remote sensing

MDPI

Article

A Population Spatialization Model at the Building Scale Using Random Forest

Mengqi Wang [1], Yinglin Wang [1], Bozhao Li [1], Zhongliang Cai [1,*] and Mengjun Kang [1,2]

[1] School of Resource and Environmental Sciences, Wuhan University, No. 129 Luoyu Rd., Wuhan 310029, China; mqwang@whu.edu.cn (M.W.); 2018282050148@whu.edu.cn (Y.W.); libozhao@whu.edu.cn (B.L.); mengjunk@whu.edu.cn (M.K.)
[2] Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100045, China
* Correspondence: zlcai@whu.edu.cn

Abstract: Population spatialization reveals the distribution and quantity of the population in geographic space with gridded population maps. Fine-scale population spatialization is essential for urbanization and disaster prevention. Previous approaches have used remotely sensed imagery to disaggregate census data, but this approach has limitations. For example, large-scale population censuses cannot be conducted in underdeveloped countries or regions, and remote sensing data lack semantic information indicating the different human activities occurring in a precise geographic location. Geospatial big data and machine learning provide new fine-scale population distribution mapping methods. In this paper, 30 features are extracted using easily accessible multisource geographic data. Then, a building-scale population estimation model is trained by a random forest (RF) regression algorithm. The results show that 91% of the buildings in Lin'an District have absolute error values of less than six compared with the actual population data. In a comparison with a multiple linear (ML) regression model, the mean absolute errors of the RF and ML models are 2.52 and 3.21, respectively, the root mean squared errors are 8.2 and 9.8, and the R^2 values are 0.44 and 0.18. The RF model performs better at building-scale population estimation using easily accessible multisource geographic data. Future work will improve the model accuracy in densely populated areas.

Keywords: population spatialization; random forest model; building scale

Citation: Wang, M.; Wang, Y.; Li, B.; Cai, Z.; Kang, M. A Population Spatialization Model at the Building Scale Using Random Forest. *Remote Sens.* **2022**, *14*, 1811. https://doi.org/10.3390/rs14081811

Academic Editors: Hua Liu and Victor Mesev

Received: 25 February 2022
Accepted: 31 March 2022
Published: 8 April 2022

1. Introduction

Population spatialization data reflect the population distribution in the objective world, and fine-scale population distribution data are essential in public health and urban planning [1–7] and facilitate mobile population monitoring, resource allocation optimization, and urban structure analysis [3,8–11]. Most countries obtain detailed population distribution maps through censuses, which have limited utility and suffer from the following problems: (1) High costs prohibit underdeveloped countries and regions from conducting large-scale population censuses [12,13]. (2) Long census intervals, changing administrative boundaries, and uneven distributions of the population within administrative units do not allow the censusing approach to reflect population changes promptly at a fine spatial resolution [12]. Since census data cannot reveal the spatial heterogeneity of population density in detail [6,14], the use of general multisource GIS datasets for decomposing census data to map population distribution at fine scales has become a research hotspot [2–4,13,15].

Research on the spatial decomposition of census data into grid cells began in the 1990s [16,17]. Early studies focused on spatial interpolation problems of populations, such as pycnophylactic interpolation [18], areal interpolation [19], intelligent interpolation [20], and dasymetric mapping [21–26]. These methods have produced many meaningful global datasets from typical GIS datasets, such as the Gridded Population of the World (GPW) with a resolution of 2.5 arc-min [27,28], the Global Rural Urban Mapping Project (GRUMP)

with a resolution of 30 arc-s [29,30], Landscan [6,31], WorldPop [32] and the Global Human Settlements Population Grid dataset [33]. However, these global-scale population datasets can only be used to study population change at the macro scale. Remote sensing products, satellite image products such as land cover/land use types, and nighttime light (NTL) images are widely used as auxiliary data to map population distribution at fine scales [5,13,15,22,23,34–36].

However, remote sensing data with medium spatial resolution lack semantic information and cannot directly indicate land use or human presence [37]. These data have a limited ability to extract demographic and socioeconomic characteristics associated with human activities in complex urban environments [38–43]. Geospatial big data such as point of interest (POI) data can compensate for these drawbacks of remotely sensed imagery. POI data contain location information and textual descriptions that extract detailed information about cities or social systems [44,45]. Many studies [35,38,40,46,47] have used POIs to define functional urban areas and land use types. Moreover, the results have shown a correlation between POI categories and population density [4,40]. Bakillah used voluntary geographic information (VGI) to map the population distribution at the building level in Hamburg; this approach relies only on POI and fine-grained land use/land cover data and does not consider the spatial heterogeneity of the population distribution when calculating the population within buildings [4]. Different geospatial big data can capture different aspects of the ground truth [39]. Additionally, some studies have combined remote sensing products with residential building footprints and census data to build empirically weighted models to map building-scale population distributions [15,48]. These methods are still challenging to apply to fine mapping of the population in China, where the urban spatial structure is diverse and the population distribution is complex [35]. Therefore, this study integrates multiple sources of easily accessible geospatial data to construct a population prediction model, revealing the population distribution at the building scale.

Machine learning, which has advanced considerably in the past few decades, has provided more efficient tools for population spatialization, and the random forest (RF) algorithm is one of the most common and powerful supervised learning algorithms. RF is a classification tree-based machine learning algorithm first proposed in 2001 by Leo Breiman and Cutlery Adele [49]. It has many appealing properties, such as high classification accuracy; the ability to model complex interactions among predictor variables; and the flexibility to perform several types of statistical data analysis, including regression, classification, survival analysis, and unsupervised learning [50–52]. Stevens et al. used the RF algorithm to build a nonparametric predictive model to map the fine-scale population distribution in Kenya, Vietnam, and Cambodia at a reduced scale of census data [13]. Methods using RF have been successfully applied to map the population density of China at a 100 m resolution [53]. Yao et al. used the RF algorithm to analyze POIs and real-time user density (RTUD) to reduce the street-level population distribution to the grid level [35]. Ye et al. combined POIs with multisource remote sensing data in an RF model to produce a 100 m spatial resolution gridded population map of China with a higher accuracy than the WorldPop dataset [37]. The results of the above studies all show that the RF algorithm is reliable and has good performance in fine-scale population spatialization; therefore, the RF algorithm was chosen to construct the prediction model in this paper.

Although many researchers have fused remote sensing images with geospatial data such as POI data and used machine learning algorithms to map population distributions at a fine scale, the existing methods still have drawbacks. Countries and regions with poor economic conditions cannot conduct large-scale population censuses, and some high-resolution remote sensing images are confidential data. This study addresses these drawbacks by fusing multiple sources of easily accessible geospatial data, such as building data, land use data, NTL data, administrative district data, road data, water system data, and POI data, using the RF algorithm to train a model to predict the population distribution at the building scale. Section 2 includes an overview of the study area and a description of the multiple sources of easily accessible geospatial data. Section 3 focuses on the methodology;

it introduces how to build a multidimensional feature library using seven easily accessible sources of geographic data, filter 30 features related to population distribution by feature engineering, and train the prediction model using the RF algorithm. In Section 4, the prediction results are mainly presented and compared with a multiple linear (ML) regression model trained with the same dataset. The model performance is analyzed by evaluating metrics such as absolute error compared to the actual value, mean absolute error (MAE), root squared mean error (RSME), and R^2. Section 5 presents the feature importance and contribution analysis, and the results show that building area and self-service POI are the two most essential features in this model. Finally, Section 6 summarizes the paper and discusses the implications for future research in this area.

2. Data and Preprocessing

The study area in this paper is Lin'an District, Hangzhou city, Zhejiang Province, located from 118°21′ to 120°30′E longitude and 29°11′ to 30°33′N latitude, as shown in Figure 1. Lin'an District contains 18 subareas. In addition, CGCS2000 and the 3-degree Gauss–Kruger zone 40 are used in the experiment.

