, to gain insights into the performance differences, we compared the above indicators of the new gridded population map with the 2018 United Nations-adjusted WorldPop population count product.

4. Results and Discussion

4.1. Results

4.1.1. Accuracy Assessment

Figure 5 displays the RMSE and MAE indicators for the new gridded population map and WorldPop. Group 1's MAE values for the new gridded population map and WorldPop are 20,698.79 and 31,985.07, respectively. For Group 2, the corresponding MAE values are 7716.43 and 15,620.94, while in Group 3, they are 7861.94 and 9049.98. Regarding the RMSE, Group 1 shows values of 37,914.52 for the new gridded population map and 56,201.27 for WorldPop. For Group 2, the RMSE values are 13,175.15 and 26,652.89; for Group 3, they are 13,856.28 and 18,143.99. Compared to WorldPop, the MAE values of the new gridded population map were reduced by 35.28%, 50.60%, and 13.13% for the three groups, respectively. The RMSE values of each group were reduced by 32.54%, 50.57%, and 23.63%, respectively. For the new gridded population map, the R^2 for the three groups are 0.64, 0.78, and 0.67, respectively; for WorldPop, they are 0.56, 0.70, and 0.51, respectively. Therefore, the new gridded population map has higher overall accuracy than WorldPop.

Table 3 presents the ratios of MAE and RMSE to the mean of the fourth-level population counts. The results reveal that WorldPop exhibits the lowest accuracy in regions with medium population density and economic development (Group 2) (ME = 0.60, RE = 1.03). This lower accuracy in Group 2 may be attributed to the complex urban living environments and diverse population distribution patterns, posing challenges for accurate modeling. In contrast, the new gridded population map demonstrates improved accuracy in all groups, particularly in Group 2, where WorldPop performs least effectively (R_{ME} = 0.30, R_{RE} = 0.52). This enhancement may be attributed to the inclusion of building height and POI data, which help better describe complex living environments, particularly in urban areas. Additionally, the MGWR model's robust applicability in simulating diverse population distribution patterns contributes to the overall improvement in accuracy.

Table 3. The ratio of MAE and RMSE to the mean of the four-level population counts.

		WorldPop	**New**	**Residual (R)**
Group 1	MAE/Pop (ME)	0.51	0.33	0.18
	RMSE/Pop (RE)	0.90	0.60	0.30
Group 2	MAE/Pop (ME)	0.60	0.30	0.30
	RMSE/Pop (RE)	1.03	0.51	0.52
Group 3	MAE/Pop (ME)	0.48	0.41	0.07
	RMSE/Pop (RE)	0.96	0.74	0.22

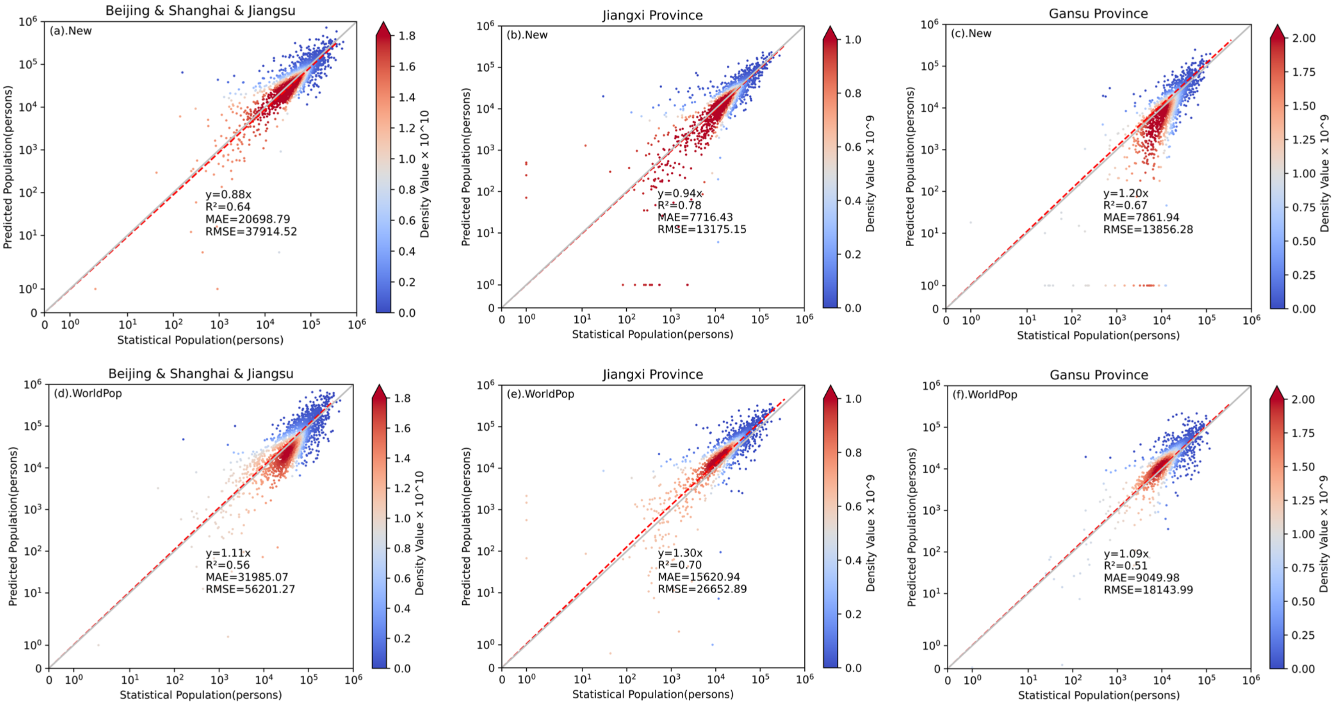

Figure 5. Scatterplots and accuracy indices between predicted and statistical populations at the fourth-level administrative units. Each point denotes a fourth-level administrative unit.

