International Journal of Geo-Information

MDPI

Article

Population Density Prediction at Township Scale Supported by Machine Learning Method: A Case Study in Inner Mongolia

Chenxi Cui [1,2,3], Yunfeng Hu [2,3], Yuhai Bao [1,*] and Hao Li [2,3]

[1] College of Geographic Sciences, Inner Mongolia Normal University, Hohhot 010022, China; 20224016026@mails.imnu.edu.cn
[2] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; huyf@lreis.ac.cn (Y.H.); lihao8037@igsnrr.ac.cn (H.L.)
[3] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: baoyuhai@imnu.edu.cn

Abstract: With the acceleration in population migration and urbanization, accurate population density prediction has become increasingly important for regional planning and resource management. This study focuses on predicting population density at the township level in Inner Mongolia. By integrating multi-source data, such as nighttime light indices and road network density, various machine learning models—including random forest, XGBoost, and LightGBM—were employed to significantly improve prediction accuracy. Interpretable machine learning techniques were utilized to quantitatively analyze the contribution of various variables to population distribution. The results indicate that nighttime light indices and road network density are key influencing factors, revealing their complex nonlinear relationships with population density. This study provides new methodological support for predicting population density in Inner Mongolia and similar regions, demonstrating the potential of machine learning in regional population research. While machine learning models effectively capture correlations between variables, they do not reveal causal relationships. Future research should introduce more detailed data and causal inference models to deepen our understanding of population distribution and its influencing factors.

Keywords: Inner Mongolia; population density; machine learning; SHAP

Citation: Cui, C.; Hu, Y.; Bao, Y.; Li, H. Population Density Prediction at Township Scale Supported by Machine Learning Method: A Case Study in Inner Mongolia. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 426. https://doi.org/10.3390/ijgi13120426

Academic Editor: Wolfgang Kainz

Received: 18 October 2024
Revised: 18 November 2024
Accepted: 27 November 2024
Published: 29 November 2024

1. Introduction

In the context of increasing population migration and accelerating urbanization, accurate population density forecasting has become a critical research topic in regional planning, socio-economic development, and resource management [1]. High-precision population density predictions can assist decision-makers in effective resource allocation, infrastructure planning, and emergency management, thereby providing data support for sustainable development.

As the complexity of spatial population distribution continues to increase, researchers have developed various methods to address the challenges of population density modeling across different scales [2–6]. These methods range from large-scale global predictions to fine-scale modeling, covering a wide array of applications [7,8]. Traditionally, statistical models based on census data are a commonly used approach for population density forecasting. These models primarily rely on macro-level statistical data and employ spatial interpolation, regression analysis, or region-based weighting methods to distribute populations across various spatial units [9,10]. Such methods are widely applied in large-scale predictions, such as at the national and provincial levels, due to their simplicity and low computational cost. However, with evolving research demands, particularly in smaller-scale applications such as towns or communities, the limitations of traditional statistical models have become increasingly evident. These models often struggle to capture complex

spatial heterogeneity and interactions among multiple variables, resulting in inadequate predictive accuracy at finer scales. To address these limitations, researchers have begun to adopt population density forecasting methods based on multi-source data, including remote sensing data, land use data, and nighttime light data [11–13]. These high-resolution data sources offer more detailed spatial information about population distribution within regions. For example, remote sensing data, such as DMSP/OLS nighttime light data and VIIRS data, can provide valuable insights into population density patterns [14–17], capture the spatial variations of economic activities [18], and reflect population density in urban areas. By integrating these multi-source data, models can more accurately reflect the actual distribution of populations while better capturing significant factors such as urban-rural differences and geographical influences [18]. In recent years, forecasting methods based on remote sensing technology [19,20] have been widely applied in global and regional studies of population distribution, demonstrating significant potential in exploring correlations between population and economic activities.

With the increasing volume of spatial data, researchers have developed various modeling methods to facilitate the integration and processing of multi-source data [18,21]. The primary modeling approaches can be categorized into top-down, bottom-up, and hybrid methods. Top-down methods rely on macro-level statistical data, which are allocated to smaller spatial units according to specific rules. These methods, such as GPW [22] and LandScan [23], are widely used for global-scale population density forecasting and are particularly suited to large-scale regional estimations. However, these methods often struggle to effectively handle spatial heterogeneity at smaller scales. Bottom-up methods, in contrast, begin with local areas and gradually infer population distribution using detailed data, such as building distribution and land use. These methods demonstrate exceptional performance in population density forecasting at smaller scales, such as cities or communities. A notable example of this approach is the Global Human Settlement Layer (GHSL) project in Europe, which leverages high-resolution regional data to generate more accurate predictions. Hybrid methods combine the strengths of both top-down and bottom-up approaches, utilizing the extensive coverage of macro-level statistical data and the granularity of micro-level data to provide high-precision population forecasts across different scales. By integrating data from multiple levels, hybrid methods ensure data integrity at larger scales while maintaining predictive accuracy at smaller scales, making them suitable for a wide range of applications.

To further enhance the accuracy of population density forecasting, researchers have increasingly incorporated machine learning techniques in recent years [24,25]. Unlike traditional rule-based or linear assumption models, machine learning methods learn complex relationships between variables in a data-driven manner, enabling more effective handling of nonlinear features within multi-source data. Commonly used machine learning algorithms, such as Random Forest [26], XGBoost [27], and LightGBM [28], excel at addressing nonlinear relationships and intricate variable interactions, making them particularly suitable for fine-scale regional modeling. However, it is important to note that while machine learning models can reveal strong correlations between variables and population density, these relationships do not imply causation.