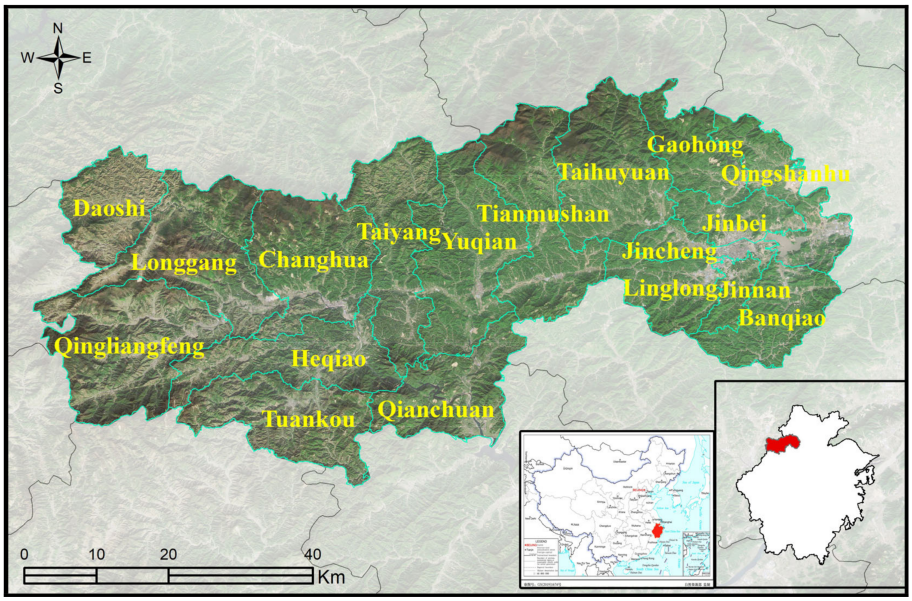
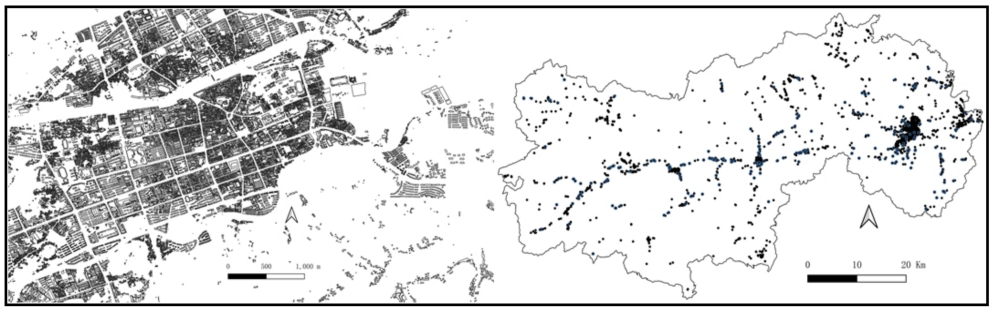


Figure 1. Geographic location of the study area.

The following datasets are used in this study, Table 1 lists the format and source of these data, Figures 2 and 3 show the visualization of datasets:

Table 1. Datasets used in the study area.

Dataset	Format	Source
Population (2017)	Table	Hangzhou Public Security Bureau
Buildings (2017)	Polygon vector features	Basic geographic information database for Hangzhou, China
Finer Resolution Observation and Monitoring of Global Land Cover (2017)	Grid, 30 m spatial resolution	Tsinghua University Open Data Set (http://data.ess.tsinghua.edu.cn/, accessed on 1 December 2021)
DMSP-OLS NTL imagery (2013)	Grid, 1 × 1 km spatial resolution	National Geophysical Data Center, USA (https://ngdc.noaa.gov/eog/, accessed on 1 December 2021)
Road network (2017)	Line vector features	Basic geographic information database for Hangzhou, China
POIs (2019)	Point features	Baidu Map API, China
Water system	Polygon vector features	Basic geographic information database for Hangzhou, China
Administrative districts (2017)	Polygon vector features	Basic geographic information database for Hangzhou, China



(**a**) (**b**)

Figure 2. Subsets of (**a**) building and (**b**) POI data.

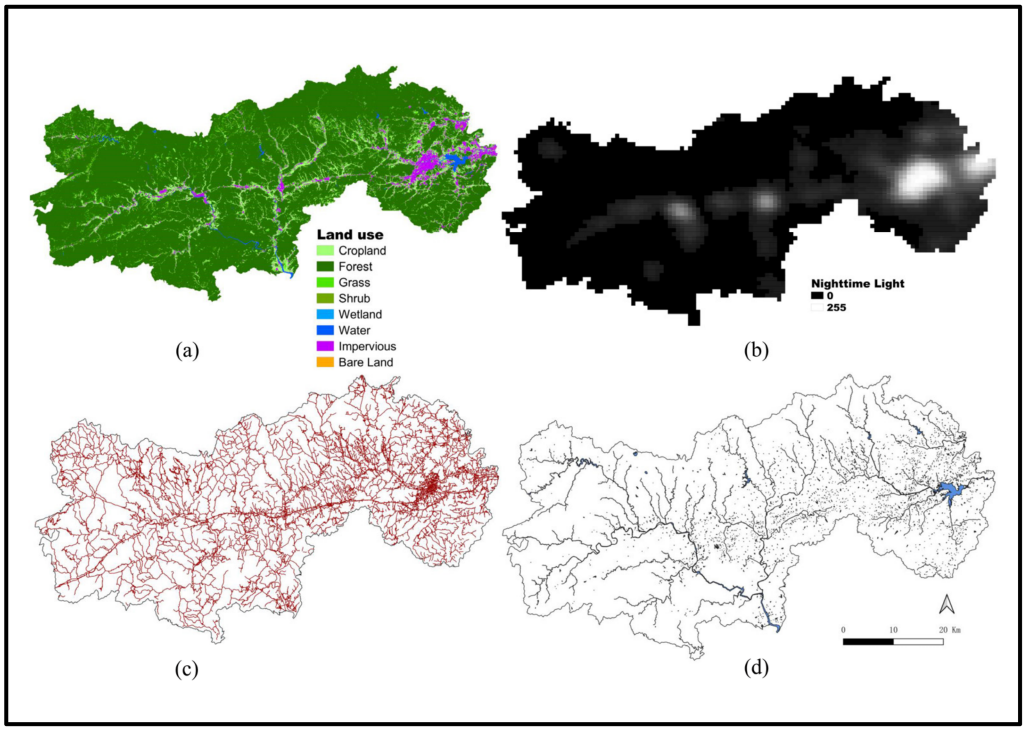


Figure 3. (**a**) Land use data, (**b**) NTL data, (**c**) road system data, and (**d**) water system data.

(1) Population data. The actual population data are used as model validation data in this experiment, and the study area contains 300,722 population records.
(2) Building data. Buildings are the basic units for the experiment, and the dataset contains 117,116 residential buildings.
(3) Land use dataset. The Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) map set is used as the auxiliary data. The data resolution is approximately 30 m in the maps. There are 8 categories included in the dataset: agricultural land, forest, grassland, shrubland, wetland, water, impervious surfaces, and bare ground.
(4) DMSP-OLS NTL imagery. The fourth version of the DMSP-OLS (Defense Meteorological Satellite Program) NTL remote sensing dataset synthesized in 2013 is used as auxiliary data for population spatialization. The resolution is approximately 1 km, and the data were resampled to 100 m.
(5) Water systems, road networks, and POIs also affect the population distribution to a certain extent. The study area includes 85,876 rivers, 27,706 roads, and 4524 POI records. The detailed information for each POI is shown in Table 2. We calculated the closest Euclidean distance from each building type to the same type of POI and used the results as model inputs.

Table 2. Types and counts of POIs.

No.	POI Type	Count	No.	POI Type	Count
1	Medical	295	9	Nursing homes	9
2	Sports	62	10	Self-service	35
3	Education	439	11	Recreation	721
4	Parks	22	12	Government agencies	396
5	Markets	186	13	Shopping	1416
6	Gas stations	56	14	Factories	246
7	Museums	5	15	Banks	221
8	Retail	123	16	Corporations	292

3. Methods

The RF concept is based on a bagging algorithm, and a method involving the random selection of independent variables is used in the training process of decision trees [49]. An RF contains multiple decision trees trained with a bagging-based integrated learning technique [54]. When building an individual classification tree, for each splitting point in the tree, a random sample containing q ($1 \leq q \leq p$) independent variables is selected as a candidate from among p total independent variables, and the independent variables associated with a splitting point can only be selected from q variables. The similarity between individual classification trees with highly correlated output prediction results can be reduced by considering only a subset of the independent variables at each splitting point. The independent variables that have the most significant impact on performance do not influence the $(p - q)/p$ ratio because they are not selected as splitting points, and other independent variables have an equal chance of being selected as split points, thus decorrelating the effects of single trees [55].

