applied sciences

MDPI

Article

Modeling the Spatial Distribution of Population Based on Random Forest and Parameter Optimization Methods: A Case Study of Sichuan, China

Yunzhou Chen [1], Shumin Wang [1,*], Ziying Gu [1] and Fan Yang [2]

[1] Institute of Earthquake Forecasting, Beijing 100036, China; chenyz@ief.ac.cn (Y.C.); guziying2023@163.com (Z.G.)
[2] Information Center of Ministry of Natural Resources, Beijing 100812, China; fyang@infomail.mnr.gov.cn
* Correspondence: wangsm@ief.ac.cn

Abstract: Spatial population distribution data is the discretization of demographic data into spatial grids, which has vital reference significance for disaster emergency response, disaster assessment, emergency rescue resource allocation, and post-disaster reconstruction. The random forest (RF) model, as a prominent method for modeling the spatial distribution of population, has been studied by many scholars, both domestically and abroad. Specifically, research has focused on aspects such as multi-source data fusion, feature selection, and data accuracy evaluation within the modeling process. However, discussions about parameter optimization methods during the modeling process and the impact of different optimization methods on modeling accuracy are relatively limited. In light of the above circumstances, this paper employs the RF model to conduct research on population spatialization with multi-source spatial information data. The study primarily explores the differences in model parameter optimization achieved through random search algorithms, grid search algorithms, genetic algorithms, simulated annealing algorithms, Bayesian optimization based on Gaussian process algorithms, and Bayesian optimization based on gradient boosting regression tree algorithms. Additionally, the study investigates the influence of different optimization algorithms on the accuracy of population spatialization modeling. Subsequently, the model with the highest accuracy is selected as the prediction model for population spatialization. Based on this model, a spatial population distribution dataset of Sichuan Province at a 1 km resolution is generated. Finally, the population dataset created in this paper is compared and validated with open datasets such as GPW, LandScan, and WorldPop. Experimental results indicate that the spatial population distribution dataset produced by the Bayesian optimization-based random forest model proposed in this paper exhibits a higher fitting accuracy with real data. The Coefficient of Determination (R^2) is 0.6628, the Mean Absolute Error (MAE) is 12,459, and the Root Mean Squared Error (RMSE) is 25,037. Compared to publicly available international datasets, the dataset generated in this paper more accurately represents the spatial distribution of the population.

Keywords: population spatialization; random forest; model parameter optimization

Citation: Chen, Y.; Wang, S.; Gu, Z.; Yang, F. Modeling the Spatial Distribution of Population Based on Random Forest and Parameter Optimization Methods: A Case Study of Sichuan, China. *Appl. Sci.* **2024**, *14*, 446. https://doi.org/10.3390/app14010446

Academic Editor: Hari Mohan Srivastava

Received: 14 December 2023
Revised: 29 December 2023
Accepted: 30 December 2023
Published: 3 January 2024

1. Introduction

Population spatial distribution data serves as critical geographic information, effectively aiding governmental optimization of social resource allocation, environmental management, and urban development [1–5]. Additionally, integrating population information with other data offers a scientific basis for risk assessment, emergency disaster response, and post-disaster reconstruction [6–12]. Traditional population data is primarily obtained through comprehensive national surveys and sampling, typically using census divisions as statistical units. However, these boundaries do not always reflect the natural distribution of data. In practical applications, there are drawbacks such as low temporal

and spatial resolution, lack of support for spatial operations and analysis, and poor intuitiveness. With the support of geographic information technology, geographers enhance the study of the spatial distribution of socio-economic data by adding clear and detailed geographic references and employing spatial grid methods for quantification. This method of spatializing statistical data stores population data in a grid format, further enhancing computational efficiency and storage capacity. It promotes the integration of population statistics with other environmental data, making predictions and analyses of population-related issues more accurate. At the same time, it effectively addresses the limitations of traditional population data, improves the constraints of population census data in Earth science applications, enhances the spatial resolution of population data, and provides a more intuitive representation of spatial population distribution patterns. Traditional population spatialization methods primarily fall into two categories: spatial interpolation and statistical modeling. Spatial interpolation is a method of converting population data from a large spatial range to a small area, such as point interpolation [13] and areal interpolation [14]. These methods make the scale conversion of population data convenient but often overlook scale and boundary effects [15], leading to suboptimal performance at the boundaries of regions. Moreover, due to the presence of assumption conditions, it becomes challenging to account for spatial heterogeneity issues [16], resulting in inaccurate estimates in heterogeneous areas. The presence of outliers or extreme values can also significantly impact interpolation results, leading to inaccurate estimations. Statistical modeling methods establish population spatialization models based on the weight relationships between various auxiliary data and the spatial distribution of the population, allowing for estimates of population quantity or density in small spatial units. Compared to spatial interpolation methods, statistical models can consider the intricate relationships between various factors and population density more comprehensively, but they need to face the challenge of multi-source heterogeneous data fusion [17]. Representative methods include multiple linear regression (MLR) [18], geographically weighted regression (GWR) [19,20], spatial lag regression model [21], kriging regression model [22], etc. Moreover, many statistical regression models are based on the assumption of linear relationships, but the distribution of populations and related factors may exhibit non-linear associations. This can lead to inaccurate modeling of the true relationships. Statistical regression models often fail to capture the complexity of geographical spatial structures, such as variations between urban centers and suburbs. This may result in models that are overly smooth in space, overlooking local variations. With the development of machine learning techniques, scholars have applied ensemble learning and neural networks to population spatialization exploration to further explore the complex relationships between multi-source geographic information and demographic statistical features. Ensemble learning combines multiple learners into a unified entity through certain strategies to jointly complete tasks and enhance decision accuracy through collective decision-making, primarily involving boosting and bagging algorithms. The Boosting algorithm follows the principle of gradient boosting [23], and updates the model by feeding back the information of each round of model training to the next round, obtaining a better model based on the residual iterative training of the previous round of models. On this basis, Extreme Gradient Boosting (XGBoost) employs weighted fusion to average the results of each tree for final output, effectively enhancing model accuracy [24]. Zhao Xin et al. [25] estimated the population distribution of Shenzhen in 2019 based on five ensemble learning models, and the XGBoost model achieved the best results. Bagging algorithms combine results from multiple learners through averaging or voting to obtain predictive results [26]. As a typical bagging algorithm, random forest (RF) is widely used in population spatialization, and it possesses several advantages compared to other algorithms. It features a more flexible and stable framework. Random Forest integrates predictions from multiple decision trees, with each tree learning from the data in a different way, thereby reducing the risk of overfitting and improving the overall model's generalization ability. This leads to more accurate predictions, helps avoid overfitting, and exhibits higher tolerance to outliers and noise [27]. Population spatialization studies

often involve various types of data, encompassing a large number of features. Random Forest's flexible and stable framework allows it to effectively handle high dimensional feature spaces, enabling the model to thoroughly consider the impact of different data. Furthermore, since each decision tree is trained independently, Random Forest inherently benefits from parallelization, which accelerates the model training process. This capability is useful in efficiently handling the extensive geographical, social, and economic data involved in the study. Stevens et al. [28] utilized the random forest model in regions such as Vietnam, Cambodia, and Kenya to generate high-precision population grid data with 100 m resolution. Li et al. [29], based on 25 m nighttime light (NTL) data and point of interest (POI) data captured by the International Space Station (ISS), proposed a population spatialization approach for constructing high-resolution urban population distribution data using the random forest method. Ye et al. [30] utilized the random forest model and integrated POI data and multi-source remote sensing data to map China's 2010 population data to 100 m grids. Liu et al. [31] integrated POI data and other multi-source data, and used the random forest algorithm to conduct refined mapping of the population of Zhengzhou City at three scales: 50 m, 300 m, and 500 m, achieving excellent spatialization results. Taking the spatialization of population in Beijing as an example, He et al. [32] compared and analyzed different methods, such as RF, MLR, XGBoost, support vector machine (SVM), back propagation neural network (BPNN), and least absolute shrinkage and selection operator (LASSO), and the results show that RF is superior to other methods. The neural network is a mathematical model that emulates the structure and function of biological neural networks, often used to model complex relationships between inputs and outputs. Although neural networks have made initial attempts and achieved certain effects in population spatialization, their interpretability is limited, and the generalization of the model requires further testing and evaluation [33,34]. Additionally, methods based on neural networks needs end-to-end mapping training, and obtaining population data at a fine grid scale is challenging; this paper does not delve into a detailed exploration of such methods.

There are publicly available population grid datasets, such as the Gridded Population of the World (GPW) [35], Global Rural Urban Mapping Project (GRUMP) [36], LandScan [37], WorldPop [38], etc. These population datasets are mainly created using areal weighting method, intelligent interpolation method, random forest algorithm, etc. Since these datasets simulate population distribution on the global scope, the modeling conditions vary significantly in different regions, making it difficult to ensure model accuracy in areas with complex environments [39]. Gunasekera et al. [40] found that LandScan performs well in modeling urban population distribution but has lower reliability in rural areas. Sabesan et al. [41] compared the differences between LandScan and GPWv3 datasets in many regions and found that LandScan has a better ability to represent the heterogeneity of population spatialization. Bai et al. [42] compared and analyzed the errors of GPWv3, GRUMPv1, WorldPop, and China Specific Population Grid (CnPop) at the township scale. The results show that the WorldPop dataset has the highest accuracy, but it also has large errors in hilly areas such as the Hengduan Mountains.

In summary, the random forest algorithm can integrate multi-source geographic information data, effectively modeling the complex relationships between population data and spatial distribution indicators, and perform population grid predictions at various resolutions. However, the work of domestic and international scholars primarily focuses on aspects such as multi-source data fusion during the random forest modeling process, remote sensing data processing, and the generation of population datasets with different resolutions. There has been limited exploration into the impact of different parameter optimization methods within the random forest modeling process on population spatialization. Furthermore, the publicly available datasets still require further improvements in accuracy in certain regions.

Therefore, this paper will further enrich data sources, improve data quality, and conduct population spatialization modeling research by combining multi-source remote

sensing geographic information data and utilizing the currently prevalent random forest algorithm. The primary focus of the article lies in the meticulous exploration of methodologies for refining and adjusting the parameters of the random forest model. Special attention is given to scrutinizing the impact of various parameter optimization techniques on the model's accuracy. Subsequently, by combining cross-validation methods, the optimal model parameters will be selected to enhance the model's structure and improve predictive accuracy. Finally, the population spatialization model is constructed based on optimal parameters, and a spatial population distribution dataset of Sichuan Province at the 1 km resolution is generated. At the same time, the dataset developed in this study is compared with public datasets such as GPW, LandScan, and WorldPop for verification.

2. Study Area and Data

2.1. Study Area

Sichuan Province is located in the southwest of China, situated in the upper reaches of the Yangtze River. It spans between 26°03′ to 34°19′ north latitude and 97°21′ to 108°12′ east longitude, bordered by seven provinces including Chongqing, Guizhou, Yunnan, Tibet, Qinghai, Gansu, and Shaanxi. Sichuan Province exhibits significant differences in its eastern and western topography, with a complex and diverse terrain characterized by a west-high-east-low topographical feature. It encompasses mountains, hills, plains, basins, and plateaus, covering a total area of 486,000 square kilometers. The province is administratively divided into 21 prefecture-level administrative regions, 183 county-level divisions, and 3101 township-level divisions. According to the Seventh National Population Census of China, the permanent population of Sichuan Province in 2020 was 83.67 million, ranking fifth in the nation in terms of total population. In recent years, Sichuan has experienced rapid socio-economic development, but it has also faced frequent geological disasters. This poses a significant challenge for urban management and emergency disaster preparedness in densely populated areas. Therefore, this study selected Sichuan Province as the study area (Figure 1).

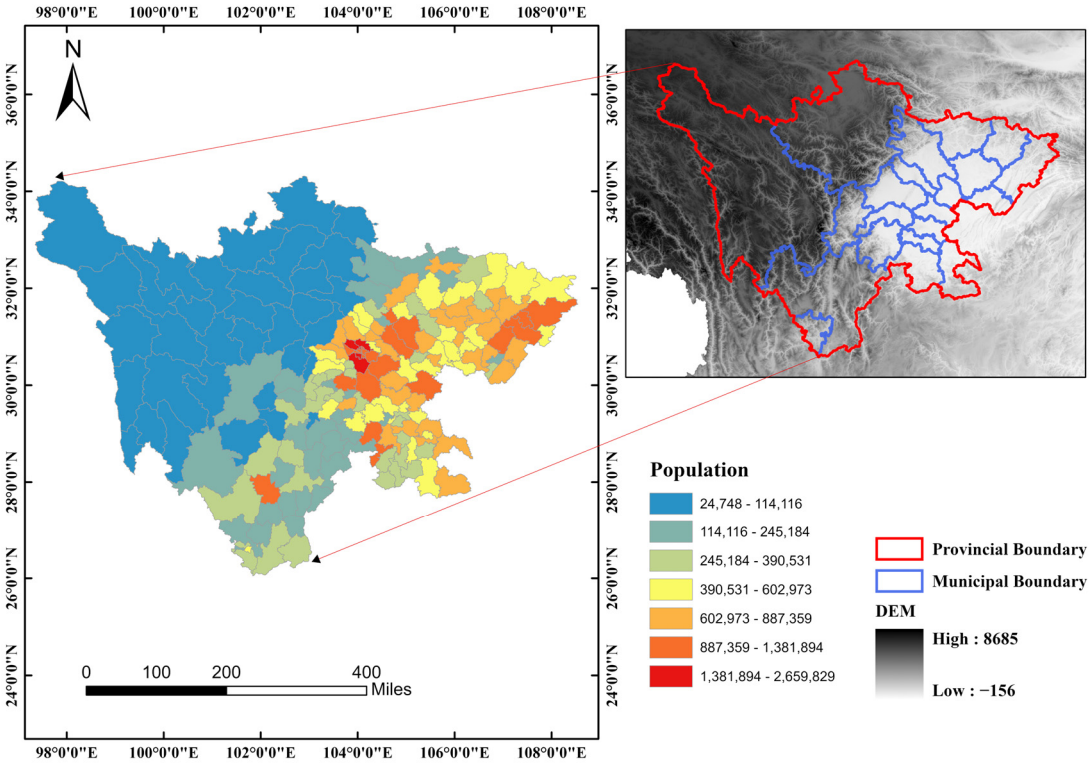


Figure 1. Population distribution map of the study area in 2020.