4.1.2. The Differences between the New Gridded Population Map and WorldPop

Using a consistent symbol system, we conducted a thorough investigation comparing the impact of the new gridded population map and WorldPop in three cities: Shanghai, Nanchang, and Lanzhou. The residuals between the two datasets were computed, revealing notable differences in Figure 6c. Both datasets exhibit concentrated population hotspots (depicted in red) in central districts, and cities with lower overall population density and economic development have fewer hotspots outside the central areas, indicating a stronger population attraction towards the city centers. However, WorldPop and the new gridded population map clearly differ in describing the transition between population hotspots and coldspots (depicted in blue). WorldPop showcases a discontinuous, cliff-like variation, creating a visual effect of the square-shaped and fragmented hotspots, which deviates from reality. Conversely, the new gridded population map presents a smoother and more realistic transition with milder variations at the edges of the hotspots. Furthermore, it becomes evident that the central hotspot area of the new gridded population map is slightly larger than that of WorldPop.

At the grid level, as shown in Figure 6c, the residual map further highlights the differences between the two datasets. The discrepancies are particularly evident at the micro-level, mainly in the hotspots and surrounding areas of the new gridded population map. The residual map depicts a mixed state where cold spots (in red) surround many hotspots (in blue). This could be attributed to the new gridded population map allocating more population in residential-intensive areas within the same third-level administrative unit. In comparison, WorldPop allocates more population in areas where residential buildings are sparse and absent. Additionally, as a city's overall population density and economic development level decrease (i.e., figures from left to right), the mixed area of cold spots and hotspots tends to shrink and decrease towards the central urban area. This observation aligns with the fact that cities with lower overall population density and economic development tend to concentrate residential buildings in the central district and its surrounding areas.

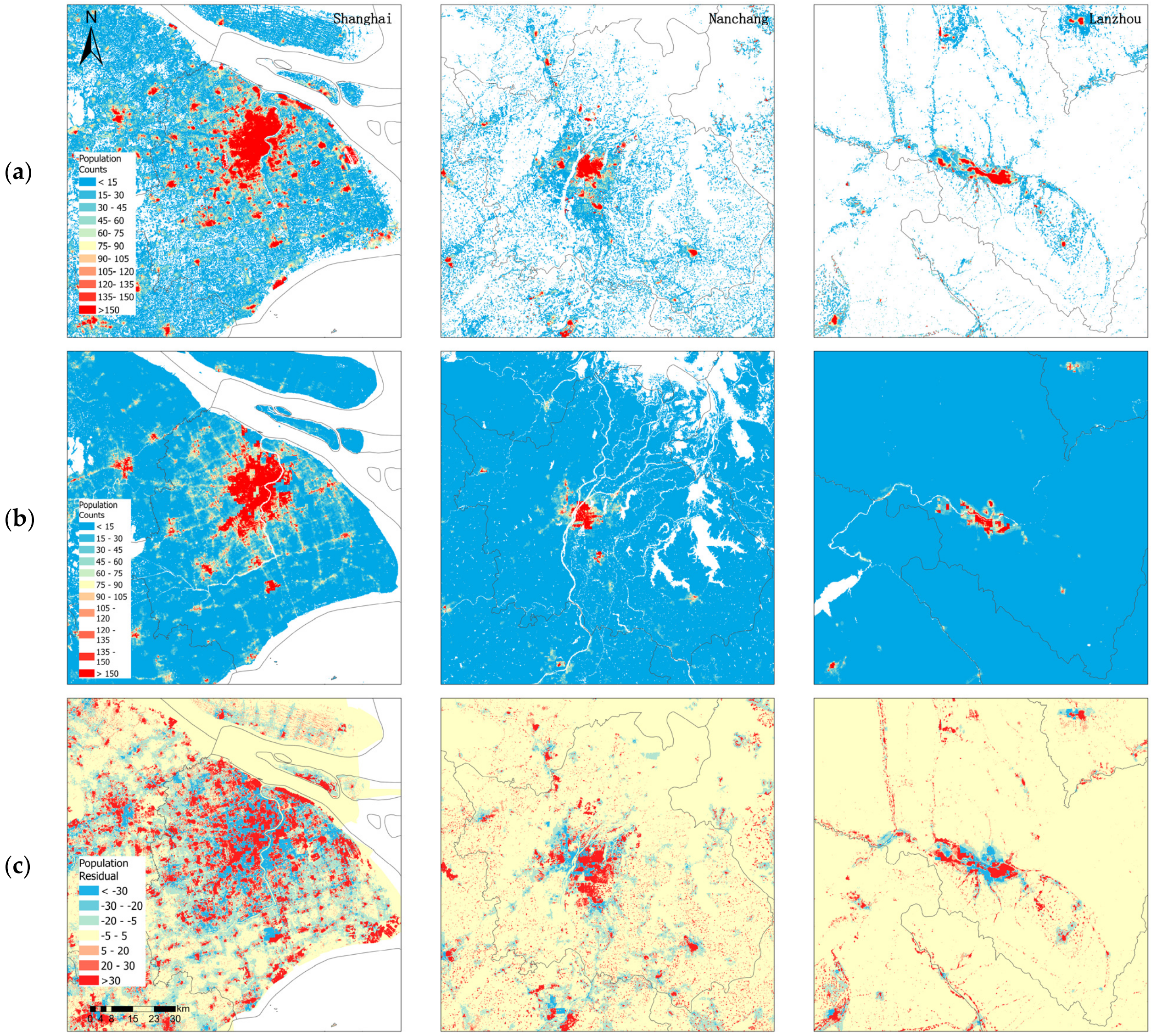

Figure 6. The new population gridded map and WorldPop in three cities and their differences. Row (**a**) displays the new gridded population map, row (**b**) displays WorldPop, and row (**c**) displays the difference between the new gridded population map and WorldPop.

4.2. Discussion

4.2.1. Performance Evaluation of Different Models

Table 4 provides two essential indicators, Adjusted R^2 and AICc, which were utilized to assess the fitting performance of all regression models in the comparative experiment involving building height information. Throughout the addition of building height information, the Adjusted R^2 of the MLR, GWR, and MGWR models demonstrated consistent increments, and their respective AICc values decreased. This indicates that adding building height information effectively enhances the effects of population distribution simulation. The MGWR model exhibited the highest overall accuracy, albeit with limited improvement over the GWR model. This suggests that the spatial relationship between the selected factors and the population has significant nonstationarity and some multiscale nature.