First, a multidimensional feature library is constructed. Second, feature engineering steps, such as filtering and standardization, are performed for the feature library, and features related to the spatial distribution of the population are selected as explanatory variables and included in model construction. Then, based on the RF regression algorithm, population spatialization is performed, and the grid search method and cross-validation are applied to adjust and optimize the model to improve performance. Finally, the population spatialization results at the building scale in the study area are obtained and compared with the results of an ML model to evaluate the accuracy and performance of each model. The flowchart in Figure 4 shows the entire workflow.

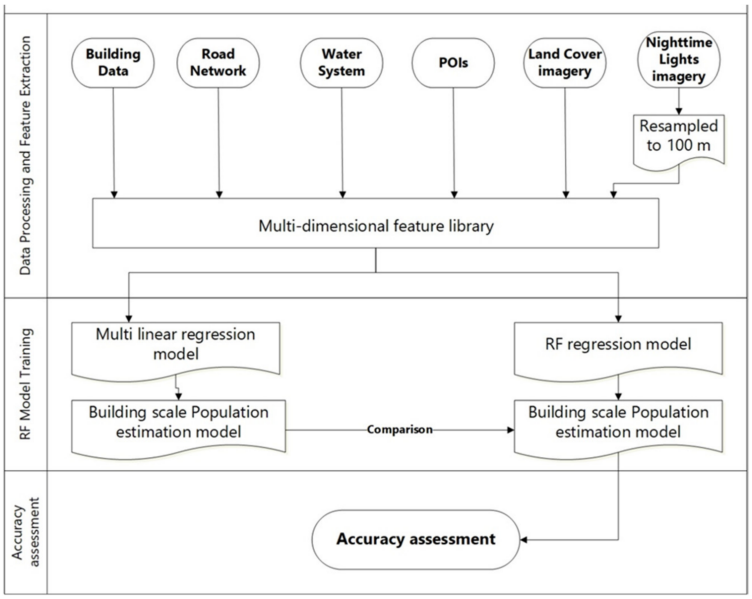


Figure 4. Flowchart of RF model construction and accuracy assessment.

3.1. Feature Engineering

3.1.1. Feature Filtering

Population spatialization requires the use of feature values related to the population distribution as independent variables for training. For an established feature set, manual screening is performed, and the influence on the spatial distribution of the population is used as the feature selection criterion. Text-based feature values, such as the name of the administrative district in which a building is located, the address of the building, the population of each sex, the names of roads, and the name of the water system, are removed due to the algorithmic factors used in model fitting. Furthermore, numerical-type feature values are used as the feature data. Among these extracted feature values, numerical-type feature values, such as those for road length and the number of floors in a building, are removed due to issues with missing values.

Feature selection is performed to facilitate the evaluation of features, mainly with a screening method; specifically, the importance of each dimensional feature is assessed according to an evaluation index, and then the features are ranked and selected based on their importance scores. To construct the model, if the variance of a dimensional feature is minimal, the feature provides limited information, the variability in the feature is minimal, the contribution to the model is limited, and the impact on model performance is negligible; therefore, these types of features are removed with a low-variance filtering feature selection method. In this method, the variance of each feature in the sample is determined, and the features are ranked according to their magnitude of variance. By default, features with a variance of 0 are removed, i.e., the sample features do not change. In this paper, the threshold value is 0.8, and the number of features is not set. The final result is that all features with variance values between 0.9 and 1 are more significant than the set threshold value, so no features are removed.

In summary, 30 features were selected from the multidimensional feature library to form the auxiliary dataset for model training and prediction. The specific features are shown in Table 3.

Table 3. List of filtered features.

No.	Feature Name	Feature Source	No.	Feature Name	Feature Source
1	Building footprint	Building	16	Factory_EDIST	POI
2	Night lighting_Min	Night lighting	17	Company_EDIST	POI
3	Night lighting_Max	Night lighting	18	Park_EDIST	POI
4	Night lighting_Ave	Night lighting	19	Store_EDIST	POI
5	Night lighting_Sum	Night lighting	20	Gas station_EDIST	POI
6	Land use type	Land Use	21	Education agency_EDIST	POI
7	River system_Cnt	River system	22	Retail_EDIST	POI
8	River system length_Min	River system	23	Market_EDIST	POI
9	River system length_Max	River system	24	Sports facility_EDIST	POI
10	River system length_Sum	River system	25	Entertainment_EDIST	POI
11	Water area_Min	River system	26	Nursing home_EDIST	POI
12	Water area_Max	River system	27	Medical institution_EDIST	POI
13	Water area_Sum	River system	28	Bank_EDIST	POI
14	Road_EDIST	Road	29	Government agency_EDIST	POI
15	Museum_EDIST	POI	30	Self-service_EDIST	POI

Min in the feature name indicates the minimum value of the feature in the building range, Max indicates the maximum value, Avg indicates the average value, Sum indicates the total value, Cnt indicates the count, and EDIST indicates the calculation of the closest Euclidean distance from each building data to the same type of data.

3.1.2. Feature Standardization

After filtering of the features, the value data are normalized, and the dataset is scaled to the interval of [0, 1] to eliminate the possible problems caused by the unit differences and different magnitudes among the multidimensional feature values. The standardization process is divided into two main steps: decentering the mean (setting the mean to 0) and scaling the variance (setting the variance to 1). To fairly assess the role of eigenvalues

in population spatialization, the eigenvalues are standardized before building the model according to the following equation.

$$X = \frac{x - x_{mean}}{\sqrt{\frac{(x-x_{mean})^2+(x+x_{mean})^2}{n}}} \quad (1)$$

where x_{mean} denotes the average value of the data and n represents the number of data points, which in the formula indicates the amount of data associated with one eigenvalue.

3.2. Model Building and Training

A total of 31 feature values in six categories, including building features, NTL features, land use type features, water system features, road features, and POI features, are used as the independent variable dataset, and the number of people in buildings is used as the dependent variable. Python and Scikit-learn [56], a third-party open-source machine-learning algorithm package, are used as the basis for programming, and an RF regression algorithm is used to construct a population spatialization model and perform model training.

In constructing the model, a sampling method with replacement is used. Of the original samples, 85% are used as the training dataset, and the remaining 15% constitute the validation dataset, which is divided by a fixed random number to ensure the randomness of the dataset division and the reproducibility of the experiment for the adjustment and optimization of the model parameters.

When constructing the model, it is necessary to determine the optimal hyperparameters, and in this experiment, a grid search approach was chosen as the parameter optimization method. The grid search method has a relatively slow run time, but after cross-validation, the results are highly reliable. Based on the optimal parameter values returned from the grid search method, the optimal values of each parameter applied in the RF regression algorithm to construct the population spatialization model are shown in Table 4.

Table 4. Optimal values of model parameters.

No.	Parameter Value	Value Range	Optimal Value
1	*bootstrap*	True, False	True
2	*oob_score*	True, False	True
3	*n_estimators*	100, 200, . . . , 1500	1100
4	*max_features*	auto, sqrt, log2	auto
5	*max_depth*	1, 2, . . . , 20	16
6	*min_samples_leaf*	1, 2, . . . , 20	19
7	*min_samples_split*	2, 4, . . . , 20	18

In this RF model, the first three parameters are the RF framework parameters, among which the *bootstrap* parameter indicates whether bootstrap samples are used when building trees. *n_estimators* is the number of trees in the forest, which impacts model performance; when this parameter is too small, underfitting will occur, the run time will be long, and the modeling efficiency will be reduced. *oob_score* indicates whether out-of-bag samples are used to estimate the generalization score and can be used to evaluate the model's strengths and weaknesses, validate the model, reduce time consumption, and improve the modeling accuracy.

The last four parameters are decision tree parameters, which mainly control the growth process of a single decision tree in the RF. *max_features* is the maximum number of features to consider when constructing a decision tree. *max_depth* is the maximum depth of the decision trees. *min_samples_split* controls subtree splitting; when the number of samples at the middle node is lower than the selected parameter value, the tree stops growing, i.e., no more features are selected for division. *min_samples_leaf* controls the tree depth of decision

trees; when the number of samples at a leaf node is lower than a given threshold, the tree is pruned.