2.2. Data Preparation

Scholars have made many attempts and explorations to enrich data sources and optimize data processing on population spatialization. The data sources mainly include land use and land cover data [18], terrain data [43], vegetation index data [44], nighttime light data [45], social perception big data such as point of interest (POI) [46], as well as building height data [47]. The data sources show a trend of diversification and refinement. The population spatialization-related data collected for this study primarily include the population census data of Sichuan Province in 2020, basic geographic information data, natural environmental data, building height data, nighttime light (NTL) data, and the social perception data. The details and sources of the data used in this study are presented in Table 1.

Table 1. The details of data used in this study.

Category	Name	Time	Spatial Scale	Format	Source
(a) Demographic data	Census data	2020	county-level and township-level	Table (csv)	http://www.citypopulation.de/ (accessed on 21 February 2022)
(b) Basic geographic information data	Administrative boundaries data	2020	1:100 w	Vector (Polygon)	https://www.webmap.cn/ (accessed on 19 October 2022)
	Road network	2021 Version	1:100 w	Vector (Polyline)	https://www.webmap.cn/ (accessed on 19 October 2022)
	River network	2021 Version	1:100 w	Vector (Polyline)	https://www.webmap.cn/ (accessed on 19 October 2022)
(c) Natural environmental data	ASTER GDEM v3	2000	30 m	Raster	https://www.gscloud.cn/ (accessed on 23 October 2022)
	Globeland 30	2020	30 m	Raster	http://www.globallandcover.com/ (accessed on 7 April 2022)
	NDVI (Landsat 8)	2020	100 m	Raster	https://www.usgs.gov/ (accessed on 20 October 2022)
(d) Building height data	CNBH-10m	2020	10 m	Raster	https://zenodo.org/ (accessed on 6 June 2023)
(e) Nighttime light (NTL) data	NPP-VIIRS v2.1	2020	500 m	Raster	https://eogdata.mines.edu/ (accessed on 23 October 2022)
	LJ-01	2019	130 m	Raster	http://datasearch.hbeos.org.cn:3000/ (accessed on 25 October 2022)
(f) Social perception data	Point of interest (POI)	——	——	Vector (Point)	https://ditu.amap.com/ (accessed on 9 March 2022)

The demographic data is the seventh national census data in 2020, which mainly includes the population data of 183 county-level administrative divisions and the population data for a partial township-level administrative division in Sichuan Province, totaling

2122 units. This data is primarily obtained from the official announcements released by local governments regarding the Seventh National Population Census. Census agencies ensure the representativeness of the data through sampling and statistical methods, ensuring that the data accurately reflects the characteristics and distribution of the entire population.

Basic geographic information data originates from the National Catalogue service For Geographic Information (http://www.webmap.cn/ (accessed on 19 October 2022)). It includes vector data for administrative boundaries, road networks, railway networks, and water systems within Sichuan Province. The overall temporal reference for this dataset is 2019, using the 2000 National Geodetic Coordinate System and the 1985 National Elevation Datum, expressed in latitude and longitude coordinates.

Natural environmental data primarily include elevation, slope data, land cover data, and normalized vegetation index data of Sichuan Province. The elevation data is sourced from the ASTER GDEM V3 product of the Geospatial Data Cloud (https://www.gscloud.cn/ (accessed on 23 October 2022)), and the slope data is calculated from elevation data. The land cover data comes from GlobeLand30, which is the first 30 m global land cover data product on the world developed by the National Geomatics Center of China. It mainly has 10 land cover types, such as cropland, forest, grassland, shrubland, wetland, water bodies, tundra, artificial surfaces, bare land, glaciers, and permanent snow and ice. Normalized vegetation index data is calculated using Landsat 8 SR data through the Google Earth Engine platform.

Building height data is sourced from the CNBH dataset produced and released by the GC3S team of the School of Life Sciences at Fudan University. This dataset has demonstrated strong correlation between simulated results and actual building observations [48], making it a crucial element in studying urban environments and population distribution.

Nighttime light data consists of the annual data products NPP-VIIRS V2.1 from the Earth Observation Group of the National Centers for Environmental Information in the United States, as well as data from the Luojia-01 satellite released by the High-Resolution Earth Observation System Hubei Data and Application Network (http://www.hbeos.org.cn (accessed on 25 October 2022)).

Social perception data is acquired through batch extraction of Points of Interest (POI) using the application programming interface (API) provided by Amap (Amap is a mapping, navigation, and location-based service provider). More than 3,340,000 records are obtained. POI data typically includes information such as name, address, longitude, latitude, and category. After data cleaning and filtering, 11 categories of POI data are selected as study objects. The details of POI data used in this study are listed in Table 2.

Table 2. The details of POI categories.

NO.	Reclassified Label	Examples	Quantity
1	poi_administration	Authority, federation, committee, fire station, etc.	103,208
2	poi_company	Factories, farms, bases, warehouses, business units, etc.	168,788
3	poi_education	Kindergartens, schools, art institutions, conference centers, training institutions, etc.	73,534
4	poi_hospital	Hospitals, pharmacies, health stations, emergency centers, clinics, etc.	107,438
5	poi_hotel	Inns, hotels, apartments, villas, homestays, clubs, etc.	53,831
6	poi_lifestyle services	Maintenance site, logistics distribution, beauty salon, service center, etc.	371,364
7	poi_residence	Community, dormitory, villa, residential area, etc.	54,992
8	poi_restaurant	Restaurant, restaurant, noodle shop, bakery, tea house, etc.	406,194
9	poi_scenery	Ethnic areas, scenic spots, ancient towns, temples, resorts, old sites, squares, etc.	12,716
10	poi_shopping	Supermarkets, shopping malls, specialty stores, building materials markets, hardware markets, etc.	921,792
11	poi_sportleisure	Swimming pools, gymnasiums, clubs, gyms, theaters, living halls, etc.	60,976

3. Methods

This study is primarily divided into four parts, including data collection, data processing, modeling and optimization, and validation, as shown in Figure 2. Firstly, data was collected from various aspects such as remote sensing, natural environment, urban construction, and social perception. These geospatial data often originate from different

sources and have varying resolutions. In order to unify the geographic reference information during the modeling process and facilitate the sampling of features from different layers at the same locations, operations such as calibration, projection, and resampling were performed on these data. Additionally, to ensure that data features can be directly used for spatial modeling, vector data needed to be transformed into quantifiable raster layer information. Furthermore, by combining various feature selection methods, features that more effectively capture the spatial distribution of the population were chosen as input data for the model. Random Forest (RF) was chosen as the modeling method. Simultaneously, various optimization algorithms like random search, grid search, genetic algorithm, simulated annealing, and Bayesian optimization were applied to optimize the model parameters, aiming to build the best model. Finally, using the random forest model with optimized parameters, we predicted and modeled the spatial distribution of the population in Sichuan Province for the year 2020. We obtained the population spatialization dataset with 1 km resolution. Subsequently, the obtained population grid dataset was compared with GPW, LandScan, and WorldPop at the township scale for analysis and accuracy verification.

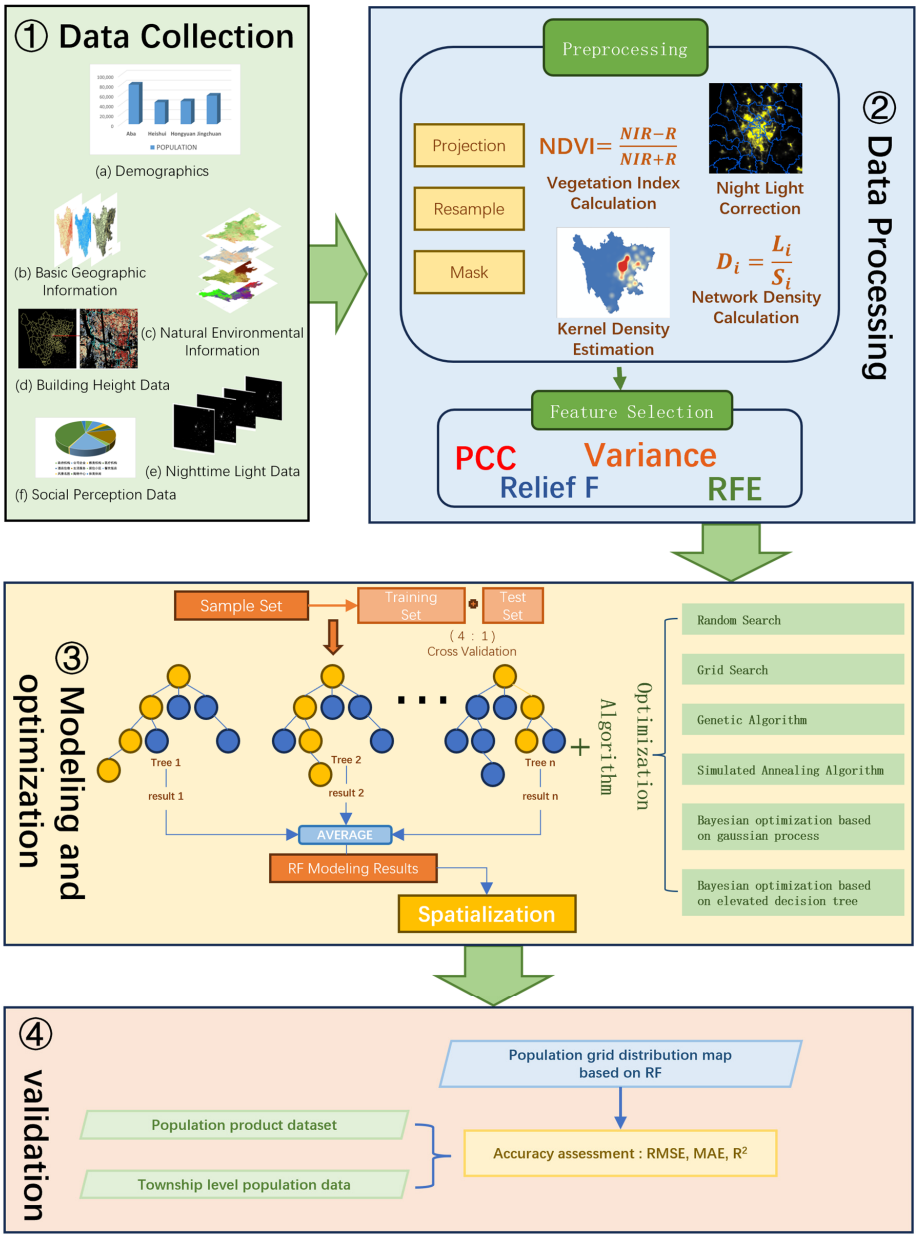


Figure 2. Flowchart of population spatialization.

3.1. Data Processing

3.1.1. Geographical Information Data Processing

Due to the differences in data sources, spatial resolutions, and projection coordinates, preprocessing operations are applied to all data in order to mitigate the impact of different image resolutions and coordinate references on modeling. These operations mainly include following aspects. (1) Unified Coordinate System: To ensure consistent area preservation during calculations, all data are transformed into the Asia North Albers Equal Area Conic projection coordinate system. (2) Unified Spatial Resolution: For data with spatial resolutions smaller than 1 km, bilinear interpolation is employed to up sample the data to a 1 km resolution.

In order to transform vector data into quantifiable and meaningful spatial features, we perform density calculations on traffic network data and water system data. Through these operations, we obtain the aggregation degree of traffic and water systems in the kilometer grid [30] (in this article, represented as D_railway, D_Road, and D_Hydrological). The specific calculations are as follows:

$$D_i = \frac{L_i}{S_i} \tag{1}$$

In Formula (1), D_i is the network density of line vector data in region i, L_i is the sum of the length of line data in region i, and S_i is the land area of region i.