Table 4. Overall fitting results of the models.

		MLR	GWR	MGWR
Building area	Adjusted R^2	0.758	0.916	0.920
	AICc	77,548.582	75,318.649	75,056.819
Building volume	Adjusted R^2	0.797	0.926	0.928
	AICc	77,182.356	74,997.368	74,779.339

Table 5 shows the overall accuracy information of the gridded population datasets generated by the six models for the five regions of Beijing, Shanghai, Jiangsu, Jiangxi, and Gansu. This represents the models' ability to predict population at the grid level. The ratios of RMSE and MAE to the mean population counts at the fourth-level administrative units (RMSE/Pop, MAE/Pop) are used as accuracy indicators. As shown in Table 5, incorporating building height information significantly improves the accuracy of population predictions. Furthermore, improvements in the model parameter estimation methods (from MLR to GWR to MGWR) also result in varying degrees of enhanced population prediction accuracy. These findings are consistent with the model fitting performance shown in Table 4.

Table 5. The accuracy of gridded population datasets generated by different models.

		MLR	GWR	MGWR
Building area	RMSE/Pop	1.19	0.86	0.80
	MAE/Pop	0.69	0.45	0.42
Building volume	RMSE/Pop	1.07	0.73	0.68
	MAE/Pop	0.58	0.37	0.33

4.2.2. Scale Analysis

Table 6 shows the bandwidth information for the GWR and MGWR models using building volume variables. MGWR can directly reflect the different action scales of different variables, while GWR can only reflect the average spatial scales of all variables. The bandwidth of GWR is 19, which is only 0.67% of the total sample size. In MGWR, the spatial scales of different variables vary considerably, with most of (i.e., more than 50% of administrative units at the third level) the regression coefficients for the eleven variables, namely, intercept, UIV, NTLI, IReV, UReV, IRuV, NTLRu, URuV, IUV, UUV, and POI, being significant, whereas most of (i.e., more than 50% of administrative units at the third level) the regression coefficients for the three variables, namely, IIV, NTLRe, and NTLU, are not significant.

Specific manifestations are:

(1) Intercept, UIV, NTLI, IReV, UReV, IRuV, NTLRu, IUV, and POI have bandwidths ranging from 11 to 29, exhibiting micro-scale features in the model. The spatial scale is close to the level of prefecture-level cities in mainland China. On the one hand, this indicates that they exhibit considerable spatial nonstationarity in the model. Once the spatial range is exceeded, the coefficients will change dramatically. On the other hand, it also proves that the population is sensitive to these variables in this modeling approach.

(2) The bandwidths of UUV and URuV are 115 and 116, respectively, close to the regional spatial scales of general provincial administrative units in China, suggesting that they have relatively medium spatial nonstationarity in the modeling.

Table 6. Bandwidth of GWR and MGWR models.

Variable	MGWR	GWR
Intercept	11	19
IIV	2847	19
NTLI	11	19
UIV	29	19
IReV	20	19
NTLRe	11	19
UReV	11	19
IRuV	15	19
NTLRu	11	19
URuV	116	19
IUV	11	19
NTLU	2845	19
UUV	115	19
POI	11	19

4.2.3. Analysis of Coefficient Spatial Pattern

Table 7 presents statistical information on the regression coefficients corresponding to the significant explanatory variables across the study area (note: non-significant samples are also counted). The spatial distribution of regression coefficients corresponding to these explanatory variables is shown in Figure 7 (note: the regression coefficients for non-significant coefficients are all negative). We only present UUV and POI for reasons limited by the length of the article.

Table 7. Statistical description of MGWR regression coefficient.

Variable	Mean	Standard Error	Min	Median	Max
Intercept	116,030.2	61,381.3	5774.5	106,166.0	339,670.9
IUV	0.00104	0.00095	−0.00180	0.00105	0.00304
UUV	−0.00364	0.03400	−0.17500	0.00096	0.06400
IRuV	0.00240	0.00335	−0.00390	0.00134	0.01940
URuV	0.00320	0.00129	0.00025	0.00330	0.01040
NTLRu	1.32000	6.40000	−26.60000	1.97000	21.50000
UIV	0.03100	0.19000	−0.42000	0.01080	2.90000
NTLI	1.70000	2.80000	−13.10000	1.21000	20.00000
IReV	0.00290	0.00233	−0.00520	0.00290	0.01110
UReV	0.00940	0.00860	−0.02570	0.01040	0.05500
POI	24.6000	15.5000	−8.80000	24.5000	83.0000

The regression coefficients for UUV are significant in 1474 sample units, accounting for 51.72% of all sample units. As seen in Figure 7c, the non-significant coefficients are mainly distributed over most of the northeastern region and all areas below the southwest diagonal of mainland China. The high values of the significant coefficients are mainly distributed in Shandong, Jiangsu, Southern Anhui, Eastern Henan, part of Inner Mongolia, Northern Hebei, and Western Liaoning. From Table 7, the UUV coefficients in the study area range from −0.175 to 0.064, with a mean value of −0.00364 and a standard error of 0.034. This indicates that the impact of UUV per 1 million m^3 on the population of the third-level administrative units ranges from −175,000 to 64,000 people, with an average impact of −3640 people, and there is a relatively significant difference in the impact of the UUV on the population in different regions. An increase in UUV leads to a decrease in IUV when urban residential buildings are held constant. In addition, the presence of high UUV values suggests that the local economic vitality may be relatively low, and therefore, there may be some degree of population loss. These may be the primary reasons why the average impact of UUV on the population is negative.