When model training and parameter optimization are completed, the population spatialization model is constructed according to the optimal values of the determined model parameters.

4. Results and Evaluation

4.1. RF Population Spatialization Results

The input dataset included 117,116 buildings with 30 features, and after RF model training, the regression model predicted the population in each building. According to the natural breaks grading method, the predicted result was divided into five levels, as shown in Figure 5.

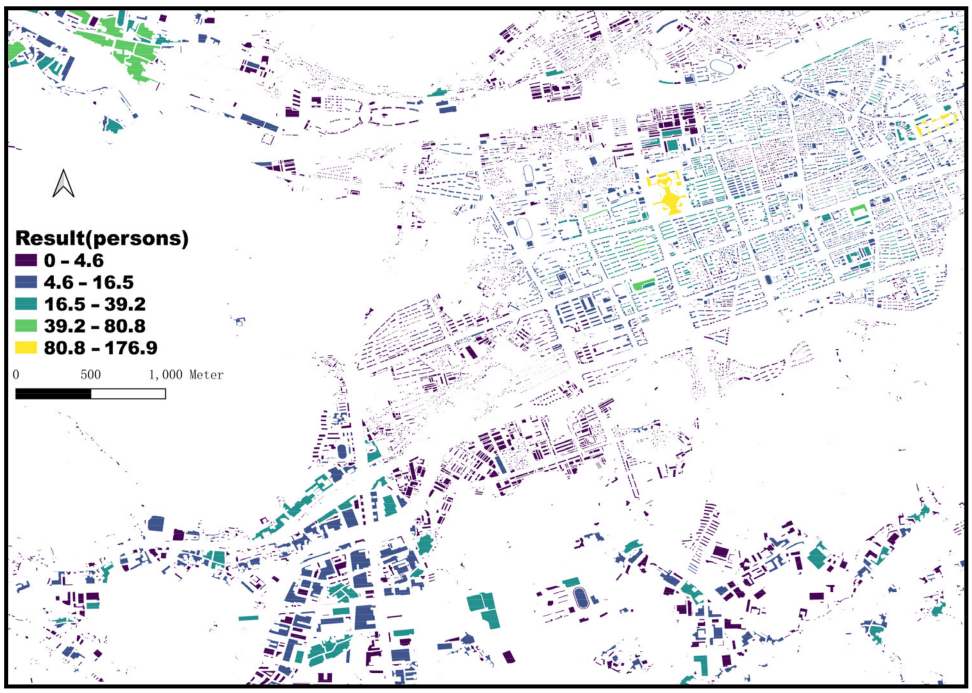
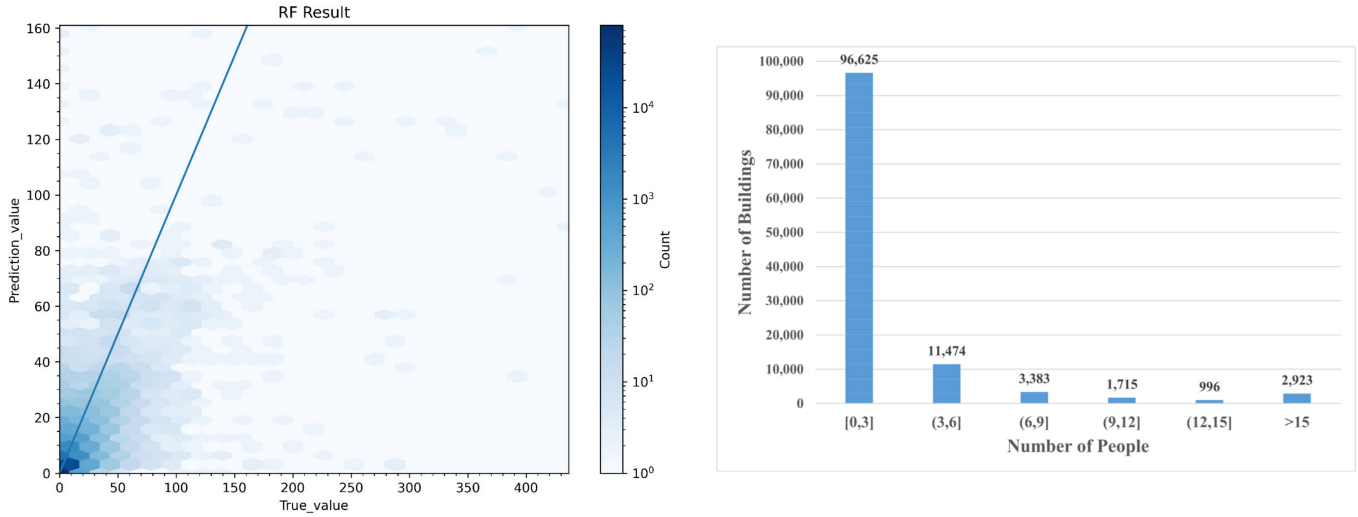


Figure 5. Results of the RF model.

Figure 6a shows the hexagonal bin plot and histogram of the prediction results. In this plot, the horizontal coordinate indicates the predicted population, the vertical coordinate indicates the actual population, and the color indicates the degree of overlap of the scattered points at the location; the darker the color, the higher the degree. Moreover, the histogram reflects the distribution of the number of people at the building scale.

Figure 6b shows that the number of buildings with a population in the interval of (0, 3] is 96,625, accounting for 83% of all buildings. The number of buildings with a population in the interval of (3, 6] is 11,474, accounting for 9% of all buildings, and the number of buildings with a population above 15 is 2923, accounting for 2% of all buildings. Some scattered points along the vertical axis indicate that the predicted population is higher than the actual population, and the prediction deviation is within 30 people. Outside the interval of [0, 30), the predicted population is significantly lower than the population denoted by the diagonal line, which suggests that the model seriously underestimates the actual population; the prediction deviation is within the range of [10, 70), which indicates that the RF regression algorithm produces a significant deviation when predicting populations for buildings with high population aggregation.



(**a**) (**b**)

Figure 6. (**a**) Hexagonal bin plot and (**b**) histogram of the RF model results.

The absolute error is calculated to evaluate the performance of the RF model, which is the difference between the predicted and true values:

$$y_{error} = y_{pred} - y_{true} \quad (2)$$

This variable reflects the absolute error in the population estimate for each building and can be used to analyze the sources of error based on local conditions. The error in the population of each building is divided into five levels and visualized by color differences, as shown in Figure 7.

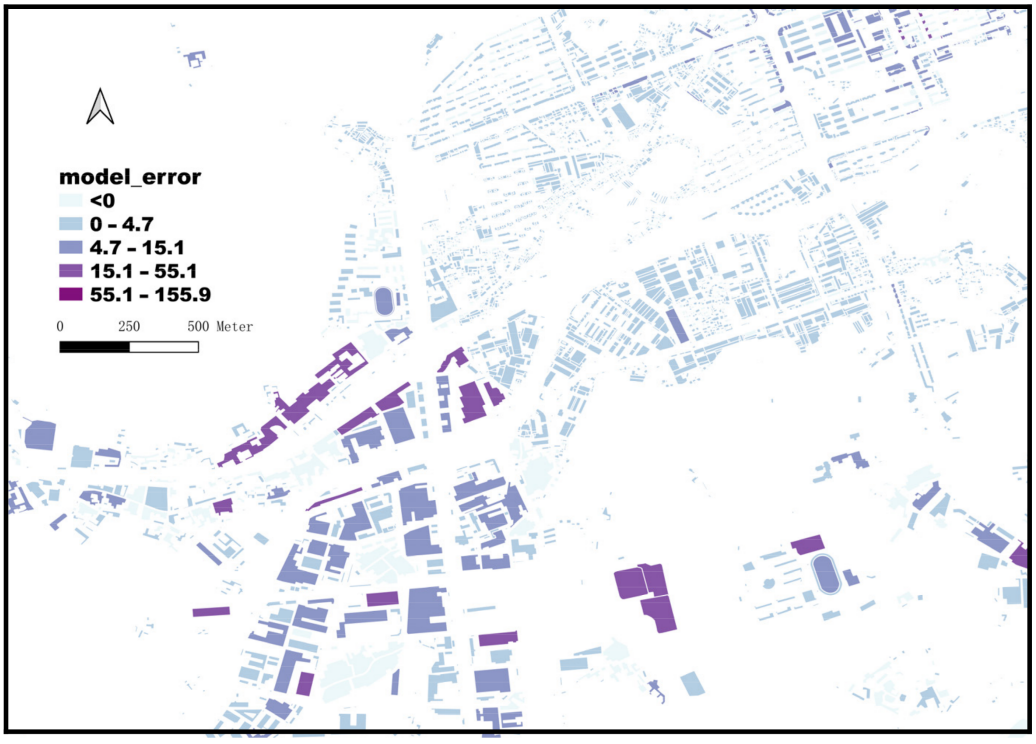


Figure 7. The error of the RF model.