Normalized Vegetation Index (NDVI) data is calculated based on the Google Earth Engine (GEE) platform. It involves computing the NDVI index for each time period using all Landsat 8 Surface Reflectance (SR) data from the year 2020. The annual maximum NDVI is derived using the Maximum Value Composite method [49], and this parameter is employed to describe the vegetation coverage in Sichuan Province.

$$NDVI_{max} = MAX(NDVI_1, NDVI_2, NDVI_3, \ldots, NDVI_n), \tag{2}$$

$NDVI_{max}$ is the maximum NDVI value of the pixel, and $NDVI_n$ indicates the NDVI value at the same pixel on the nth NDVI image of the year 2020.

3.1.2. Nighttime Light Data Processing

Although obtaining nighttime light data is relatively convenient, accurate reflection of human activity requires data correction and extraction. The following operations were carried out in this study.

1. Radiometric Correction

Since the acquired NPP-VIIRS data is a pre-processed annual composite product (Annual VNL V2.1), there is no need for additional radiometric correction. However, the obtained Loujia-01 data is recorded as the gray value of the pixel, lacking a direct physical interpretation. Therefore, it needs to be corrected using the radiance conversion formula to address radiometric distortion errors.

$$L = DN^{\frac{3}{2}} \times 10^{-10}, \tag{3}$$

where L represents the radiance value after absolute radiometric correction, measured in $W/(m^2 \cdot sr \cdot \mu m)$.

2. Background Noise and Abnormal Value Correction

Nighttime light anomalies refer to pixels with abnormally high brightness values. In this study, NPP-VIIRS data and Loujia-01 data were processed using threshold extraction and median filtering. Threshold extraction is using nighttime light image pixel values collected in the most economically developed region of the study area as the maximum DN value; other bigger pixel values are anomalies caused by phenomena like fishing lights. Additionally, zero is set as the minimum value in the study area, and the negative values are removed. Since Chengdu city is the most economically developed city in Sichuan,

we extract the maximum value of nighttime light data from Chengdu as the maximum threshold. Median filtering is a common denoising technique in image processing. It replaces the pixel value with the median value of the adjacent pixels. The method is achieved through template traversal and is effective in removing isolated noise points from images.

3. Stable Light Source Extraction

Due to the lack of filtering and removal of volcanic activity-related background noise in NPP-VIIRS data, Loujia-01 data presents a more stable signal. Given that stable light sources between neighboring years do not exhibit significant variations, the calibrated 2019 Loujia-01 nighttime light data was used as a reference standard to correct the unstable light sources in the 2020 NPP-VIIRS data. Based on the assumption that both the 2019 Loujia-01 and 2020 NPP-VIIRS data share the same light generation areas, a mask is created by extracting pixels with non-zero DN values from Loujia-01. The data outside this mask in the NPP-VIIRS dataset is set to 0, effectively treating it as unstable nighttime light [50].

3.1.3. POI Data Processing

POI is discrete point data, which cannot be directly used in the population spatialization model. It is necessary to transform discrete data into a continuous density image. Kernel density analysis is a density analysis method based on the first law of geography, which considers spatial variations and reflects the characteristic of feature center intensity decay with distance. In this method, points closer to the center are assigned higher weights, while those farther away receive lower weights. The estimated density for each point is the weighted average density of all points in that region. In kernel density calculations, the choice of kernel function and bandwidth significantly impacts the results of kernel density analysis. We utilize the kernel density analysis tool in ArcGIS 10.7 to analyze 11 categories of POI data. The selected kernel function is the quartic kernel function, as shown in Equation (4). This function is relatively smooth. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the Search radius distance from the point. Compared to the Gaussian kernel function, it has a faster convergence rate. Bandwidth calculation is determined through Equation (5), which avoids the "ring around the points" phenomenon that often occurred with sparse datasets. Ultimately, density maps for the 11 categories of POI data are generated with a resolution of 1 km.

$$\text{Density} = \frac{1}{(Radius)^2} \sum_{i=1}^{n} \left[\frac{3}{\pi} p_i \left(1 - \left(\frac{dist_i}{Radius} \right)^2 \right)^2 \right] \tag{4}$$

$$Radius = 0.9 * min\left(SD, \sqrt{\frac{1}{\ln 2}} * D_m \right) * n^{-0.2}, \tag{5}$$

Here, $i = 1, \ldots, n$ is the number of input points. Only include points in the sum if they are within the radius distance of the (x, y) location. *Radius* is the search radius. p_i is the population field value of point i, which is an optional parameter. $dist_i$ is the distance between point i and the (x, y) position. *SD* is the standard distance. D_m is the median distance. n is the sum of the population field values.

3.2. Feature Selection

In this study, five types of auxiliary feature variables, including nighttime lights, natural environment, basic geographic information, social perception information, and building height data, were preprocessed to obtain 31 quantifiable grid statistical features. These features were then subjected to partitioned statistics at the county level administrative units. The density values of the auxiliary feature variables were obtained by combining the area information of the county-level administrative units. Simultaneously, by organizing the

data collected from the seventh population census, population information was aggregated at the county level to derive total population and average population density. In this context, the Pearson correlation coefficient method was employed to select one variable with higher correlation with auxiliary information as the target variable.

After collecting and preprocessing, the feature data exhibited a diverse range, with some remaining correlations and redundancies among the data. Bringing all features into a spatial model would increase the model's complexity and computational load. Feature selection helps reduce irrelevant features and improve algorithm performance [51]. Therefore, in this study, various feature selection techniques were explored from different perspectives, including the well-known Pearson correlation coefficient method, variance threshold selection method, ReliefF feature selection method, and recursive feature elimination method. These techniques were used to identify the most suitable features as modeling factors.

3.2.1. Pearson Correlation Coefficient Method

The Pearson correlation coefficient is a statistical method used to measure the strength and direction of the linear relationship between two variables. In feature selection, the Pearson correlation coefficient can be employed to assess the correlation between each feature and the target variable. The goal of feature selection is to choose the most relevant or important features from the original feature set to enhance model performance, reduce dimensionality, and mitigate the risk of overfitting. Mao et al. [52] employed the Pearson correlation coefficient (PCC) to select Points of Interest (POI) information, considering values with an absolute correlation coefficient greater than 0.5 as indicators of social factors influencing Residential Population (RPI). Lu et al. [53] utilized the Pearson correlation coefficient (PCC) to assess the correlation between Numerical Weather Prediction (NWP) variables and wind power. They selected NWP features with PCC values greater than 0.5 to predict future wind power.

3.2.2. Variance Threshold Method

The variance threshold method involves feature selection through removing features with low variance in the dataset. When constructing a model, if a feature has low variance, it indicates limited variability, and its contribution to the model is minimal. This suggests that the impact on model performance can be negligible. Therefore, low-variance filtering methods can be used to eliminate such features. Wang et al. [47] applied the low-variance filtering feature selection method to choose 30 features from a multidimensional feature library as auxiliary data for spatializing population.

3.2.3. ReliefF Method

In 1994, Kononeill extended the Relief algorithm, resulting in the ReliefF algorithm designed to address regression problems with target attributes as continuous values [54]. The ReliefF algorithm is a type of feature weighting algorithm that assigns higher weights to features highly correlated with the target variable. Features with weights below a certain threshold are removed.

3.2.4. Recursive Feature Elimination Method

Recursive Feature Elimination (RFE) is a method of feature selection that involves recursively removing features with minimal contribution to the model. It is a classical wrapper method that considers interactions between features and assesses feature importance by constructing models [55].

3.3. Random Forest Model Construction and Parameters Optimization

3.3.1. Model Construction

The random forest algorithm is a machine learning technique that involves training and predicting with numerous decision trees collectively. It is primarily based on bagging ensemble learning theory, where decision trees are constructed by randomly drawing

an equivalent training sample with replacement from the sample set. The algorithm employs a random subspace method to select a subset of features from the entire feature set and chooses the optimal features to split the nodes of the decision tree. The final result is determined collectively with multiple independent decision trees. This approach effectively constructs complex nonlinear relationships between population data and geo graphical factors. These random selection processes ensure better feature acquisition, improving model robustness and preventing overfitting. Additionally, due to the independence between decision trees, they can be generated in parallel during the tree-building process, enhancing the efficiency of the algorithm.

We constructed the population spatialization prediction model based on random forest algorithm using the Python language and the Scikit-learn open-source library. A sample dataset comprising 183 county-level administrative units of Sichuan Province will be used as the training sample. A total of 27 highly correlated features are selected as independent variables, while population data of county-level administrative units are used as the dependent variable. In total, 80% of the samples are used as training data, and the remaining 20% constitute the validation samples. The dataset is divided with fixed random numbers to ensure the randomness of dataset division and the replicability of experiments.

3.3.2. Model Parameter Optimization

The key parameters in random forest modeling are shown in Table 3. Different parameter combinations have a significant impact on the performance of the random forest model. Therefore, adjusting the parameters of the random forest algorithm is the key step to optimizing model performance. In this paper, we compare and analyze the methods of model parameter optimization, including random search, grid search, genetic algorithm, simulated annealing, Bayesian optimization based on Gaussian process, and Bayesian optimization based on gradient boosting regression trees. The R2_score is used to assess the fitness of each decision tree, serving as a critical indicator of the overall model fitting performance.

Table 3. Key parameters of the random forest model.

Parameter	Function
n_estimators	The number of decision trees in the forest. Increasing the value of this parameter can enhance the model's accuracy and robustness. However, it can also increase the computation time and memory usage.
max_depth	The maximum depth of each decision tree, which controls the complexity of the model and the risk of overfitting. If the depth of the tree is too large, the model might overfit. If it's too small, the model might underfit.
max_features	The number of features randomly chosen at each node, which can make the model more randomized and reduce the risk of overfitting when increased. However, it might also decrease the accuracy of the model prediction.
min_samples_split	The minimum number of samples required to continue splitting at each node, which can prevent overfitting by increasing its value. But this might reduce the sensitivity of the model.
min_samples_leaf	The minimum number of samples required in each leaf node, which can prevent overfitting. Increasing this value can stabilize the model, but it might affect the sensitivity of the model.

In the process of random forest construction, it is common to utilize methods like random search and grid search to select parameters such as the depth of decision trees, the number of leaf nodes in each tree, and the number of nodes for splitting in each tree. The random search is a method to randomly select parameter combinations in the parameter space, allowing for exploration of a larger parameter space. While it efficiently explores different combinations, it may require more iterations to find the optimal parameter combination. Grid search systematically traverses the parameter space by training and evaluating the model for each possible parameter combination. It ensures comprehensive parameter space exploration but might face computational resource challenges, particularly in large parameter spaces.

Genetic algorithm is an optimization algorithm inspired by natural selection and genetic mechanisms. It simulates the genetic process in biological evolution, searching the solution space of the problem through genetic operations. Genetic algorithm possesses global search capabilities, allowing for extensive exploration within the solution space. By employing crossover and mutation operations on multiple individuals within populations, genetic algorithms can escape local optima and strive to find the global optimum. Simulated annealing algorithm is derived from the physical process of annealing in solids. The idea comes from the fact that in solid-state annealing, gradually lowering the temperature allows atoms to move towards an equilibrium state, resulting in the optimal crystalline structure. This method avoids falling into local optimal solutions by accepting suboptimal solutions with a certain probability in the search space.

Bayesian optimization is a method of optimizing black-box functions by building models, suitable for scenarios, where minimizing or maximizing an objective function is challenging or unknown. We compare two Bayesian optimization algorithms; one is based on Gaussian process (GP) modeling and the other is based on gradient boosting regression trees (GBRT). Gaussian Process is a probabilistic model assuming that the objective function is a continuous and smooth function, estimating the distribution of the objective function across the entire parameter space based on available data points. GBRT is an ensemble learning method that combines multiple decision trees to model the target function. It iteratively fits the model and optimizes the loss function to improve predictive performance. Both methods optimize iteratively, updating models based on feedback from the objective function at each iteration, and find the optimal solution within a limited number of iterations. However, their strategies for selecting the next sampling point differ slightly. We compare these two strategies and ultimately choose the most suitable approach for modeling. Compared to traditional grid search and random search methods, Bayesian optimization is more efficient, especially in high-dimensional parameter spaces, allowing it to find the optimal solution more quickly. Compared with genetic algorithm and simulated annealing, Bayesian optimization can model and infer uncertainties using prior information and posterior probabilities, and have better ability to handle noise and uncertainty cases. Furthermore, this method can quickly find the global optimal solution without manually setting parameters.

In the process of parameter optimization, this paper takes into account the complexity of the research and the size of the sample set, and sets the parameters for the above-mentioned parameter optimization methods based on empirical adjustment. Among them, the range of hyperparameters for the random forest is constrained as shown in Table 4.

Table 4. The range of random forest hyperparameters.

Parameter	Min	Max
n_estimators	100	500
max_depth	3	27
max_features	1	27
min_samples_split	2	15
min_samples_leaf	1	10

Cross-validation is a method for evaluating model performance. It involves repeatedly splitting the dataset into different training and testing sets, training and testing the model multiple times, and obtaining scores for the model on different dataset splits. This approach ensures that all data is thoroughly used for training and testing, helps understand how the model handles different data subsets, effectively prevents overfitting [56], and improves the accuracy and stability of the model.

Specifically, parameter selection is performed according to the following steps:

(a) Data Splitting: We partition the training dataset into 5 independent and equally sized subsets, referred to as "folds," and no replacement is made.