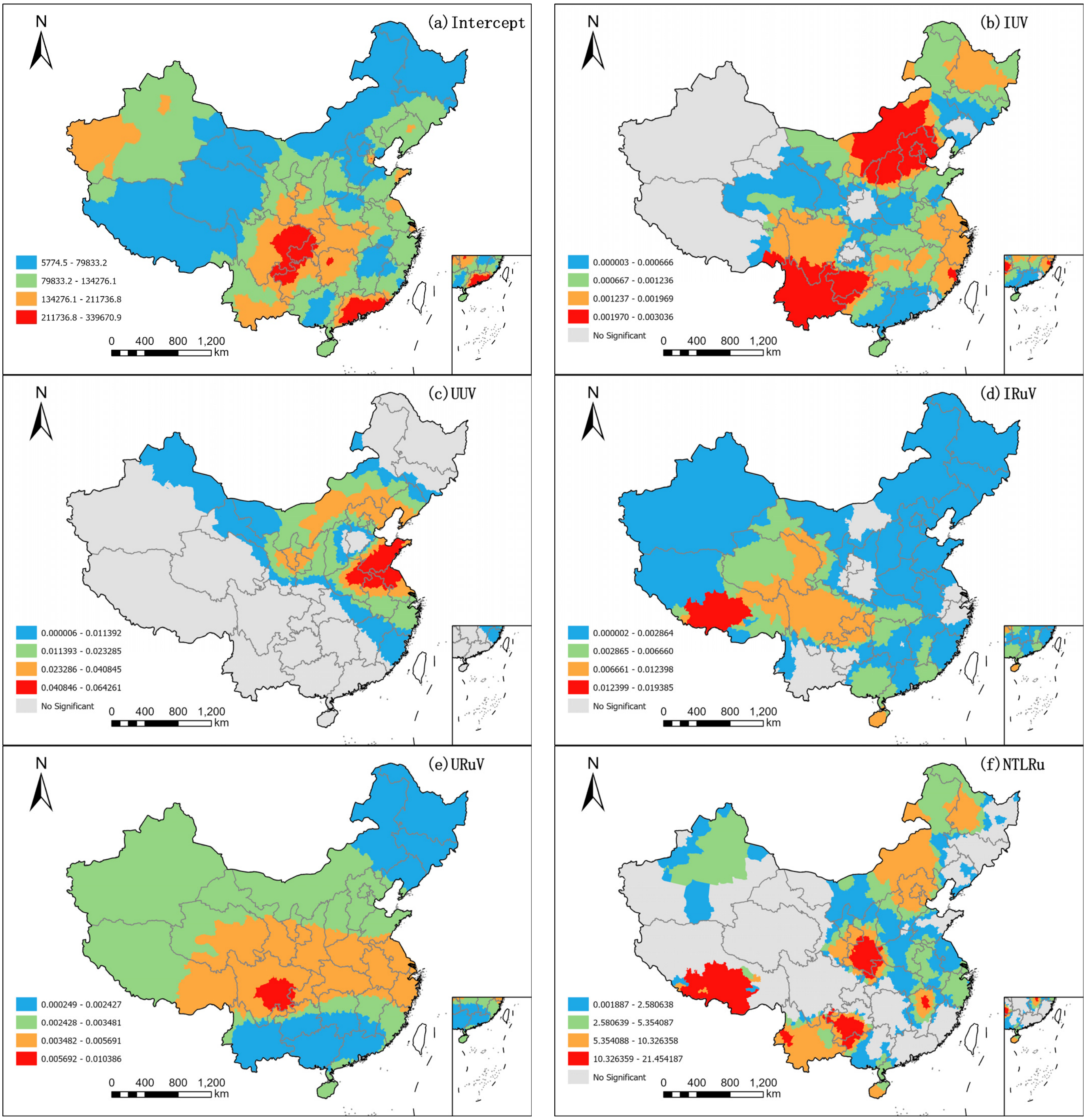

Figure 7. *Cont.*

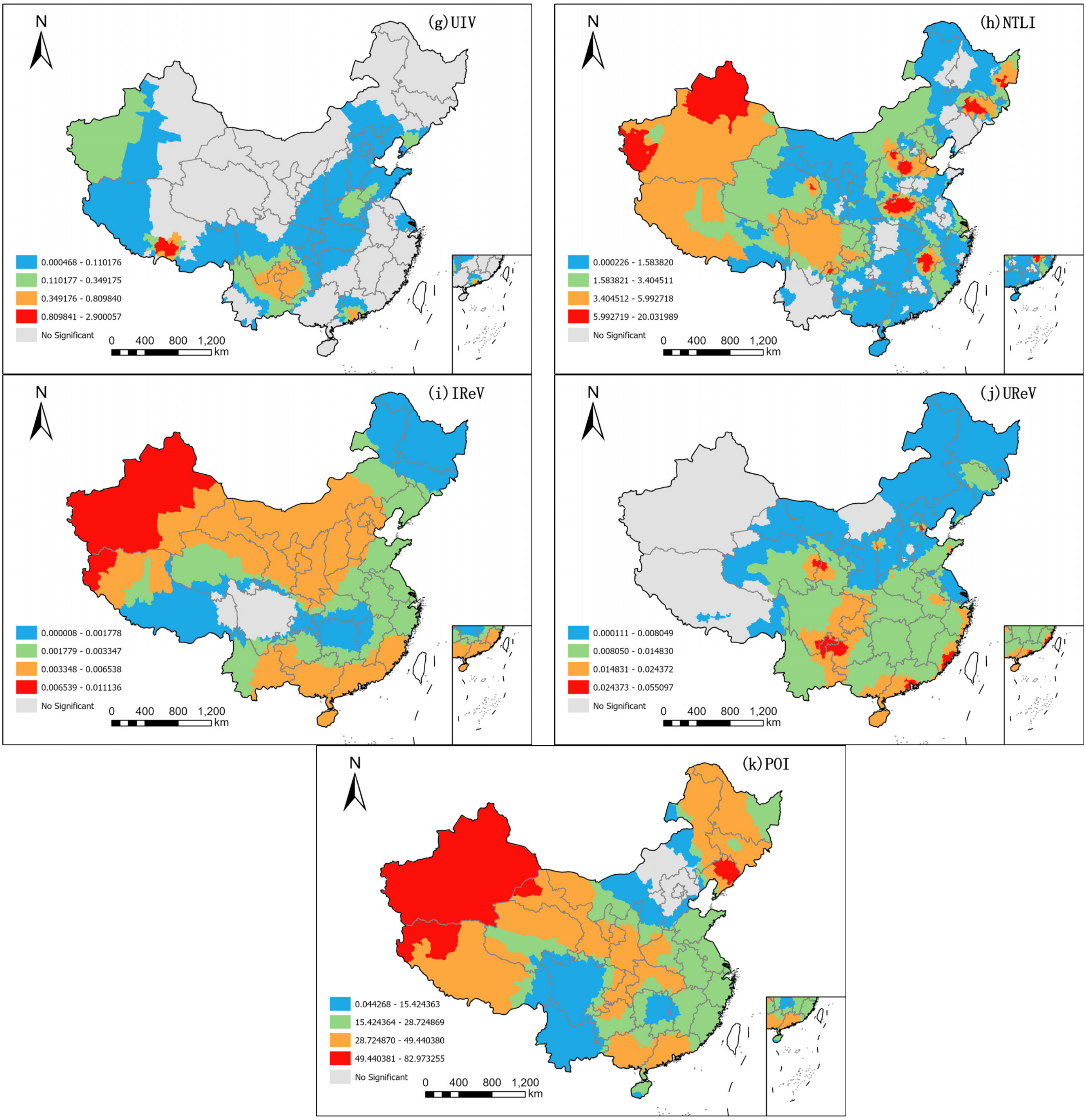

Figure 7. Spatial distribution of coefficients corresponding to explanatory variables in MGWR fitting results.

In total, 2613 sample units show significant POI coefficients, representing 91.68% of the total samples. From Figure 7k, the areas with non-significant POI coefficients are mainly concentrated in Beijing–Tianjin–Hebei and parts of Inner Mongolia. The high values of the significant POI coefficients are mainly in most of Northwest China, most of Northeast China, parts of Central China, Tibet, most of Guangdong, and most of Guangxi. Surprisingly, the POI regression coefficients are higher in many economically underdeveloped regions (e.g., Tibet, Northwest China, and Northeast China) than in economically developed regions (e.g., southeastern coastal region). This may be due to differences in the spatial equality of economic development. Compared to economically developed regions, economically underdeveloped regions tend to have higher spatial inequality in the level of economic development, and thus, POI, which represents economic dynamism, is more attractive to people, i.e., the population is more dependent on POI. From Table 7, the POI coefficients for the whole study area range from −8.8 to 83, with a mean value of 24.6 and a standard error of 15.5, which represents an average impact of 2460 people per 100 values of composite POI kernel density on the population of the third-level administrative units in the whole study area, with relatively significant differences in the impact of composite POI kernel density on the population in different regions.

In conclusion, most regression coefficients are significant in the above modeling, indicating the validity of our modeling approach. In addition, the distribution of the regression coefficients for each explanatory variable shows noticeable regional differences, indicating that the spatial relationship between the population and these variables is nonstationary. The varying scale of the coefficients indicates that the spatial relationship between the population and these variables is multiscale. Therefore, MGWR is more reasonable than MLR and GWR for modeling population distribution at mainland China's third-level administrative unit scale.

4.2.4. Residential Buildings Versus Land Use/Cover

Land use/cover data are primarily auxiliary data for population distribution modeling [5,20,23]. However, such data may not accurately reflect the actual distribution range of populations, i.e., human settlement areas, nor quantitatively represent population density at specific locations within a land class. Instead, it can only provide a qualitative representation of the average population density among different land classes (e.g., urban land's average population density is higher than rural land's). Directly utilizing land use/cover data for population distribution modeling may fail to reveal variations in population density within each land class, resulting in significant inconsistencies with reality [8]. In contrast, residential building data, serving as human settlement data, can provide a more precise and realistic representation of population distribution ranges. Including height attributes allows quantifying population density differences among buildings of the same type.