The number of buildings with prediction deviations in the interval of [0, 6) is 106,643, accounting for 91% of all buildings. The number of buildings with errors within the interval of [6, 20) is 8592, accounting for 7% of all buildings, and the number of buildings with errors above 20 people is 1881, accounting for 2% of all buildings. The buildings with large deviations from the predictions are concentrated in the eastern part of the study area, the core area of population aggregation, indicating that the model yields poor predictions in densely populated areas.

Moreover, three indicators are chosen to evaluate the model and predicted values: the mean absolute error (MAE) is the average of the absolute error, which is used to evaluate the predicted data and actual data directly; the root mean squared error (RMSE) is the square root of the mean of the squared difference between the predicted value and the actual observation, which is often used as a measure of the prediction results of machine learning models; and the goodness-of-fit (R^2) is used to evaluate the fit of the model to the observed data, with a range of [0, 1].

The MAE indicator for the RF model is 2.52, indicating that the model is relatively accurate in predicting the population at the building scale in the study area. The RMSE indicator is 8.2, indicating that the variance of the population at the building scale is explained by the characteristics extracted from the multisource data. The R^2 indicator is 0.44, indicating that the model accurately fits 44% of the population data at the building scale in the study area.

4.2. Comparison with an ML Regression Model

To investigate the advantages of the RF population spatialization model at the building scale, we constructed an ML regression model [57] with the same feature set. The regularized Lasso method [58] was chosen to construct the ML regression model. Then, we compared and analyzed the population prediction results obtained with the two algorithms. Table 5 shows the variable information and corresponding coefficients used in the model.

Table 5. Coefficients of variables.

Variable Name	Coefficient	Variable Name	Coefficient
Building footprint	3.45154761	Factory_ EDIST	−0.09081666
Night lighting_Min	0	Company_ EDIST	0.28783718
Night lighting_Max	−0.32496302	Park_ EDIST	0.17310836
Night lighting_Ave	0	Store_ EDIST	−0.58551197
Night lighting_Sum	−0.06787263	Gas_ EDIST	−0.29694836
Land Use Type	0.1270625	Education_ EDIST	0.20848317
River system_Cnt	−0.02700125	Retail_ EDIST	0.78328459
River systemlength_Min	0.16721964	Market_ EDIST	−0.7959491
River systemlength_Max	0.3384865	Sports_ EDIST	0.47081309
River systemlength_Sum	−0.51151035	Leisure_ EDIST	−0.09139956
Water area_Min	0	Nuring_ EDIST	0.29575162
Water area_Max	0.39497887	Medical_ EDIST	−0.39804979
Water area_Sum	0	Bank_ EDIST	0.26058286
Road_EDIST	−0.30963247	Government_ EDIST	−0.02576417
Museum_ EDIST	−0.21837239	Self-service_ EDIST	−0.98357816

Based on these parameters, the prediction results of the multiple linear regression model are shown in Figure 8.

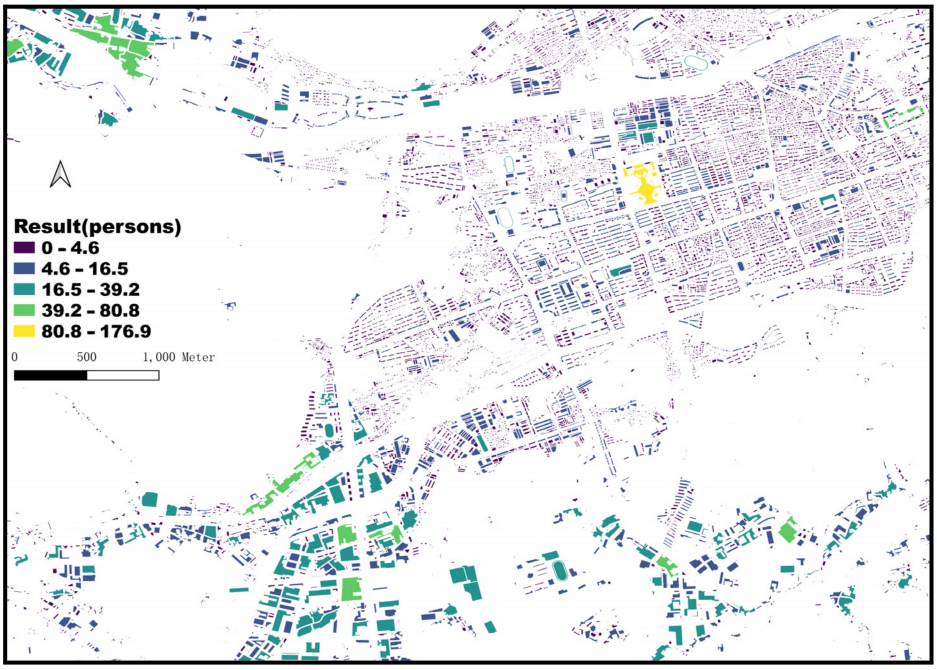


Figure 8. Results of the ML model.

The hexagonal bin plot of the multiple linear regression results is shown in Figure 9. As seen from the figure, most of the points are concentrated in the interval of [0, 40) and limited to [0, 20). Some of the data in the interval are distributed along the vertical axis, indicating that the model overestimates the actual population, and the prediction deviates from the observations by 20–30 people. The scatter points outside this interval are significantly lower than the actual population, and the deviation is in the range of [30, 80), indicating that the multiple linear regression algorithm underestimates the population; however, the degree of deviation is slightly higher than that of the RF regression algorithm.

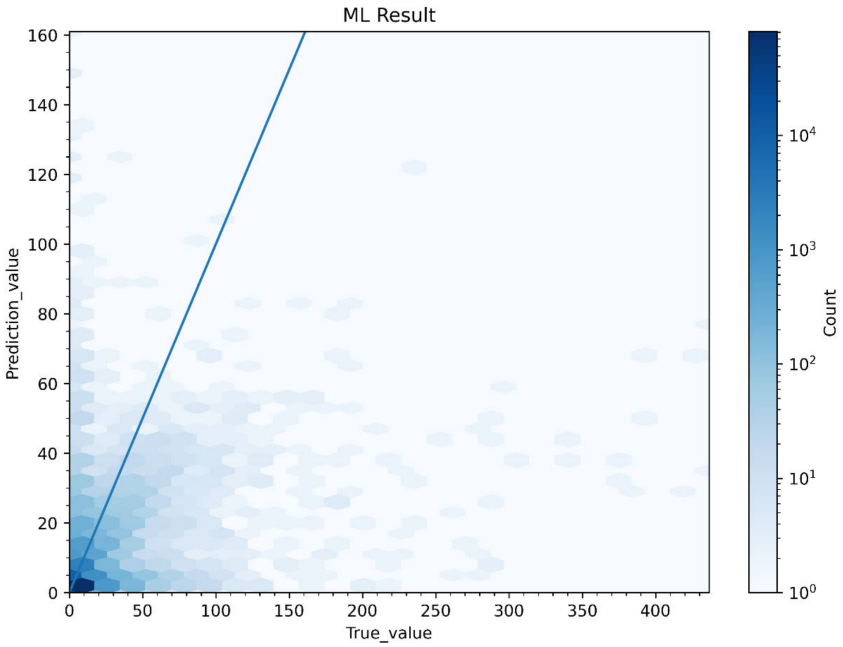


Figure 9. Hexagonal bin plot of the ML model.