(b) Model Training and Validation: In each training round, we perform 5 iterations, where in each iteration, 4 of the folds are chosen, and we combine them with parameter optimization methods to train the model. The remaining 1 fold is used as the validation set, and we calculate the R2_score to evaluate the model produced during this training session. After 5 iterations, we obtain 5 performance evaluation metrics.
(c) Model Evaluation: We compute the average of the performance metrics from all 5 iterations and use it as the performance metric for that round of model training.

In the end, we select the optimal parameter set obtained through cross-validation evaluation in each method's training to construct the population spatialization model.

3.4. Population Spatialization

We collect the feature values of all variables of each grid and create a 1 km grid dataset with multidimensional attribute features. Then, we utilize the parameter optimized random forest model to predict population spatialization, generating the estimated population weight value of each grid. Finally, Equation (6) is employed to calculate the population within each grid.

$$Pop2020_{ij} = \frac{Pre_{ij}}{\sum_{j=1}^{n} Pre_{ij}} \times county_i \,, \tag{6}$$

In Formula (6), $Pop2020_{ij}$ represents the final predicted population in grid j of county i, Pre_{ij} is the predicted value obtained by random forest model for grid j in county i, $County_i$ is the total resident population of the county i in the seventh population census, and n is the total number of grids in the county i.

4. Results

4.1. Feature Selection Results

This study employs Pearson correlation coefficient to assess the linear correlation between each feature and the target variable. Specifically, by calculating the Pearson correlation coefficient between each feature and the target variable, the absolute value of the coefficient is used as a measure of feature importance. A value close to 1 indicates a strong correlation, while a value close to 0 suggests almost no correlation. This approach is utilized to select features that are most correlated with the target variable. The study ultimately presents the correlation between different population information and various auxiliary feature variables, including Pearson correlation coefficients and significance levels, in the form of a heatmap, as shown in Figure 3a.

Through the comprehensive analysis of the correlation between population information and feature variables, population density exhibits higher correlation compared to total population. Therefore, population density is selected as the target variable for modeling. Additionally, Points of Interest (POI) data, terrain characteristics, nighttime lights data, building height data, and transportation and water system data show strong correlations with population density, with p-values less than 0.001, indicating statistical significance and suitability for model construction.

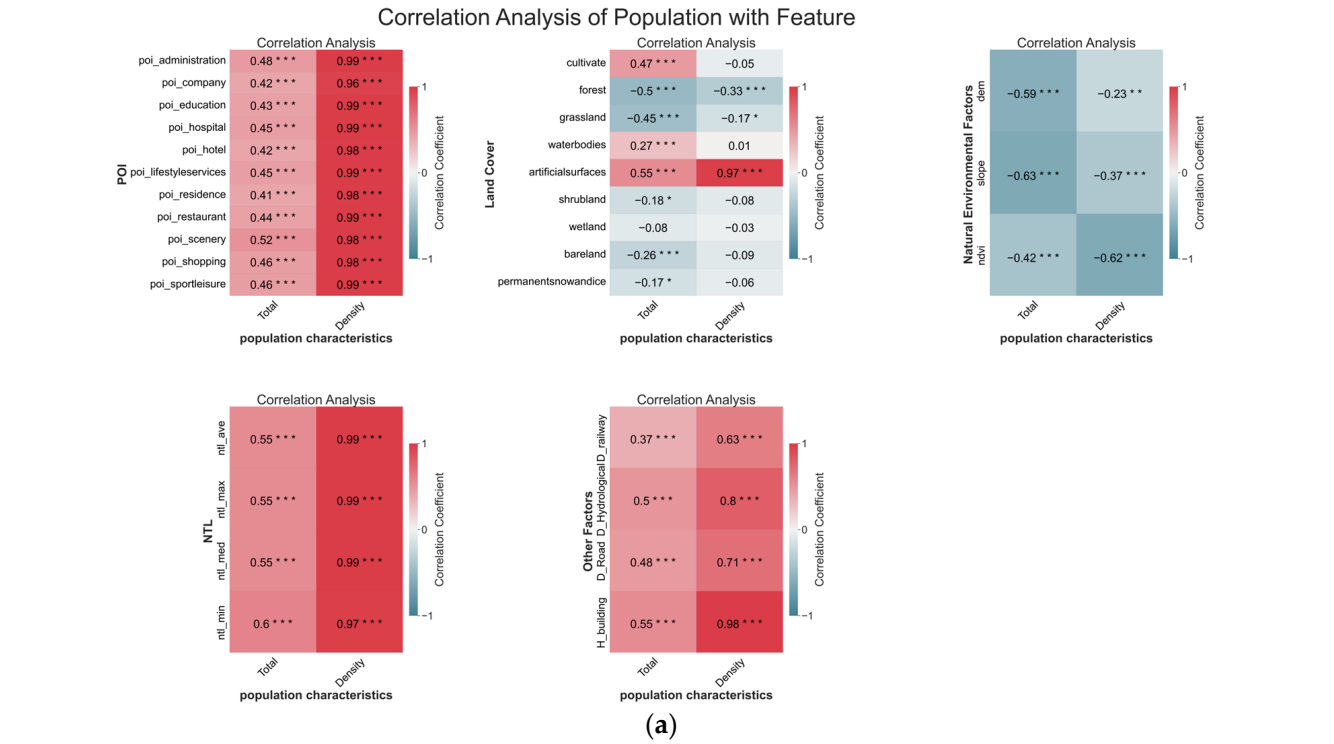
According to the variance threshold selection method, Figure 3b is obtained by calculating the variance of each feature in the dataset. From the figure, it is evident that wetland has low variability in the dataset, providing limited information to the model. Therefore, it may be considered for removal during the modeling process.

Based on the ReliefF algorithm, Figure 3c is obtained by assessing the degree of correlation between features and the importance of each feature to the target variable, population density. The figure reveals that shrubland, waterbodies, bareland, grassland, permanent snow and ice, and dem have the lowest weights and importance levels with respect to the target variable. These features can be considered for removal.

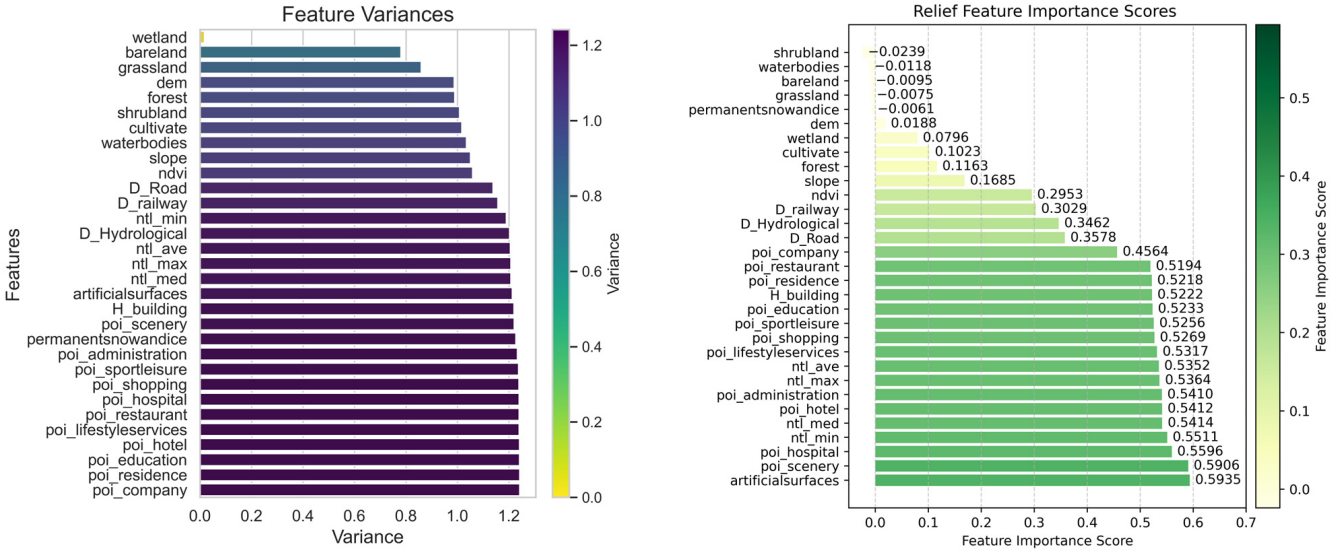
Utilizing the recursive feature elimination method with a random forest regression model as the feature selection estimator, Figure 3d shows the ranking of feature importance obtained through iterative calculations. Among them, shrubland, wetland, bareland, and

permanent snow and ice exhibit the lowest feature importance and may be considered for removal during modeling.

Considering the results from various feature selection methods, the study decides to remove the features shrubland, wetland, bareland, and permanent snow and ice from the land cover data. The remaining 27 features are selected as auxiliary data for spatializing population modeling.

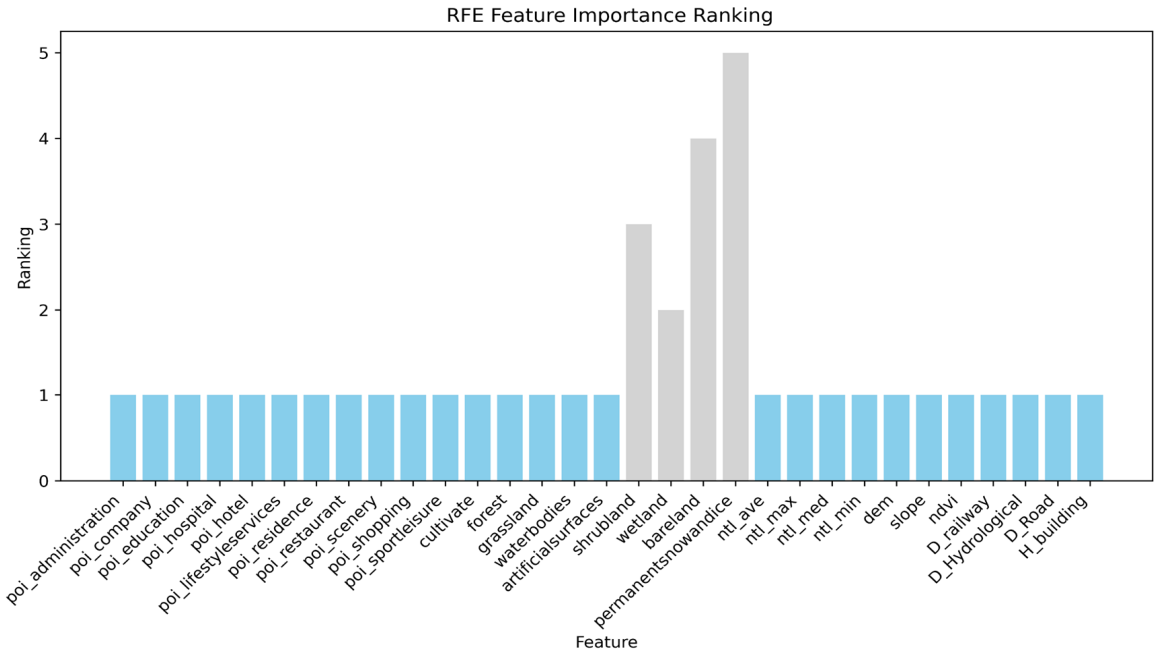


(**a**)



(**b**) (**c**)

Figure 3. *Cont.*



(**d**)

Figure 3. The result of different feature selection methods: (**a**) feature selection results based on Pearson correlation coefficient; (**b**) feature selection results based on variance threshold method; (**c**) feature selection results based on ReliefF; (**d**) feature selection results based on recursive feature elimination. In (**a**), $p < 0.001$ is represented as ***, $p < 0.01$ is represented as **, and $0.01 < p < 0.05$ is represented as *.

4.2. Parameter Optimization Results

This article combines the cross-validation method to improve the performance and generalization ability of the model, and compares and analyzes the effects of the above six different parameter adjustment methods on the parameter optimization of the random forest regression model. The following figures illustrate the process of combining the six parameter optimization methods with five-fold cross-validation to determine the optimal parameters. The color aggregation of parameter values in the parallel coordinate plot reflects the tendencies in constructing the score of the random forest model during modeling. It can be observed that n_estimators, max_depth, max_features, and min_samples_split do not exhibit obvious regulations or trends within their specified value ranges. The colors of the connecting lines between these parameters are evenly distributed, indicating that these four parameters do not display any specific patterns in the optimization of the random forest regression model. As for min_samples_leaf, when the parameter is less than 3, the color of the lines approaches yellow, indicating that the constructed random forest regression model exhibits better fitting performance in this range of parameter values.

In the parallel coordinate plot, lines connecting model parameters that are closer to yellow indicate higher R2_score values, suggesting better fitting performance of models composed of those parameter combinations. Lines with colors closer to deep green indicate lower fitting scores, indicating poorer model fitting. When the color approaches dark blue, the fitting score of the model is negative, indicating that the model constructed with this set of parameters has no practical significance. The overall color presentation of the figures demonstrates the efficiency of the corresponding parameter optimization methods.