Table 8 presents statistical information on various types of residential buildings and land in this study. Regarding land use/cover data, the vast majority of the area is classified as remaining land (97.31%), followed by rural (1.40%) and urban (0.81%) land, with the smallest percentage of industrial land (0.48%). Correspondingly, the proportion of area covered by each type of residential building follows a similar decreasing order (40.50%, 30.42%, 24.84%, and 4.24%). The row 'Density of various residential buildings (area ratio of various residential buildings to corresponding land)' demonstrates a significant downward trend in building density on urban, rural, industrial, and remaining land (23.48%, 16.85%, 6.67%, and 0.32%, respectively). These findings indicate that only a tiny portion of all land classes is occupied by residential buildings (representing human settlement areas). At the 100 m raster level, this is manifested by the presence of residential buildings in less than 30% of the pixels in the study area, accounting for more than 70% of the pixel values in the final population distribution map being null (Figure 4). Consequently, using land class variables (i.e., land class area, land class proportion) as the sole or primary variables for modeling population distribution at night can result in a large number of people in residential building areas being incorrectly assigned to uninhabited non-residential building areas.

Table 8. Statistical information on various residential buildings and land classes.

		Urban	Rural	Industrial	Remaining	All
Residential building	Area (km^2)	18,277.48	22,384.06	3122.69	29,806.80	73,591.03
	Proportion	24.84%	30.42%	4.24%	40.50%	100%
Land class	Area (km^2)	76,296.54	132,178.94	45,814.05	9,203,427.64	9,457,717.17
	Proportion	0.81%	1.40%	0.48%	97.31%	100%
Density of various residential buildings (area ratio of various residential buildings to corresponding land)		23.48%	16.85%	6.67%	0.32%	0.78%

Table 9 presents the Pearson correlation coefficients (R) between the population and various residential building volumes, building areas, and land areas at the third-level administrative units in this study. As human settlements, the correlation between residential building areas and population is generally significantly higher than between the corresponding land areas and population. Specifically, the correlation coefficients between the areas of urban, rural, industrial, and remaining residential buildings and population (0.733, 0.385, 0.423, and 0.559, respectively) are notably higher than those of land areas (0.697, 0.378, 0.208, and −0.175, respectively). This observation indirectly reflects that residential buildings can more accurately and reasonably represent the population distribution, thus exhibiting a stronger correlation with population. The remaining land area negatively correlates with the population (R = −0.175). This is primarily due to an increase in the area of remaining land, leading to a decrease in land classes (urban, rural, industrial) that are more closely related to population distribution.