Figure 10 shows the error for the ML regression model. In this regression model, the number of buildings with prediction deviations in the [0, 6) interval is 98,832, accounting

for 84% of all buildings. Additionally, the number of buildings with errors in the [6, 20) interval is 12,544, accounting for 11% of all buildings, and the number of buildings with errors of 20 or more people is 5740, accounting for 5% of all buildings. From the perspective of relative error, the prediction deviation for the multiple linear regression model in the [0, 6) interval is similar to that of the RF regression model (Figure 11). However, the number of buildings with a prediction deviation of more than 20 people is significantly higher than that of the RF model. In the prediction results of the multiple linear regression model, the buildings with large deviations were concentrated in the northeastern part of the study area, and the RF algorithm does not simulate a similar result.

Figure 10. Error in the ML model.

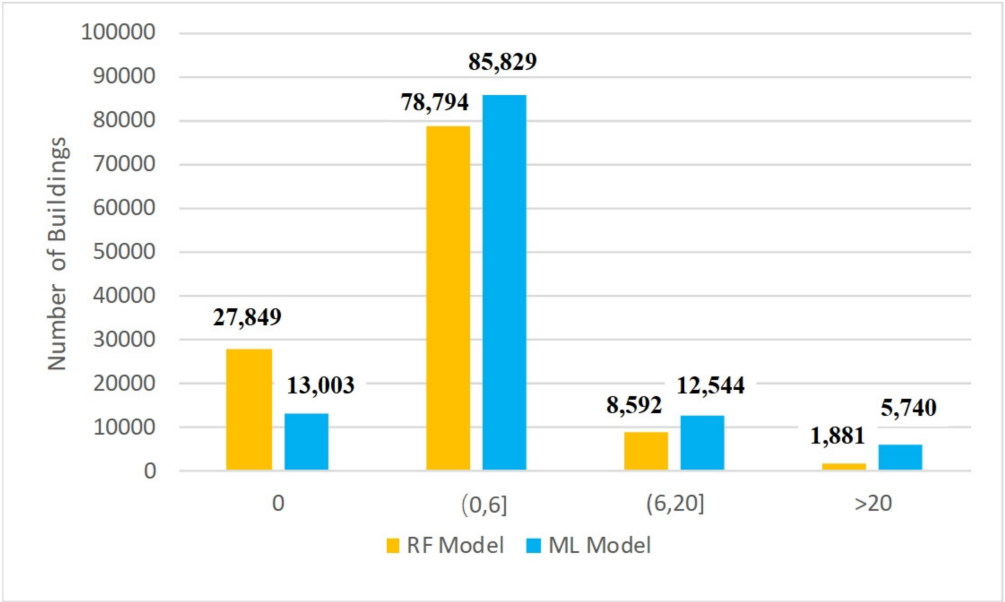


Figure 11. Histogram of error comparison between the RF model and ML model.

The image above shows the error comparison between the RF model and ML model. The vertical coordinate indicates the number of buildings, while the horizontal coordinates indicate the different intervals of absolute error values. The number of buildings with zero error in the RF model is more than twice the number in the ML model. In addition, based on the selected standard model metrics, the MAE for the ML model was 3.21, the RMSE was 9.8, and the R^2 was 0.18. These results were compared with the RF regression model, as shown in Table 6.

Table 6. Comparison of model metrics.

Model Name	MAE	RMSE	R^2
Random forest	2.52	8.2	0.44
Multiple linear regression	3.21	9.8	0.18

The result indicates that the ML regression model performs poorly in terms of data prediction accuracy and model goodness of fit compared to the RF regression model from metric evaluation. In summary, compared with the ML regression algorithm, the population spatialization model of population data constructed based on the RF regression algorithm yields a better fitting result, smaller prediction bias values, and better relatively accurate results for population prediction in densely populated areas, thus reflecting a better overall model performance.

5. Discussion

We quantified the involvement of features in the model construction process from two perspectives: feature importance and feature contribution. Then, the relationship between features and the model was established to assess the model.

5.1. Feature Importance Analysis

The importance of each feature in the constructed population spatialization model was evaluated based on the mean decrease impurity (MDI) or Gini index, a commonly used evaluation index [59]. Suppose there are m features $X_1, X_2, \ldots, X_m$. The Gini coefficient score VIM_j for the jth feature X_j at the nodes of each decision tree in the RF is calculated as

$$GI = 1 - \sum_{k=1}^{|k|} \rho_{mk}{}^2 \quad (3)$$

where GI is the Gini coefficient value, k is the feature category, m is the node, and ρ_{mk} is the proportional contribution of category k at node m. Thus, the amount of change in the Gini index generated by the branching of feature X_j at node m is

$$VIM_{jm} = GI_m - GI_l - GI_r \quad (4)$$

where VIM is the feature importance score and GI_l and GI_r represent the amount of change in the Gini coefficient after feature j branches at node m, respectively. Assuming that the nodes for which feature X_j influences the construction of decision tree i are in set M, the importance of feature X_j to decision tree i is

$$VIM_{jm} = \sum_{m \in M} VIM_{jm} \quad (5)$$

On this basis, assuming that there are n trees in the RF, the importance of the influence of feature X_j on all decision trees in the RF construction process is

$$GVIM_j = \sum_{i=1}^{n} VIM_{ij} \quad (6)$$

After the importance scores of all the features are calculated, the scores are normalized to obtain the final score for each feature, as shown below for the example of feature *Xj*:

$$\bar{VIM}_J = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad (7)$$

According to the MDI method, the importance of the constructed RF model was calculated, and the results are shown in Figure 12. Notably, the building footprint has the highest importance score and has the most significant influence on the features; the next-most-important features are self-service features, sports facilities features, and road features. The three feature types that have the lowest influence on the model are the total water area features, total system length features, and land use type features.

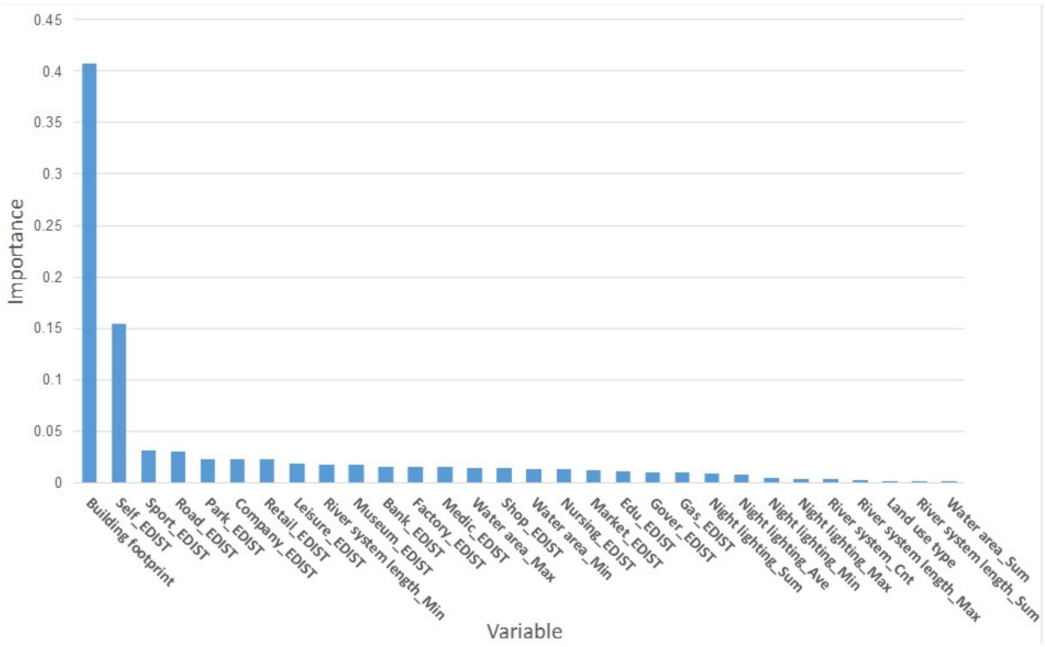


Figure 12. Feature importance in the RF model.