Specifically, for tuning methods based on random search, grid search, and simulated annealing, as shown in Figures 4–6 the lines exhibit a predominantly green color, with numerous and dispersed patterns. This indicates that the majority of models resulting from these tuning methods have poor fitting performance and lower tuning efficiency. Additionally, for tuning methods based on genetic algorithms and Bayesian optimization,

as depicted in Figures 7–9, the lines have an overall yellowish color, indicating better fitting performance and a higher efficiency in parameter optimization. This is conducive to selecting parameter combinations that yield more effective modeling results.

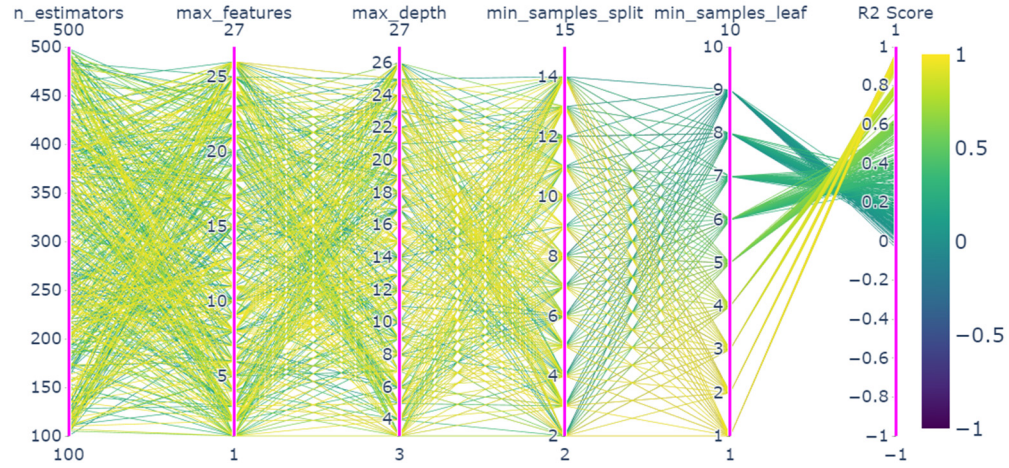


Figure 4. The parallel coordinate plot of random search.

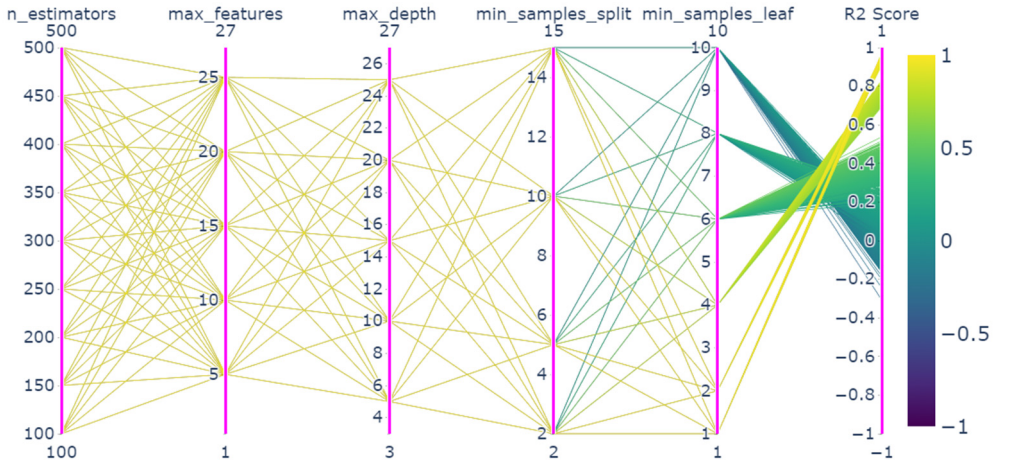


Figure 5. The parallel coordinate plot of grid search.

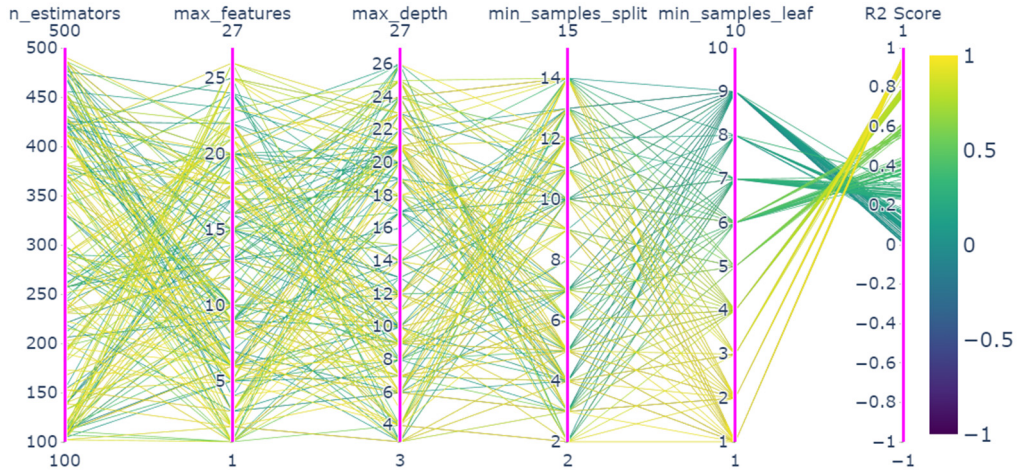


Figure 6. The parallel coordinate plot of simulated annealing algorithm.

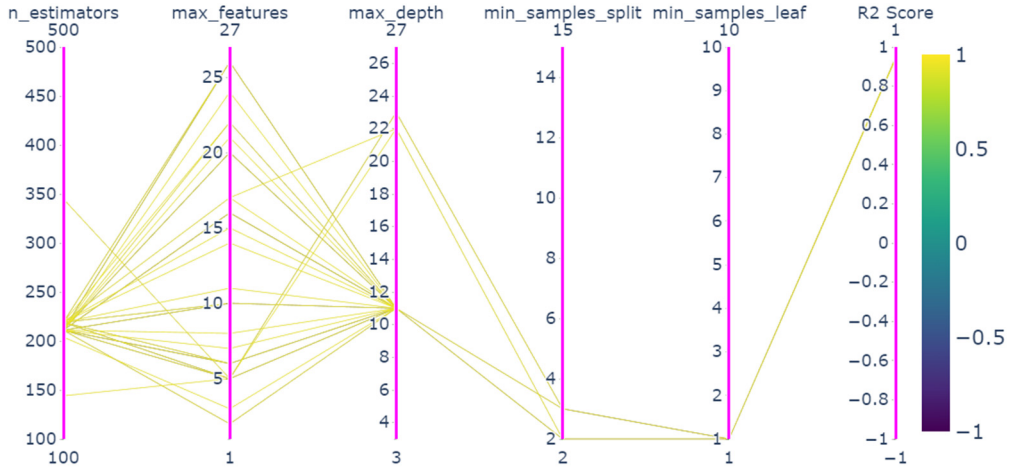


Figure 7. The parallel coordinate plot of genetic algorithm.

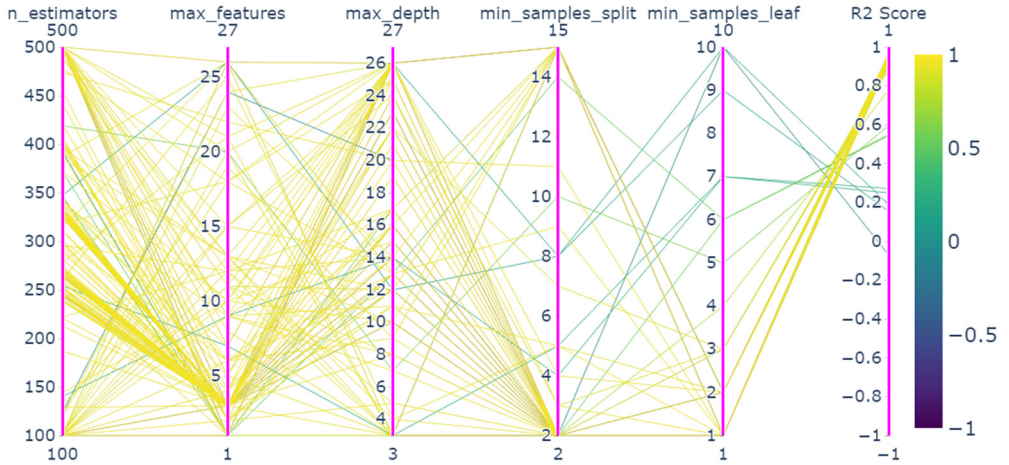


Figure 8. The parallel coordinate plot of Bayesian optimization based on Gaussian process.

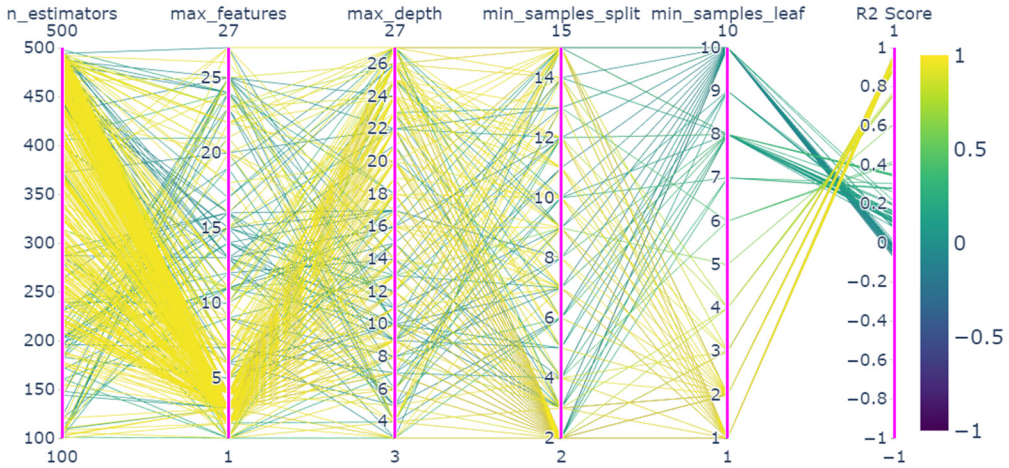


Figure 9. The parallel coordinate plot of Bayesian optimization based on gradient boosting regression trees.

The optimal parameter results obtained by the six parameter tuning algorithms for random forest models on the training set are summarized in Table 5.

Table 5. Results of different parameter optimization methods on the training set.

	n_Estimators	Max_Features	Max_Depth	Min_Samples_Split	Min_Samples_Leaf	R2_Score
Random search	267	3	14	2	1	0.9690
Grid search	150	5	15	2	1	0.9707
Genetic algorithm	211	6	11	2	1	0.9444
Simulated annealing algorithm	153	3	21	4	1	0.9619
Bayesian optimization based on Gaussian process	100	3	23	2	1	0.9699
Bayesian optimization based on gradient boosting regression trees	102	3	17	2	1	0.9701

4.3. Model Accuracy Validation

This paper utilizes the remaining 20% of the sample data as a validation set and employs Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) as evaluation metrics to assess the accuracy of the optimal models constructed using different tuning methods. Specifically, RMSE is commonly used to measure the proximity between the predicted values and the actual values. A smaller RMSE value indicates more accurate predictions by the model. MAE is the average of the absolute differences between predicted values and actual values, commonly used to measure the average magnitude of errors in model predictions. R^2 can be used to measure how well a model explains the target variable. Its values range from 0 to 1. The closer it is to 1, the better the explanation degree of the model. The formula of each indicator is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{8}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{9}$$

In Formulas (7)–(9), y_i is the true value, $\hat{y}_i$ is the predicted value, $\overline{y}$ is the average of the true values, and n is the total number of samples.

The results are summarized in Table 6, revealing that the model obtained through Bayesian optimization algorithm based on Gaussian processes exhibits the highest R^2 and the lowest RMSE, with values of 0.936 and 116.034 million people per square kilometer, respectively. Meanwhile, the Bayesian optimization algorithm based on gradient boosting regression trees achieves the lowest MAE; its value is 76.403 million people per square kilometer.

Table 6. Accuracy validation results of RF model based on different optimization methods.

Tuning Method	R^2	MAE	RMSE
Random search	0.9209	84.6478	129.2998
Grid search	0.9151	85.6674	133.9270
Genetic algorithm	0.9027	93.4762	143.3880
simulated annealing algorithm	0.9295	80.8391	122.0234
Bayesian optimization based on Gaussian process	0.9363	76.4410	116.0342
Bayesian optimization based on gradient boosting regression trees	0.9361	76.4304	116.1700

In order to analyze the accuracy of the models and the distribution of errors, we performed fitting regression analysis and error statistics on the model prediction results and the real values. The outcomes are depicted as scatter plots of predicted vs. actual values and histograms of errors distribution, as shown in Figures 10–15. In the scatter plots,

blue points represent the errors between predicted and actual values, with darker colors indicating smaller errors. The blue solid line is the regression line fitted through points composed of predicted and actual values, and the red dashed line is the ideal regression line when predicted and actual values are equal. The deviation between the blue and red lines signifies the discrepancy between the predictive model and actual values.