Table 9. Correlation coefficient (R) of the population with various residential buildings and land.

R	Urban	Rural	Industrial	Remaining
Residential building volume	0.749	0.438	0.432	0.602
Residential building area	0.733	0.385	0.423	0.559
Land area	0.697	0.378	0.208	−0.175

Regarding human settlement with added height information, the correlation between the volumes of various residential buildings and the population is the highest: the correlation coefficients of urban, rural, industrial, and remaining residential buildings are 0.749, 0.438, 0.432, and 0.602, respectively. In contrast to the remaining land, the correlation coefficients of remaining residential building areas or volumes with population are relatively high positive values (0.559 and 0.602, respectively). This is mainly due to the remaining land accounting for the majority (97.31%, Table 8). Despite the low density of residential buildings (0.32%, Table 8), the total amount of residential buildings is quite substantial (40.50%, Table 8), leading to a relatively strong correlation with population.

Using buildings as the range of population distribution can effectively mitigate the uncertainty caused by nighttime light data. Nighttime lights exhibit a 'blooming' effect, leading to illuminated areas extending beyond the actual concentration areas of lights (such as city centers). Moreover, bodies of water, like lakes and rivers, also contribute significant intensity values in nighttime light images [18]. Table 10 presents statistical information on relevant indicators of nighttime lights in residential and non-residential building areas. Notably, the illuminated area of non-residential building areas (857,278.71 km^2) is more than twice that of residential building areas (415,124.20 km^2), and approximately 30% of the nighttime light intensity values are located in non-residential building areas. When nighttime light intensity values are solely or primarily used as variables without constraining the range of population distribution, a significant portion of the population may be allocated to underdeveloped (non-residential building) areas, resulting in insufficient allocation in urban areas with high population density and excessive allocation in rural and suburban areas with sparse population. This phenomenon can significantly impact the accuracy of population distribution results [18]. However, by adopting buildings as the scope of population distribution, the adverse effects of the 'blooming' effect in nighttime light data can be significantly reduced, resulting in more accurate and reliable population distribution results.

Table 10. Statistics related to nighttime light in residential building areas and non-residential building areas.

	The Sum of Nighttime Light Intensity Values	Percentage of Total Nighttime Light Intensity Values	Illuminated Area (km^2)	Percentage of the Illuminated Area in the Study Area
Residential building area	264,359,623	70.36%	415,124.20	4.39%
Non-residential building area	111,346,418	29.64%	857,278.71	9.06%

5. Conclusions

In this study, we applied the MGWR model to disaggregate population data by integrating 3D residential building, nighttime light, POI, and land use/cover data, creating a 100 m gridded population map for mainland China. As far as we know, this is the first time the MGWR model has been used in the context of GPM and the first instance of employing 3D residential building data for national-level GPM in China. The resulting gridded population map exhibits higher accuracy than the existing WorldPop dataset. This improvement can be attributed to utilizing 3D residential building data and the MGWR model. Unlike land use/cover data, residential building data can more accurately reflect the extent of population distribution and show a stronger correlation with the population; its height information can reflect the vertical distribution of the population within the building and is an excellent auxiliary variable. In addition, for large-scale countries or regions like China, the MGWR model, which takes into account the nonstationarity and multiscale nature of the spatial relationship between population and variables, is very suitable for use in GPM in such study areas because of the relatively significant differences in the population distribution patterns among regions.

This study can be a significant reference for developing the next-generation global gridded population product datasets. As GHSL-3D Building and nighttime light data are globally available, and alternatives to land use/cover and POI data, such as ESA/CCI and OpenStreetMap data, exist, the approach presented here can be applied globally. Regarding global population input data, GPWv4 can be a viable substitute for census data, as population grid products like GHS-POP have employed GPWv4 as input data [41]. In contrast to the RF model used by WorldPop, the MGWR model allows for uniform modeling of all input units globally, eliminating the need for zonal modeling by country or region to control the accuracy of population predictions. As a result, the modeling method in this study significantly reduces the complexity and time required for global population modeling.

However, there is still much room for improvement in this study. The MGWR model used in this study can reflect the spatially localized relationship between population and explanatory variables, but it fails to reveal their nonlinear relationship. The relationship between population and influencing factors is usually nonlinear. In addition, to minimize the problem of collinearity among variables, we combined the kernel density layers of the 13 categories of POI into a single one, resulting in the loss of a large amount of semantic (category) information in the POI data. To address these shortcomings, we plan to combine the local regression idea with nonlinear machine learning algorithms (e.g., RF) to build a new model in our following research. Similar to GWR, for each location point, only some nearby observations are used to build a local model [52,53]. This model can express the spatially nonstationary and nonlinear relationship between the population and the variables and is less sensitive to the problem of covariance between the variables. This is expected to produce results with higher accuracy.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijgi13090335/s1, Table S1: Accuracy of the GHSL building height dataset, Table S2: summary of Pearson's correlation coefficient. Figure S1: Distribution map of BBH, CNBH and GHSL datasets in the core built-up areas of the four cities.

Author Contributions: Zhen Lei: conceptualization, data curation, formal analysis, methodology, validation, visualization, writing—original draft, writing—review and editing. Shulei Zhou: formal analysis, visualization, validation, writing—original draft, writing—review and editing. Penggen Cheng: conceptualization, funding acquisition, project administration, supervision, writing—review and editing. Yijie Xie: data curation, visualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, CASM under Grant 4 February 2022 and the National Natural Science Foundation of China (Grant Nos. 41861052, 42161060).

Data Availability Statement: Most of the data in the paper are available by accessing the Mendeley Data repository (https://data.mendeley.com/datasets/hwz54s535n/1) (accessed on 3 September 2024).