5.2. Feature Contribution Analysis

This experiment is based on the Boruta algorithm in the R language [60] and is performed to assess the degree of feature contributions. The features used to construct the RF model are the explanatory variables, and the actual population of buildings in the study area is the explanatory variable. The Boruta algorithm is used to evaluate the correlation of features with the dependent variable, and features with high correlations with the dependent variable are filtered and removed; rather than focusing solely on the impact of features on model performance improvement, this approach enhances the understanding of feature contributions [60].

After the algorithm is used to quantify the degree of feature contributions, the correlations between the explanatory variables and the explained variables are assessed by comparing the median feature Z score (Z score) with the median Z score of the best attributes. The obtained results are visualized in graphical form, as shown in Figure 13.

In the figure, green denotes the correlation between the corresponding feature and the population distribution, and the features with high correlation rankings include the building footprint, POIs, and land use, among others. The features with the slightest influence are the total water area, the total water system length, and the maximum water system length.

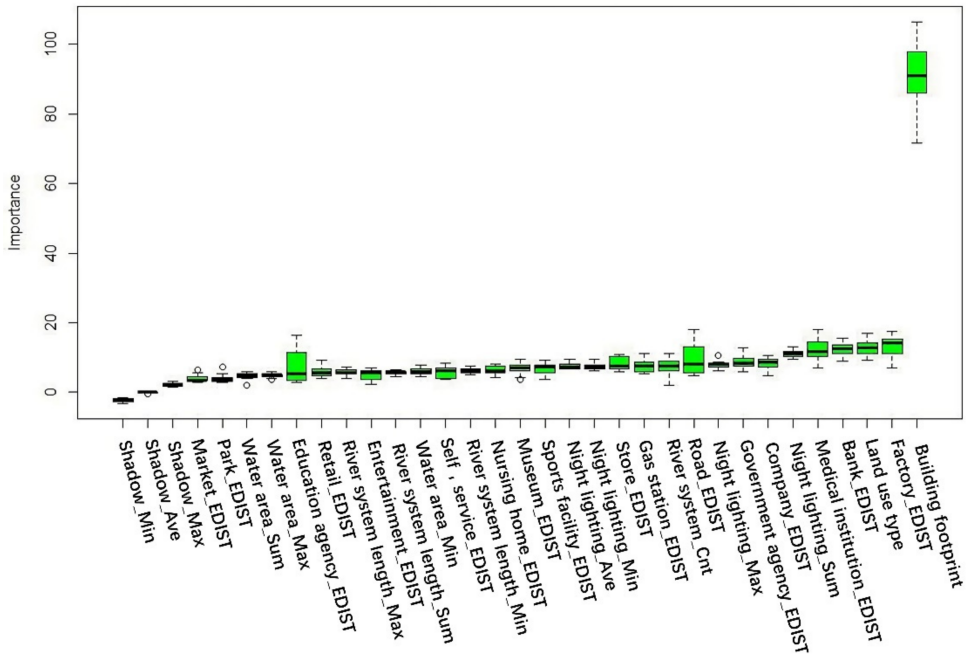


Figure 13. Evaluation of feature contributions.

6. Conclusions

Most previous studies on mapping population distributions have fused multisource remote sensing data and big geospatial data such as POI data to decompose census data. These methods have limitations for economically underdeveloped countries and regions, which do not have access to confidential remote sensing data and cannot conduct large-scale population censuses. This study addresses this problem by fusing multiple sources of easily accessible geospatial data and successfully mapping the building-level population distribution in Lin'an using the RF algorithm. First, 30 features related to population activities are extracted from the easily accessible multisource geographic dataset and trained using the RF algorithm, and the results are compared with actual values and ML models. The RF model has better performance than the ML model in terms of the absolute error, MAE, RMSE, and R^2. Furthermore, the results of feature importance and feature contribution analysis show that only a few features contribute significantly to the model prediction results, and other features play a fine-tuning role in the prediction results, among which the building footprint is the most important feature and is highly correlated with the population distribution.

This study also has some shortcomings. The multisource geographic data sources used in the experiments were released at inconsistent times: building data, road network data, water system data, and land use data were collected in 2017; POI data were updated in 2019; and NTL data were collected in 2013. The data will change over time, leading to bias in model prediction in local areas. Future work will introduce additional geospatial data and improve the algorithm to increase the accuracy of the prediction model in densely populated areas.

Author Contributions: M.W.: methodology; software; visualization; formal analysis; writing—original draft preparation. Y.W.: data curation; formal analysis; investigation. B.L.: data curation; software. Z.C.: conceptualization; supervision; writing—review and editing. M.K.: conceptualization; methodology; writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Beijing Key Laboratory of Urban Spatial Information Engineering, grant number 20210211.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Finer Resolution Observation and Monitoring-Global Land Cover data are available at the link http://data.ess.tsinghua.edu.cn/ (accessed on 1 December 2021); the MSP-OLS nighttime lights imagery data are available at the link https://ngdc.noaa.gov/eog/ (accessed on 1 December 2021); actual population data are confidential and cannot be uploaded. Other related data and codes that support the findings of this study are available with identifiers at the link https://doi.org/10.5281/zenodo.6275126 (accessed on 25 February 2022).

Acknowledgments: Thanks to Shiliang, Su for guiding the research methodology. We appreciate the insightful comments from the editor and all the anonymous reviewers.

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

References

1. Wu, C.; Murray, A.T. A Cokriging Method for Estimating Population Density in Urban Areas. *Comput. Environ. Urban Syst.* **2005**, *29*, 558–579. [CrossRef]
2. Langford, M. An Evaluation of Small Area Population Estimation Techniques Using Open Access Ancillary Data: Small Area Population Estimation Techniques. *Geogr. Anal.* **2013**, *45*, 324–344. [CrossRef]
3. Deville, P.; Linard, C.; Martin, S.; Gilbert, M.; Stevens, F.R.; Gaughan, A.E.; Blondel, V.D.; Tatem, A.J. Dynamic Population Mapping Using Mobile Phone Data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15888–15893. [CrossRef] [PubMed]
4. Bakillah, M.; Liang, S.; Mobasheri, A.; Jokar Arsanjani, J.; Zipf, A. Fine-Resolution Population Mapping Using OpenStreetMap Points-of-Interest. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1940–1963. [CrossRef]
5. Gaughan, A.E.; Stevens, F.R.; Linard, C.; Jia, P.; Tatem, A.J. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **2013**, *8*, e55882. [CrossRef]
6. Bhaduri, B.; Bright, E.; Coleman, P.; Urban, M.L. LandScan USA: A High-Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. *GeoJournal* **2007**, *69*, 103–117. [CrossRef]
7. 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]
8. Jia, P.; Qiu, Y.; Gaughan, A.E. A Fine-Scale Spatial Population Distribution on the High-Resolution Gridded Population Surface and Application in Alachua County, Florida. *Appl. Geogr.* **2014**, *50*, 99–107. [CrossRef]
9. Ahola, T.; Virrantaus, K.; Krisp, J.M.; Hunter, G.J. A Spatio-temporal Population Model to Support Risk Assessment and Damage Analysis for Decision-making. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 935–953. [CrossRef]
10. Aubrecht, C.; Özceylan, D.; Steinnocher, K.; Freire, S. Multi-Level Geospatial Modeling of Human Exposure Patterns and Vulnerability Indicators. *Nat. Hazards* **2013**, *68*, 147–163. [CrossRef]
11. Hay, S.I.; Noor, A.M.; Nelson, A.; Tatem, A.J. The Accuracy of Human Population Maps for Public Health Application. *Trop. Med. Int. Health* **2005**, *10*, 1073–1086. [CrossRef]
12. Zhou, Y.; Ma, L.J.C. China's Urban Population Statistics: A Critical Evaluation. *Eurasian Geogr. Econ.* **2005**, *46*, 272–289. [CrossRef]
13. 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]
14. Mao, H.; Ahn, Y.-Y.; Bhaduri, B.; Thakur, G. Improving Land Use Inference by Factorizing Mobile Phone Call Activity Matrix. *J. Land Use Sci.* **2017**, *12*, 138–153. [CrossRef]
15. 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]
16. Deichmann, U. *A Review of Spatial Population Database Design and Modeling*; Technical Report 96-3; National Center for Geographic Information and Analysis: Santa Barbara, CA, USA, 1996.
17. Jones, H.R. *Population Geography*, 2nd ed.; Guilford Press: New York, NY, USA, 1990; ISBN 978-0-89862-464-9. [CrossRef]
18. Tobler, W.R. Smooth Pycnophylactic Interpolation for Geographical Regions. *J. Am. Stat. Assoc.* **1979**, *74*, 519–530. [CrossRef]
19. Langford, M.; Maguire, D.; Unwin, D. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In *Handling Geographical Information: Methodology and Potential Applications*; Longman Pub Group: London, UK, 2014.
20. Mennis, J.; Hultgren, T. Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartogr. Geogr. Inf. Sci.* **2006**, *33*, 179–194. [CrossRef]
21. Holt, J.B.; Lo, C.P.; Hodler, T.W. Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartogr. Geogr. Inf. Sci.* **2004**, *31*, 103–121. [CrossRef]
22. Eicher, C.L.; Brewer, C.A. Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartogr. Geogr. Inf. Sci.* **2001**, *28*, 125–138. [CrossRef]
23. Briggs, D.J.; Gulliver, J.; Fecht, D.; Vienneau, D.M. Dasymetric Modelling of Small-Area Population Distribution Using Land Cover and Light Emissions Data. *Remote Sens. Environ.* **2007**, *108*, 451–466. [CrossRef]
24. Mennis, J. Generating Surface Models of Population Using Dasymetric Mapping. *Prof. Geogr.* **2003**, *55*, 31–42. [CrossRef]