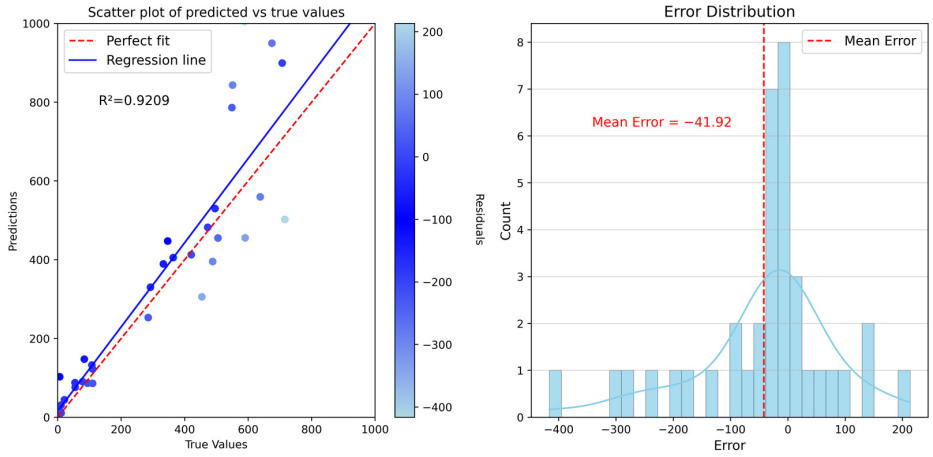


Figure 10. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Random search.

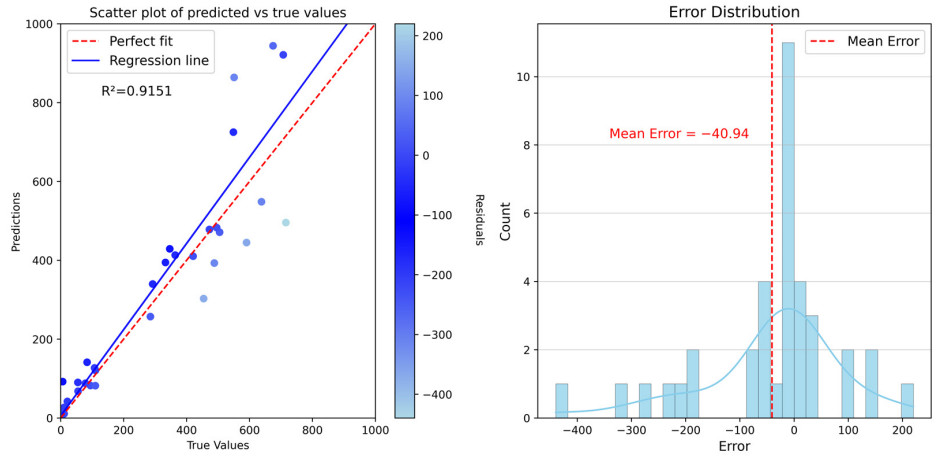


Figure 11. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Grid search.

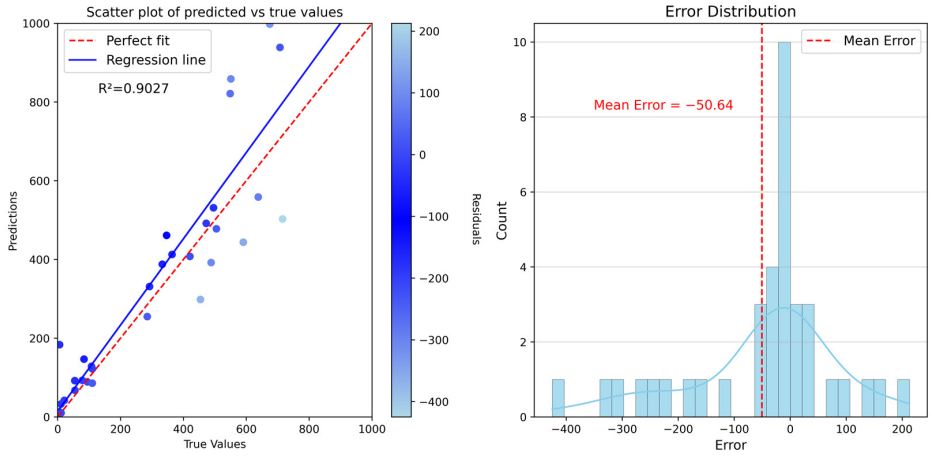


Figure 12. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Genetic algorithm.

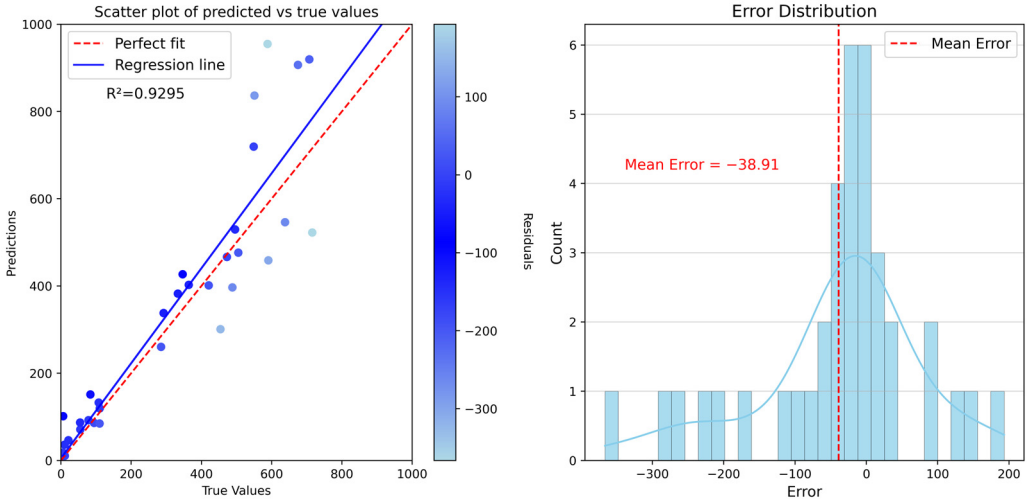


Figure 13. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Simulated annealing algorithm.

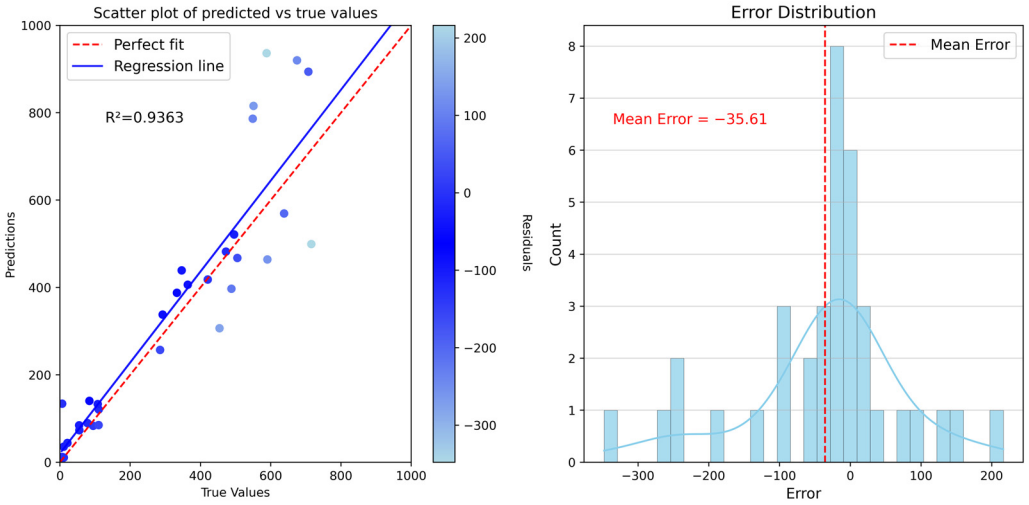


Figure 14. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Bayesian optimization based on Gaussian process.

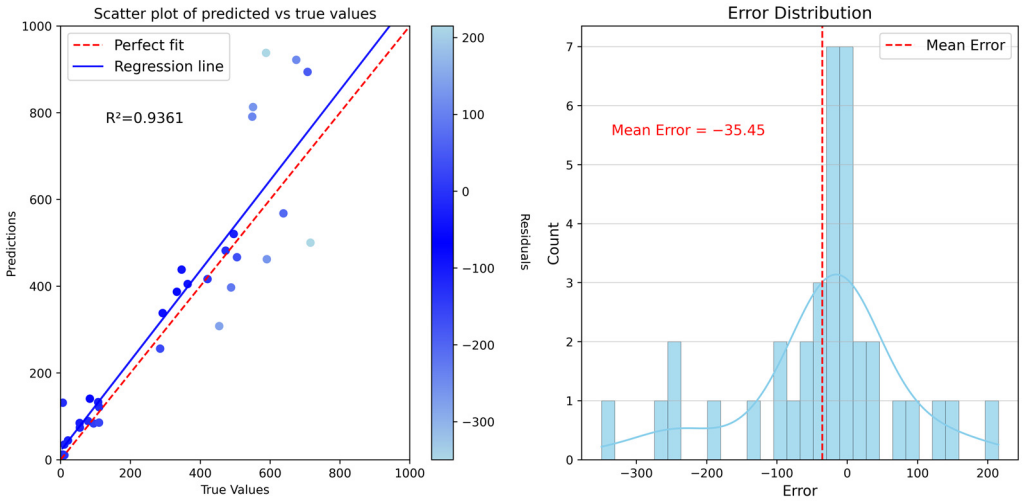


Figure 15. The prediction accuracy scatter diagram (**left**) and error distribution histogram (**right**) of RF model optimized with Bayesian optimization based on gradient boosting regression trees.

From Figures 10–15, it can be observed that, in the low-density region, the predictions of population density by all six methods are relatively close to the true values compared to the high-density region, with smaller absolute errors. Overall, the scatter points in the graphs are evenly distributed on both sides of the line, indicating the reliability of the model's accuracy. From the regression fitted line, it can be observed that the blue solid line and the red dashed line obtained from Bayesian optimization methods (Figures 14 and 15) are closer to the other methods. This indicates that these two Bayesian optimization methods have smaller deviations from the true values, suggesting better predictive capabilities.

In the error distribution histograms, the horizontal axis displays the range of model errors, where a larger numerical range indicates a greater deviation in model prediction. The vertical axis represents the number of the prediction error. The majority of the errors of the model predicted based on the six mentioned optimization methods are concentrated around zero. It indicates that random forest models have good prediction accuracy and stable prediction ability. From Figures 10–12, it can be seen that the models optimized with random search, grid search, and genetic algorithms have a wider prediction error range. The model optimized with these methods have relatively unstable prediction capabilities, and there may be large prediction errors in individual counties. From Figures 13–15, it can be observed that the models optimized with simulated annealing, Bayesian optimization based on Gaussian process, and Bayesian optimization based on gradient boosting regression trees have narrower horizontal ranges, implying smaller prediction error ranges and more stable prediction capabilities.

In conclusion, models constructed using parameters from the Bayesian optimization method exhibit higher coefficients of determination and smaller average errors, indicating better fitting performance and accuracy. This indicates that the Bayesian optimization algorithm is superior to commonly used methods for selecting random forest modeling parameters.

4.4. Population Spatialization Result Validation and Error Analysis

In this study, we use the random forest regression model obtained with the Bayesian optimization based on gradient boosting regression to spatialize the population of Sichuan Province in 2020, and obtain a 1 km grid population distribution map of Sichuan Province (Figure 16). From the map, it is evident that the population in Sichuan Province is primarily concentrated in the eastern Chengdu Plain and various urban centers. There is a noticeable trend of population density decreasing gradually from the urban centers towards the surrounding areas. In the western Qinghai-Tibet Plateau region, the population is sparse, with only a few areas having small populations. This distribution aligns closely with the actual population distribution in Sichuan Province. This indicates that the population spatialization model proposed in this article is effective.

This study's population spatialization model was trained using county-level administrative population census data. In order to validate the accuracy of population distribution at smaller spatial scales after gridification, we selected the seventh national population census data at the township level, which is smaller than the county-level administrative division, as the validation dataset. We also conducted an analysis of areas with significant errors to provide insights for future research. Due to limited access to township-level census data, we collected population data for 2122 townships and streets. In areas where population census data was missing, we used gridded population data as a base map. Then, we count the populations of the 2122 townships based on the RF-BOA Population Map, and compare them with the data collected from the seventh national population census. The evaluation of the results was conducted using the Relative Error (RE), which was calculated using Formulas (10) and (11) below. The error distribution chart was also created, as depicted in Figure 17.

$$E = Pop_{pred} - Pop_{true}, \tag{10}$$

$$RE = \frac{E}{Pop_{true}} \tag{11}$$

In Formulas (10) and (11), Pop_{pred} is the predicted population of townships obtained in this study, Pop_{true} corresponds to the actual population of townships obtained from the population census, and E is the difference between the predicted population and the actual population.