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

References

1. Hales, S.; de Wet, N.; Maindonald, J.; Woodward, A. Potential Effect of Population and Climate Changes on Global Distribution of Dengue Fever: An Empirical Model. *Lancet* **2002**, *360*, 830–834. [CrossRef] [PubMed]
2. Hay, S.I.; Guerra, C.A.; Tatem, A.J.; Noor, A.M.; Snow, R.W. The Global Distribution and Population at Risk of Malaria: Past, Present, and Future. *Lancet Infect. Dis.* **2004**, *4*, 327–336. [CrossRef]
3. Guo, W.; Liu, J.; Zhao, X.; Hou, W.; Zhao, Y.; Li, Y.; Sun, W.; Fan, D. Spatiotemporal dynamics of population density in China using nighttime light and geographic weighted regression method. *Int. J. Digit. Earth* **2023**, *16*, 2704–2723. [CrossRef]
4. Lin, C.-H.; Wen, T.-H. Using Geographically Weighted Regression (GWR) to Explore Spatial Varying Relationships of Immature Mosquitoes and Human Densities with the Incidence of Dengue. *Int. J. Environ. Res. Public Health* **2011**, *8*, 2798–2815. [CrossRef]
5. Wang, L.; Wang, S.; Zhou, Y.; Liu, W.; Hou, Y.; Zhu, J.; Wang, F. Mapping Population Density in China between 1990 and 2010 Using Remote Sensing. *Remote Sens. Environ.* **2018**, *210*, 269–281. [CrossRef]
6. Zhu, C.; Zhang, X.; Zhou, M.; He, S.; Gan, M.; Yang, L.; Wang, K. Impacts of Urbanization and Landscape Pattern on Habitat Quality Using OLS and GWR Models in Hangzhou, China. *Ecol. Indic.* **2020**, *117*, 106654. [CrossRef]
7. Zandbergen, P.A. Dasymetric Mapping Using High Resolution Address Point Datasets. *Trans. GIS* **2011**, *15*, 5–27. [CrossRef]
8. Lei, Z.; Xie, Y.; Cheng, P.; Yang, H. From Auxiliary Data to Research Prospects, a Review of Gridded Population Mapping. *Trans. GIS* **2023**, *27*, 3–39. [CrossRef]
9. Qiu, Y.; Zhao, X.; Fan, D.; Li, S.; Zhao, Y. Disaggregating population data for assessing progress of SDGs: Methods and applications. *Int. J. Digit. Earth* **2022**, *15*, 2–29. [CrossRef]
10. Doxsey-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]
11. 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]
12. Clark, C. Urban Population Densities. *J. R. Stat. Society. Ser. A* **1951**, *114*, 490–496. [CrossRef]
13. Tian, Y.; Yue, T.; Zhu, L.; Clinton, N. Modeling Population Density Using Land Cover Data. *Ecol. Model.* **2005**, *189*, 72–88. [CrossRef]
14. Martin, D. Mapping Population Data from Zone Centroid Locations. *Trans. Inst. Br. Geogr.* **1989**, *14*, 90–97. [CrossRef]
15. Martin, D. An Assessment of Surface and Zonal Models of Population. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 973–989. [CrossRef]
16. Martin, D.; Tate, N.J.; Langford, M. Refining Population Surface Models: Experiments with Northern Ireland Census Data. *Trans. GIS* **2000**, *4*, 343–360. [CrossRef]