25. Su, M.-D.; Lin, M.-C.; Hsieh, H.-I.; Tsai, B.-W.; Lin, C.-H. Multi-Layer Multi-Class Dasymetric Mapping to Estimate Population Distribution. *Sci. Total Environ.* **2010**, *408*, 4807–4816. [CrossRef]
26. Langford, M. Rapid Facilitation of Dasymetric-Based Population Interpolation by Means of Raster Pixel Maps. *Comput. Environ. Urban Syst.* **2007**, *31*, 19–32. [CrossRef]
27. 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]
28. CIESIN; WRI. Gridded Population of the World (GPW), Version 2. In *Center for International Earth Science Information Network (CIESIN) Columbia University, International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI)*; CIESIN, Columbia University: Palisades, NY, USA, 2000.
29. Balk, D.L.; Deichmann, U.; Yetman, G.; Pozzi, F.; Hay, S.I.; Nelson, A. Determining Global Population Distribution: Methods, Applications and Data. In *Advances in Parasitology*; Elsevier: Amsterdam, The Netherlands, 2006; Volume 62, pp. 119–156. ISBN 978-0-12-031762-2.
30. CIESIN; CIAT. Global Rural-Urban Mapping Project (GRUMP), Alpha Version. In *Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI)*; Socioeconomic Data and Applications Center (SEDAC), Columbia University: Palisades, NY, USA, 2005.
31. Bright, E.A.; Coleman, P.R.; Dobson, J.E. LandScan: A Global Population Database for Estimating Populations at Risk. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 849–858.
32. Tatem, A.J.; Gaughan, A.E.; Stevens, F.R.; Patel, N.N.; Jia, P.; Pandey, A.; Linard, C. Quantifying the Effects of Using Detailed Spatial Demographic Data on Health Metrics: A Systematic Analysis for the AfriPop, AsiaPop, and AmeriPop Projects. *Lancet* **2013**, *381*, S142. [CrossRef]
33. European Commission, Joint Research Centre (JRC). *GHS-POP R2015A—GHS Population Grid, Derived from GPW4, Multitemporal (1975, 1990, 2000, 2015)—OBSOLETE RELEASE*; European Commission, Joint Research Centre (JRC): Brussels, Belgium, 2015. Available online: http://data.europa.eu/89h/jrc-ghsl-ghs_pop_gpw4_globe_r2015a (accessed on 1 December 2021).
34. 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]
35. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [CrossRef]
36. Azar, D.; Graesser, J.; Engstrom, R.; Comenetz, J.; Leddy, R.M.; Schechtman, N.G.; Andrews, T. Spatial Refinement of Census Population Distribution Using Remotely Sensed Estimates of Impervious Surfaces in Haiti. *Int. J. Remote Sens.* **2010**, *31*, 5635–5655. [CrossRef]
37. 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]
38. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying Urban Land Use by Integrating Remote Sensing and Social Media Data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
39. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
40. Cai, J.; Huang, B.; Song, Y. Using Multi-Source Geospatial Big Data to Identify the Structure of Polycentric Cities. *Remote Sens. Environ.* **2017**, *202*, 210–221. [CrossRef]
41. Zhang, Q.; Gao, W.; Su, S.; Weng, M.; Cai, Z. Biophysical and Socioeconomic Determinants of Tea Expansion: Apportioning Their Relative Importance for Sustainable Land Use Policy. *Land Use Policy* **2017**, *68*, 438–447. [CrossRef]
42. Su, S.; He, S.; Sun, C.; Zhang, H.; Hu, L.; Kang, M. Do Landscape Amenities Impact Private Housing Rental Prices? A Hierarchical Hedonic Modeling Approach Based on Semantic and Sentimental Analysis of Online Housing Advertisements across Five Chinese Megacities. *Urban For. Urban Green.* **2021**, *58*, 126968. [CrossRef]
43. Su, S.; Zhang, J.; He, S.; Zhang, H.; Hu, L.; Kang, M. Unraveling the Impact of TOD on Housing Rental Prices and Implications on Spatial Planning: A Comparative Analysis of Five Chinese Megacities. *Habitat Int.* **2021**, *107*, 102309. [CrossRef]
44. Yoshida, D.; Song, X.; Raghavan, V. Development of Track Log and Point of Interest Management System Using Free and Open Source Software. *Appl. Geomat.* **2010**, *2*, 123–135. [CrossRef]
45. McKenzie, G.; Janowicz, K.; Gao, S.; Yang, J.-A.; Hu, Y. POI Pulse: A Multi-Granular, Semantic Signature–Based Information Observatory for the Interactive Visualization of Big Geosocial Data. *Cartogr. Int. J. Geogr. Inf. Geovis.* **2015**, *50*, 71–85. [CrossRef]
46. Gao, S.; Janowicz, K.; Couclelis, H. Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-Based Social Networks: GAO et Al. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
47. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [CrossRef]
48. Lwin, K.; Murayama, Y. A GIS Approach to Estimation of Building Population for Micro-Spatial Analysis. *Trans. GIS* **2009**, *13*, 401–414. [CrossRef]
49. Loh, W.-Y. Classification and Regression Trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [CrossRef]
50. Goel, E.; Abhilasha, E. Random Forest: A Review. *Int. J. Adv. Res. Comput. Sci. Softw.* **2017**, *7*, 251–257. [CrossRef]

51. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random Forests: From Early Developments to Recent Advancements. *Syst. Sci. Control Eng.* **2014**, *2*, 602–609. [CrossRef]
52. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
53. Gaughan, A.E.; Stevens, F.R.; Huang, Z.; Nieves, J.J.; Sorichetta, A.; Lai, S.; Ye, X.; Linard, C.; Hornby, G.M.; Hay, S.I.; et al. Spatiotemporal Patterns of Population in Mainland China, 1990 to 2010. *Sci. Data* **2016**, *3*, 160005. [CrossRef]
54. Anyanwu, M.N.; Sajjan, S. Comparative Analysis of Serial Decision Tree Classification Algorithms. *Int. J. Comput. Sci. Secur.* **2009**, *3*, 230–240.
55. Resende, P.A.A.; Drummond, A.C. A Survey of Random Forest Based Methods for Intrusion Detection Systems. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
56. Scikit-Learn 1.0. Available online: Https://Github.Com/Scikit-Learn/Scikit-Learn (accessed on 26 December 2021).
57. Liu, Y. Mathematical Model of Multiple Linear Regression. *J. Shenyang Inst. Eng.* **2005**, 128–129.
58. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
59. Zhao, X.; Yu, B.; Liu, Y.; Chen, Z.; Li, Q.; Wang, C.; Wu, J. Estimation of Poverty Using Random Forest Regression with Multi-Source Data: A Case Study in Bangladesh. *Remote Sens.* **2019**, *11*, 375. [CrossRef]
60. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]