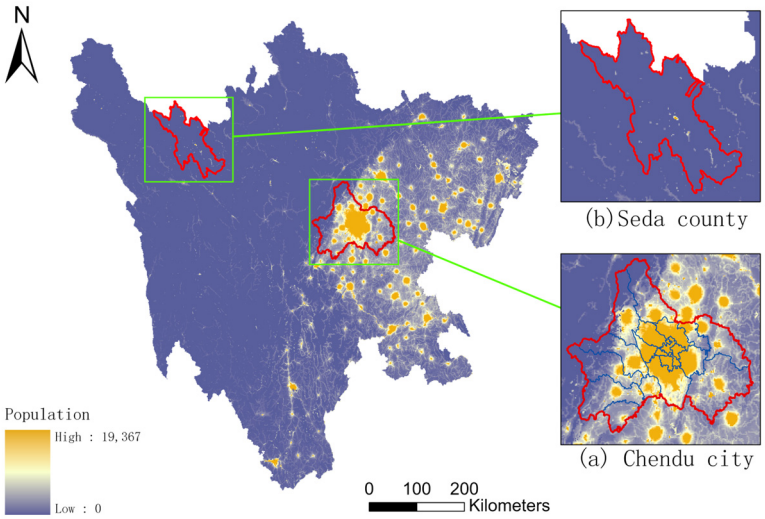


Figure 16. The population distribution map with 1 km resolution of Sichuan Province (RF-BOA Population Map).

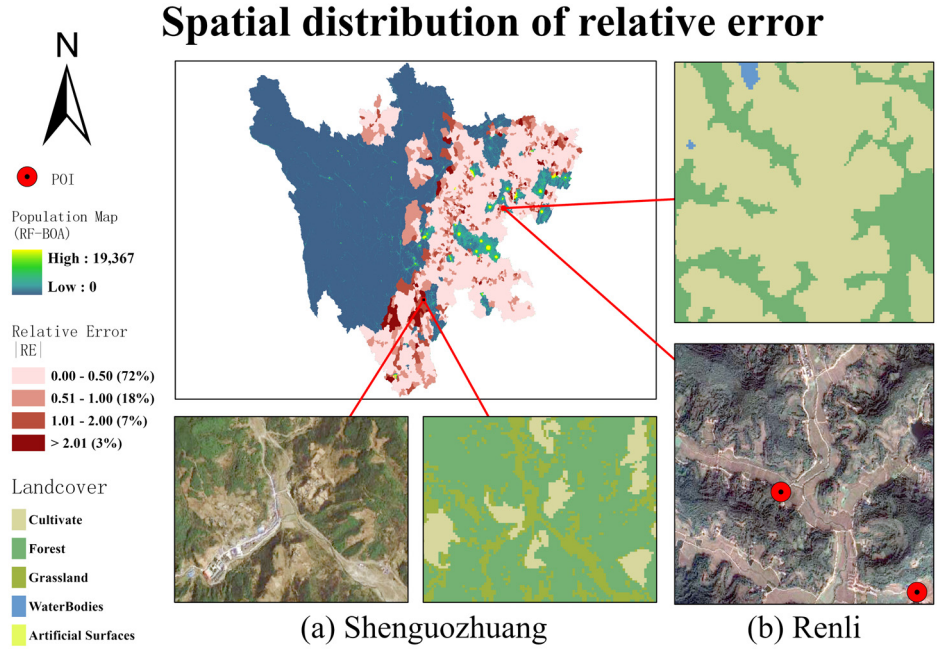


Figure 17. Spatial analysis of relative error.

Figure 17 illustrates the spatial distribution of relative errors in the population dataset for Sichuan Province created in this study. As depicted in the figure, the relative errors in the population distribution of around 70% of townships and streets are relatively small, while approximately 3% of townships and streets exhibit larger errors. These larger errors are scattered across the region and do not exhibit a specific spatial pattern. This phenomenon suggests that the gridding model proposed in this paper has a strong spatialization capability at the township level, accurately capturing population distribution. The model is effective in fine-grained population distribution.

To further investigate the causes of prediction errors in the model, the article selected two townships (Shenguozhuang Township and Renli Town) with the highest relative errors as validation areas. The relative error of population distribution in townships and streets in the verification area was analyzed combined with land cover data, satellite images of Google Maps in 2020, and POI data. Figure 17a,b show that both townships have a predominant land cover of shrubland, forests, and cultivated land, with a high vegetation cover. The satellite imagery reveals minimal human activities, such as building structures, road networks, transportation, and public facilities. Only Renli Town has a few POI data points. In summary, in the areas with larger prediction discrepancies, there is a high level of vegetation cover, sparse identifiable housing from Google imagery, and a lack of information on transportation networks and Points of Interest. Additionally, the nightlight data may have limited indicative value for rural areas. Therefore, predicting population in rural areas is more challenging compared to regions with more distinct and comprehensive features. In future research, supplementary data from various sources should be explored to obtain finer-scale population indicators for rural areas, allowing for a multi-dimensional and multi-scale characterization of rural populations.

5. Discussion

5.1. Feature Importance Analysis

Feature importance analysis can be used to explain the predictions of machine learning models and provides a deeper understanding, helping to comprehend the importance of features in the model, enhancing its interpretability and trustworthiness. SHAP (SHapley Additive exPlanations) is a method for explaining machine learning model predictions based on the concept of cooperative game theory [57]. It is used to analyze the impact of each feature on the model's predictions. The main idea behind SHAP feature importance analysis is to explain the model's output based on the contributions of each feature, and the allocation of contributions is influenced by the interactions between features. The SHAP values take into account the impact of a feature on the model output when combined with other features, providing a fair way to allocate the contribution of each feature to the model predictions. For each sample and each feature, SHAP values can explain how much influence that feature has on the prediction output for that sample [58]. A positive value indicates an increase in the predicted value, while a negative value indicates a decrease. The SHAP feature importance estimates for a random forest regression model obtained through Bayesian optimization with gradient boosting trees are shown in Figure 18.

From Figure 18, it can be observed that the SHAP importance of data related to POI, such as shopping centers, hospitals, schools, companies, government institutions, as well as data on life services, building heights, median nighttime lighting, and artificial land surface in land cover, all rank in the top ten for the random forest regression model's feature importance. This indicates that these mentioned features play an important role in revealing spatial population distribution. Among these, POI data holds a particularly high percentage, and these points of interest are closely related to people's daily lives. As urban development progresses, people's daily life needs are increasing, and the development of services, daily activities, and residential areas is closely tied to the spatial distribution of the population. Therefore, the distribution of geographic points of interest largely determines the spatial distribution of the population. Regarding building height data, with each grid having the same area size, the pixel values representing building heights determine the volume capacity of buildings, reflecting their ability to accommodate people per unit area. Places where the building capacity is larger tend to have larger populations, so building height is a good indicator of population distribution. Nighttime light data reflects human economic activity during the night. Economically developed areas usually host larger populations, which also proves the ability of nighttime light data to indicate population distribution. Human-made surface of land cover represents surfaces formed through human activities, including various human settlements, industrial areas, transportation

facilities, etc. These are closely tied to human activities and tend to better reflect human population than other types of land cover.

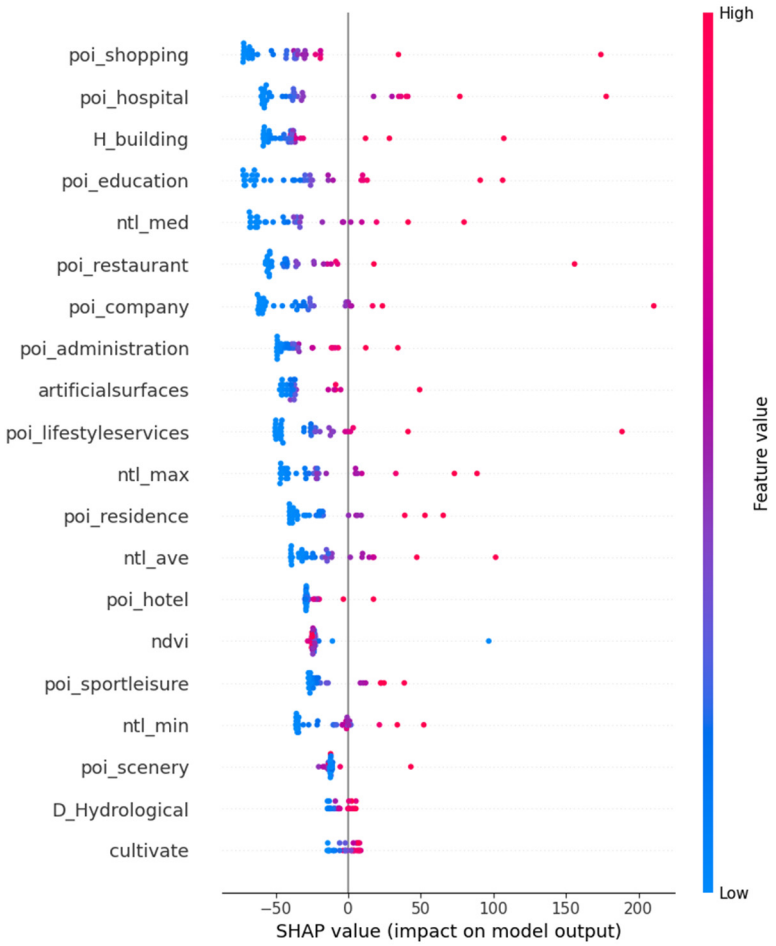


Figure 18. The SHAP feature importance of random forest model.

As urban development progresses, the need for daily life amenities increases, and the construction of life services is closely related to the spatial distribution of population. In Sichuan Province, the administrative units consistently influence the distribution of urban centers. Therefore, the shopping centers and administrative institutions in geographic POI to some extent determine the spatial distribution of the population. However, features like water bodies, grassland and forest of land cover, elevation and slope features of terrain, and railway density and road density features of geographic basic information exhibit the lowest importance in modeling. This shows that these factors have weak ability to indicate population information in Sichuan Province. Although Sichuan Province has complex topography, its natural geographical variations are quite distinct. The majority of the region experiences a warm and humid climate, with relatively minor differences in infrastructure development. This may result in features such as water bodies, farmland, terrain elements, and transportation networks having a limited impact on population distribution. Therefore, during the modeling process, these factors did not exhibit strong indicative roles.

In conclusion, features related to Points of Interest, nighttime light, and socioeconomic factors such as building height have a more significant impact on the spatial distribution of population compared to environmental features like terrain. These features more accurately reflect human activities and presence.

5.2. Comparison with Other Datasets

The WorldPop dataset, LandScan dataset, and Gridded Population of the World (GPW) dataset are widely used global population datasets. The WorldPop dataset is developed

by the University of Florida, based on population census data, nighttime light data, and land use data, and employs the random forest model for population spatialization. The LandScan dataset is developed by Oak Ridge National Laboratory and primarily based on population census data, remote sensing imagery, and geographic information data. This dataset mainly employs the Dasymetric method to estimate population distribution. The GPW dataset is developed by the Center for International Earth Science Information Network (CIESIN) at Columbia University. It spatializes global population statistical data using areal weighting methods. We assess the dataset generated in this study, comparing it with WorldPop dataset, LandScan dataset and GPW dataset at an equivalent resolution. We compared the 2020 Sichuan Province population data generated in this article with the above-mentioned data sets under the same resolution conditions.

The accuracy evaluation results are presented in Table 7. It can be observed that the RF-BOA model achieves the highest coefficient of determination (R^2) fitted with the true values, reaching a value of 0.6628. It surpasses the R^2 of the other datasets. Additionally, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are smaller for RF-BOA compared to LandScan, GPW, and WorldPop, with MAE and RMSE values of 12,459 people per township and 25,037 people per township, respectively.

Table 7. Accuracy assessment.

	R^2	MAE	RMSE
LandScan	0.3915	17,805.7163	33,634.8154
GPW	0.5384	13,401.9731	29,293.9095
WorldPop	0.6094	12,707.6707	26,947.1740
RF-BOA	0.6628	12,459.3926	25,037.1139

6. Conclusions

Population data, as a fundamental dataset capable of reflecting the socio-economic conditions of a region, can provide important information for regional resource allocation, disaster prevention, and emergency management. With the proliferation of grid-based research and applications, traditional methods are no longer sufficient to meet the spatial data accuracy requirements in this field. Therefore, it is necessary to spatially grid population statistics at the county, city, and national levels to obtain population distribution information at the kilometer grid level or even higher spatial resolutions.

In this article, we explored and conducted research on population spatialization from two aspects: data enhancement and model parameter optimization. In terms of data, we incorporated innovative social perception data, such as Points of Interest (POI), and high-precision building height data as auxiliary data. Additionally, we further processed and corrected data, such as nighttime light data, to enhance the quality of auxiliary data. Meanwhile, the feature importance analysis results indicated that high-precision socio-economic data plays a crucial role in the population spatialization.

In terms of model construction, numerous researchers have explored various methods, but there has been relatively little research on model parameter tuning. Hence, this article placed a special focus on the optimization of model parameters. To ensure model interpretability, we chose the popular Random Forest algorithm as the base model and investigated the impact of six different parameter optimization methods on model performance. The results demonstrated that Bayesian optimization algorithms significantly improved model parameter optimization, effectively enhancing model accuracy.

In the end, this article takes Sichuan Province as an example and constructs a spatial population distribution dataset at 1 km resolution based on the best parameters. It has been validated against township-level population census data and is found to closely match real-world conditions, demonstrating the effectiveness of the model. When compared to other international open-source datasets, the dataset created in this article better aligns

with the actual population distribution at the township level, exhibiting higher accuracy and showcasing the advantages of the modeling approach.