17. Leyk, S.; Gaughan, A.E.; Adamo, S.B.; de Sherbinin, A.; Balk, D.; Freire, S.; Rose, A.; Stevens, F.R.; Blankespoor, B.; Frye, C.; et al. The Spatial Allocation of Population: A Review of Large-Scale Gridded Population Data Products and Their Fitness for Use. *Earth Syst. Sci. Data* **2019**, *11*, 1385–1409. [CrossRef]
18. 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] [PubMed]
19. Chen, M.; Xian, Y.; Huang, Y.; Zhang, X.; Hu, M.; Guo, S.; Chen, L.; Liang, L. Fine-Scale Population Spatialization Data of China in 2018 Based on Real Location-Based Big Data. *Sci. Data* **2022**, *9*, 624. [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. Stevens, F.R.; Gaughan, A.E.; Nieves, J.J.; King, A.; Sorichetta, A.; Linard, C.; Tatem, A.J. Comparisons of Two Global Built Area Land Cover Datasets in Methods to Disaggregate Human Population in Eleven Countries from the Global South. *Int. J. Digit. Earth* **2020**, *13*, 78–100. [CrossRef]
22. Gaughan, A.E.; Stevens, F.R.; Huang, Z.; Nieves, J.J.; Sorichetta, A.; Lai, S.; Ye, X.; Linard, C.; Hornby, G.M.; Hay, S.I. Spatiotemporal Patterns of Population in Mainland China, 1990 to 2010. *Sci. Data* **2016**, *3*, 160005. [CrossRef] [PubMed]
23. Sorichetta, A.; Hornby, G.M.; Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. High-Resolution Gridded Population Datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci. Data* **2015**, *2*, 150045. [CrossRef]
24. Anselin, L.; Griffith, D.A. Do Spatial Effecfs Really Matter in Regression Analysis? *Pap. Reg. Sci.* **1988**, *65*, 11–34. [CrossRef]
25. Fotheringham, A.S. Trends in Quantitative Methods I: Stressing the Local. *Prog. Hum. Geogr.* **1997**, *21*, 88–96. [CrossRef]
26. Fotheringham, A.S.; Yang, W.; Kang, W. Multiscale Geographically Weighted Regression (MGWR). *Ann. Am. Assoc. Geogr.* **2017**, *107*, 1247–1265. [CrossRef]
27. Yu, H.; Fotheringham, A.S.; Li, Z.; Oshan, T.; Kang, W.; Wolf, L.J. Inference in Multiscale Geographically Weighted Regression. *Geogr. Anal.* **2020**, *52*, 87–106. [CrossRef]
28. Shi, Q.; Zhuo, L.; Tao, H.; Li, Q. Mining Hourly Population Dynamics by Activity Type Based on Decomposition of Sequential Snapshot Data. *Int. J. Digit. Earth* **2022**, *15*, 1395–1416. [CrossRef]
29. Zhang, Y.; Zhang, Y.; Huang, B.; Liu, X. A Hybrid Model for High Spatial and Temporal Resolution Population Distribution Prediction. *Int. J. Digit. Earth* **2022**, *15*, 2268–2295. [CrossRef]
30. Freire, S.; Kemper, T.; Pesaresi, M.; Florczyk, A.; Syrris, V. Combining GHSL and GPW to improve global population mapping. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015.
31. 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]
32. Tiecke, T.G.; Liu, X.; Zhang, A.; Gros, A.; Li, N.; Yetman, G.; Kilic, T.; Murray, S.; Blankespoor, B.; Prydz, E.B.; et al. Mapping the World Population One Building at a Time. *arXiv* **2017**, arXiv:1712.05839.
33. Azar, D.; Graesser, J.; Engstrom, R.; Comenetz, J.; Leddy, R.M., Jr.; 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]
34. Linard, C.; Kabaria, C.W.; Gilbert, M.; Tatem, A.J.; Gaughan, A.E.; Stevens, F.R.; Sorichetta, A.; Noor, A.M.; Snow, R.W. Modelling changing population distributions: An example of the Kenyan Coast, 1979–2009. *Int. J. Digit. Earth* **2017**, *10*, 1017–1029. [CrossRef]
35. Thomson, D.R.; Gaughan, A.E.; Stevens, F.R.; Yetman, G.; Elias, P.; Chen, R. Evaluating the Accuracy of Gridded Population Estimates in Slums: A Case Study in Nigeria and Kenya. *Urban Sci.* **2021**, *5*, 48. [CrossRef]
36. Huang, X.; Wang, C.; Li, Z.; Ning, H. A 100 m Population Grid in the CONUS by Disaggregating Census Data with Open-Source Microsoft Building Footprints. *Big Earth Data* **2021**, *5*, 112–133. [CrossRef]
37. Lwin, K.; Murayama, Y. A GIS Approach to Estimation of Building Population for Micro-Spatial Analysis. *Trans. GIS* **2009**, *13*, 401–414. [CrossRef]
38. Schug, F.; Frantz, D.; van der Linden, S.; Hostert, P. Gridded Population Mapping for Germany Based on Building Density, Height and Type from Earth Observation Data Using Census Disaggregation and Bottom-up Estimates. *PLoS ONE* **2021**, *16*, e0249044. [CrossRef]
39. Shang, S.; Du, S.; Du, S.; Zhu, S. Estimating Building-Scale Population Using Multi-Source Spatial Data. *Cities* **2021**, *111*, 103002. [CrossRef]
40. Ural, S.; Hussain, E.; Shan, J. Building Population Mapping with Aerial Imagery and GIS Data. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 841–852. [CrossRef]
41. Schiavina, M.; Melchiorri, M.; Pesaresi, M.; Politis, P.; Freire, S.; Maffenini, L.; Florio, P.; Ehrlich, D.; Goch, K.; Tommasi, P. *GHSL Data Package 2022*; Publications Office of the European Union: Luxembourg, 2022; ISBN 978-92-76-53071-8.
42. Tripathy, P.; Balakrishnan, K.; de Franchis, C.; Kumar, A. Generating megacity-scale building height maps without DGNSS surveyed GCPs: An open-source approach. *Environ. Plan. B Urban Anal. City Sci.* **2022**, *49*, 2312–2330. [CrossRef]
43. Wu, W.-B.; Yu, Z.-W.; Ma, J.; Zhao, B. Quantifying the influence of 2D and 3D urban morphology on the thermal environment across climatic zones. *Landsc. Urban Plan.* **2022**, *226*, 104499. [CrossRef]
44. Liu, M.; Ma, J.; Zhou, R.; Li, C.L.; Li, D.K.; Hu, Y.M. High-resolution mapping of mainland China's urban floor area. *Landsc. Urban Plan.* **2021**, *214*, 104187. [CrossRef]

45. Wu, W.-B.; Ma, J.; Banzhaf, E.; Meadows, M.E.; Yu, Z.-W.; Guo, F.-X.; Sengupta, D.; Cai, X.-X.; Zhao, B. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens. Environ.* **2023**, *291*, 113578. [CrossRef]
46. Che, Y.; Li, X.; Liu, X.; Wang, Y.; Liao, W.; Zheng, X.; Zhang, X.; Xu, X.; Shi, Q.; Zhu, J.; et al. 3D-GloBFP: The first global three-dimensional building footprint dataset. *Earth Syst. Sci. Data Discuss.* **2024**, *2024*, 1–28.
47. Frantz, D.; Schug, F.; Okujeni, A.; Navacchi, C.; Wagner, W.; van der Linden, S.; Hostert, P. National-Scale Mapping of Building Height Using Sentinel-1 and Sentinel-2 Time Series. *Remote Sens. Environ.* **2021**, *252*, 112128. [CrossRef] [PubMed]
48. Mei, Y.; Gui, Z.; Wu, J.; Peng, D.; Li, R.; Wu, H.; Wei, Z. Population Spatialization with Pixel-Level Attribute Grading by Considering Scale Mismatch Issue in Regression Modeling. *Geo-Spat. Inf. Sci.* **2022**, *25*, 365–382. [CrossRef]
49. Wang, L.; Fan, H.; Wang, Y. Improving Population Mapping Using Luojia 1-01 Nighttime Light Image and Location-Based Social Media Data. *Sci. Total Environ.* **2020**, *730*, 139148. [CrossRef]
50. 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]
51. Breiman, L.; Friedman, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
52. Georganos, S.; Grippa, T.; Niang Gadiaga, A.; Linard, C.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E.; Kalogirou, S. Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling. *Geocarto Int.* **2021**, *36*, 121–136. [CrossRef]
53. Grekousis, G.; Feng, Z.; Marakakis, I.; Lu, Y.; Wang, R. Ranking the Importance of Demographic, Socioeconomic, and Underlying Health Factors on US COVID-19 Deaths: A Geographical Random Forest Approach. *Health Place* **2022**, *74*, 102744. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.