In conclusion, the population spatialization model constructed using the method proposed in this paper effectively predicts population distribution. It can provide detailed population spatialization information for urban planning and natural disaster prevention and control, enabling more rational resource allocation and efficient disaster response.

Author Contributions: Conceptualization, Y.C. and S.W.; Methodology, Y.C. and S.W.; Validation, Y.C. and S.W.; Formal analysis, Y.C. and S.W.; Writing—original draft, Y.C.; Writing—review and editing, S.W., Z.G. and F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "National Natural Science Foundation of China, grant number 42271090", "Fundamental Research Funds of the Institute of Earthquake Forecasting, China Earthquake Administration, grant number CEAIEF2022050504 and CEAIEF20230202", "National High-Resolution Earth Observation Major Project, grant number 31-Y30F09-9001-20/22".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

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

References

1. Bird, J.; Lebrand, M.; Venables, A.J. The belt and road initiative: Reshaping economic geography in Central Asia? *J. Dev. Econ.* **2020**, *144*, 102441. [CrossRef]
2. Andrade-Pacheco, R.; Savory, D.J.; Midekisa, A. Household electricity access in Africa (2000–2013): Closing information gaps with model-based geostatistics. *PLoS ONE* **2019**, *14*, e0214635. [CrossRef]
3. Tusting, L.S.; Bisanzio, D.; Alabaster, G.; Cameron, E.; Cibulskis, R.; Davies, M.; Flaxman, S.; Gibson, H.; Knudsen, J.; Mbogo, C.; et al. Mapping changes in housing in sub-Saharan Africa from 2000 to 2015. *Nature* **2019**, *568*, 391–394. [CrossRef] [PubMed]
4. Eales, A.; Alsop, A.; Frame, D.; Strachan, S.; Galloway, S. Assessing the market for solar photovoltaic (PV) microgrids in Malawi. *Hapres J. Sustain. Res.* **2020**, *2*, e200008.
5. Melchiorri, M.; Pesaresi, M.; Florczyk, A.J.; Corbane, C.; Kemper, T. Principles and applications of the global human settlement layer as baseline for the land use efficiency indicator—SDG 11.3. 1. *ISPRS Int. J. Geoinf.* **2019**, *8*, 96. [CrossRef]
6. Ehrlich, D.; Melchiorri, M.; Florczyk, A.J.; Pesaresi, M.; Kemper, T.; Corbane, C.; Corbane, C.; Freire, S.; Schiavina, M.; Siragusa, A. Remote sensing derived built-up area and population density to quantify global exposure to five natural hazards over time. *Remote Sens.* **2018**, *10*, 1378. [CrossRef]
7. Dasgupta, S.; Laplante, B.; Murray, S.; Wheeler, D. Exposure of developing countries to sea-level rise and storm surges. *Clim. Chang.* **2011**, *106*, 567–579. [CrossRef]
8. 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]
9. Azar, D.; Engstrom, R.; Graesser, J.; Comenetz, J. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sens. Environ.* **2013**, *130*, 219–232. [CrossRef]
10. Lai, S.; Bogoch, I.I.; Ruktanonchai, N.W.; Watts, A.; Lu, X.; Yang, W.; Yu, H.; Khan, K.; Tatem, A.J. Assessing spread risk of COVID-19 within and beyond China in early 2020. *Data Sci. Manag.* **2022**, *5*, 212–218. [CrossRef]
11. Thomson, D.R.; Linard, C.; Vanhuysse, S.; Steele, J.E.; Shimoni, M.; Siri, J.; José Siri, M.; Caiaffa, W.T.; Rosenberg, M.; Wolff, E.; et al. Extending data for urban health decision-making: A menu of new and potential neighborhood-level health determinants datasets in LMICs. *J. Urban. Health* **2019**, *96*, 514–536. [CrossRef] [PubMed]
12. James, W.H.; Tejedor-Garavito, N.; Hanspal, S.E.; Campbell-Sutton, A.; Hornby, G.M.; Pezzulo, C.; Nilsen, K.; Sorichetta, A.; Ruktanonchai, C.W.; Carioli, A.; et al. Gridded birth and pregnancy datasets for Africa, Latin America and the Caribbean. *Sci. Data* **2018**, *5*, 180090. [CrossRef] [PubMed]
13. Cai, Q.; Rushton, G.; Bhaduri, B.; Bright, E.; Coleman, P. Estimating small-area populations by age and sex using spatial interpolation and statistical inference methods. *Trans. GIS* **2006**, *10*, 577–598. [CrossRef]
14. Goodchild, M.F.; Anselin, L.; Deichmann, U. A framework for the areal interpolation of socioeconomic data. *Environ. Plan. A* **1993**, *25*, 383–397. [CrossRef]
15. Xie, Z. A framework for interpolating the population surface at the residential-housing-unit level. *GIsci Remote Sens.* **2006**, *43*, 233–251. [CrossRef]
16. Jin, Y.; Liu, R.; Fan, H.; Li, P.; Liu, Y.; Jia, Y. Multi-Resolution Population Mapping Based on a Stepwise Downscaling Approach Using Multisource Data. *Remote Sens.* **2023**, *15*, 1947. [CrossRef]

17. Guo, H.; Zhu, W. A review on the spatial disaggregation of socioeconomic statistical data. *Acta Geogr. Sin.* **2022**, *77*, 2650–2667.
18. Zeng, C.; Zhou, Y.; Wang, S.; Yan, F.; Zhao, Q. Population spatialization in China based on night-time imagery and land use data. *Int. J. Remote Sens.* **2011**, *32*, 9599–9620. [CrossRef]
19. Lo, C.P. Population estimation using geographically weighted regression. *GIsci Remote Sens.* **2008**, *45*, 131–148. [CrossRef]
20. Huang, Y.; Zhao, C.; Song, X.; Chen, J.; Li, Z. A semi-parametric geographically weighted (S-GWR) approach for modeling spatial distribution of population. *Ecol. Indic.* **2018**, *85*, 1022–1029. [CrossRef]
21. Chi, G.; Zhu, J. Spatial regression models for demographic analysis. *Popul. Res. Policy Rev.* **2008**, *27*, 17–42. [CrossRef]
22. Liu, X.H.; Kyriakidis, P.C.; Goodchild, M.F. Population-density estimation using regression and area-to-point residual kriging. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 431–447. [CrossRef]
23. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
24. Folberth, C.; Baklanov, A.; Balkovič, J.; Skalský, R.; Khabarov, N.; Obersteiner, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. For. Meteorol.* **2019**, *264*, 1–15. [CrossRef]
25. Zhao, X.; Xia, N.; Xu, Y.; Huang, X.; Li, M. Mapping population distribution based on XGBoost using multisource data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11567–11580. [CrossRef]
26. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
27. Qiu, G.; Bao, Y.; Yang, X.; Wang, C.; Ye, T.; Stein, A.; Jia, P. Local population mapping using a random forest model based on remote and social sensing data: A case study in Zhengzhou, China. *Remote Sens.* **2020**, *12*, 1618. [CrossRef]
28. 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]
29. Li, K.; Chen, Y.; Li, Y. The random forest-based method of fine-resolution population spatialization by using the international space station nighttime photography and social sensing data. *Remote Sens.* **2018**, *10*, 1650. [CrossRef]
30. 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]
31. Liu, L.; Cheng, G.; Yang, J.; Cheng, Y. Population Spatialization in Zhengzhou City Based on Multi-source Data and Random Forest Model. *Front. Earth Sci.* **2023**, *11*, 1092664. [CrossRef]
32. He, M.; Xu, Y.; Li, N. Population spatialization in Beijing city based on machine learning and multisource remote sensing data. *Remote Sens.* **2020**, *12*, 1910. [CrossRef]
33. Doupe, P.; Bruzelius, E.; Faghmous, J.; Ruchman, S.G. Equitable development through deep learning: The case of sub-national population density estimation. In Proceedings of the 7th Annual Symposium on Computing for Development, Nairobi, Kenya, 18–20 November 2016; pp. 1–10.
34. Xing, X.; Huang, Z.; Cheng, X.; Zhu, D.; Kang, C.; Zhang, F.; Liu, Y. Mapping human activity volumes through remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5652–5668. [CrossRef]
35. 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]
36. Da Costa, J.N.; Bielecka, E.; Calka, B. Uncertainty quantification of the global rural-urban mapping project over Polish census data. In *Environmental Engineering, Proceedings of the International Conference on Environmental Engineering, Vilnius, Lithuania, 27–28 April 2017*; ICEE, Vilnius Gediminas Technical University, Department of Construction Economics & Property: Vilnius, Lithuania; Volume 10, pp. 1–7.
37. Dobson, J.E.; Bright, E.A.; Coleman, P.R.; Durfee, R.C.; Worley, B.A. LandScan: A global population database for estimating populations at risk. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 849–857.
38. Tatem, A.J. WorldPop, open data for spatial demography. *Sci. Data* **2017**, *4*, 170004. [CrossRef] [PubMed]
39. Zhou, Y.; Ma, M.; Shi, K.; Peng, Z. Estimating and interpreting fine-scale gridded population using random forest regression and multisource data. *ISPRS Int. J. Geoinf.* **2020**, *9*, 369. [CrossRef]
40. Gunasekera, R.; Ishizawa, O.; Aubrecht, C.; Blankespoor, B.; Murray, S.; Pomonis, A.; Daniell, J. Developing an adaptive global exposure model to support the generation of country disaster risk profiles. *Earth Sci. Rev.* **2015**, *150*, 594–608. [CrossRef]
41. Sabesan, A.; Abercrombie, K.; Ganguly, A.R.; Bhaduri, B.; Bright, E.A.; Coleman, P.R. Metrics for the comparative analysis of geospatial datasets with applications to high-resolution grid-based population data. *GeoJournal* **2007**, *69*, 81–91. [CrossRef]
42. Bai, Z.; Wang, J.; Wang, M.; Gao, M.; Sun, J. Accuracy assessment of multi-source gridded population distribution datasets in China. *Sustainability* **2018**, *10*, 1363. [CrossRef]
43. Zhang, J.; Zhu, W.; Zhu, L.; Cui, Y.; He, S.; Ren, H. Topographical relief characteristics and its impact on population and economy: A case study of the mountainous area in western Henan, China. *J. Geogr. Sci.* **2019**, *29*, 598–612. [CrossRef]
44. Lu, D.; Tian, H.; Zhou, G.; Ge, H. Regional mapping of human settlements in southeastern China with multisensor remotely sensed data. *Remote Sens. Environ.* **2008**, *112*, 3668–3679. [CrossRef]
45. Amaral, S.; Câmara, G.; Monteiro, A.M.V.; Quintanilha, J.A.; Elvidge, C.D. Estimating population and energy consumption in Brazilian Amazonia using DMSP night-time satellite data. *Comput. Environ. Urban. Syst.* **2005**, *29*, 179–195. [CrossRef]
46. 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]

47. 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. [CrossRef]
48. Wu, W.B.; Ma, J.; Banzhaf, E.; Meadows, M.E.; Yu, Z.W.; Guo, F.X.; Sengupta, D.; Cai, 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]
49. Holben, B.N. Characteristics of maximum-value composite images from temporal AVHRR data. *Int. J. Remote Sens.* **1986**, *7*, 1417–1434. [CrossRef]
50. Li, X.; Xu, H.; Chen, X.; Li, C. Potential of NPP-VIIRS nighttime light imagery for modeling the regional economy of China. *Remote Sens.* **2013**, *5*, 3057–3081. [CrossRef]
51. Biswas, N.; Ali, M.M.; Rahaman, M.A.; Islam, M.; Mia, M.R.; Azam, S.; Ahmed, k.; e Moni, M.A. Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *Biomed. Res. Int.* **2023**, *2023*, 6864343. [CrossRef]
52. Mao, Z.; Han, H.; Zhang, H.; Ai, B. Population spatialization at building scale based on residential population index—A case study of Qingdao city. *PLoS ONE* **2022**, *17*, e0269100. [CrossRef]
53. Lu, P.; Ye, L.; Pei, M.; Zhao, Y.; Dai, B.; Li, Z. Short-term wind power forecasting based on meteorological feature extraction and optimization strategy. *Renew. Energy* **2022**, *184*, 642–661. [CrossRef]
54. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
55. Subbiah, S.S.; Chinnappan, J. Deep learning based short term load forecasting with hybrid feature selection. *Electric Pow. Syst. Res.* **2022**, *210*, 108065. [CrossRef]
56. Wakjira, T.G.; Ibrahim, M.; Ebead, U.; Alam, M.S. Explainable machine learning model and reliability analysis for flexural capacity prediction of RC beams strengthened in flexure with FRCM. *Eng. Struct.* **2022**, *255*, 113903. [CrossRef]
57. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *NeurIPS* **2017**, *30*, 4768–4777.
58. Meng, Y.; Yang, N.; Qian, Z.; Zhang, G. What makes an online review more helpful: An interpretation framework using XGBoost and SHAP values. *J. Theor. Appl. Electron. Commer. Res.* **2020**, *16*, 466–490. [CrossRef]