In this study, the application of machine learning techniques is primarily focused on improving the efficiency of population density predictions rather than uncovering causal mechanisms between variables. By integrating multi-source data, machine learning models can enhance predictive accuracy, particularly when handling complex nonlinear relationships. Additionally, with advancements in interpretable machine learning techniques, researchers can quantify the influence of each variable on population density predictions. This interpretability analysis helps researchers better understand the decision-making processes of the models. While these variables may not have causal relationships with population density, their correlations provide strong support for accurate predictions.

In summary, this study utilizes advanced machine learning algorithms, including Random Forest (RF), XGBoost, and LightGBM, in combination with multi-faceted variable

data such as nighttime lights and transportation networks, to construct a population density prediction model at the township scale in Inner Mongolia. By employing interpretable machine learning techniques, the contribution of each variable is quantified, revealing their nonlinear correlations with population density, thereby enhancing the model's transparency and credibility. The study aims to address the following questions:

1. What is the spatial distribution of the population at the township scale in Inner Mongolia?
2. How do various machine learning models perform in population forecasting?
3. How can interpretable machine learning models elucidate the relationships between different variables and population density, thereby enhancing model transparency?

2. Study Area, Data, and Methods

2.1. Overview of the Study Area

Inner Mongolia Province (Figure 1) is in northern China (between 97°11′–126°02′ E and 37°24′–53°23′ N) and forms part of the inland area of the Eurasian continent. As of 2023, the region comprises 12 prefecture-level administrative units, including 9 cities and 3 leagues, and encompasses 103 counties. Covering a total land area of approximately 1.183 million square kilometers, it accounts for 12.3% of China's total land area, making it the third-largest province or region in China.

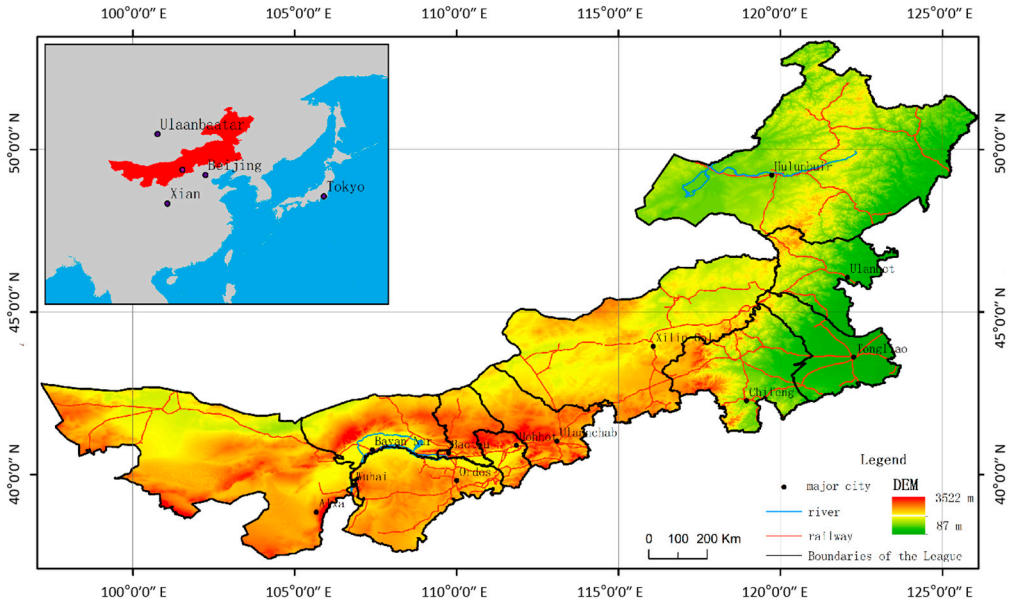

Figure 1. Location and topography of the study area.

Inner Mongolia extends diagonally from northeast to southwest, forming a narrow shape. It shares borders with Heilongjiang, Jilin, Liaoning, and Hebei to the east; Shanxi, Shaanxi, and Ningxia to the south; Gansu to the west; and Russia and Mongolia to the north. The region is characterized by relatively high elevation, with an average altitude of approximately 1000 m. The terrain is predominantly plateau and flatlands. The average annual temperature ranges from 2 °C to 14 °C, with annual precipitation averaging around 400 mm. The climate is primarily temperate continental, featuring cold winters and hot summers, with precipitation decreasing from east to west. Inner Mongolia features diverse vegetation types. From east to west, the primary land cover types include forest, arable land, grassland, and desert, with grasslands and deserts serving as the dominant ecological systems in the region.

In 2020, the permanent population of Inner Mongolia was 24.049 million, with Chifeng City having the highest population at 4.036 million, while Alxa League had the lowest population at 262,400. The gross domestic product (GDP) of Inner Mongolia in 2020

was 1.73598 trillion yuan, consisting of a primary industry value added of 202.51 billion yuan, a secondary industry value added of 686.80 billion yuan, and a tertiary industry value added of 846.67 billion yuan, reflecting a ratio of 11.7:39.6:48.8 among the three industries. The main industries include agricultural and livestock product processing, energy and chemical industries, metallurgy and building materials, tourism, and high-tech and service industries. Additionally, the region hosts the national key development urban agglomeration of "Hohhot-Baotou-Ordos-Yulin", encompassing Hohhot, Baotou, and Ordos in Inner Mongolia, as well as Yulin in Shanxi Province.

2.2. Indicators

This study employs population density (total population of the area/total area) as an indicator to characterize the population distribution in Inner Mongolia, treating it as the dependent variable. The independent variables comprise 18 elements across seven categories: economy, climate, vegetation cover, rivers, roads, topography, and land use. This selection aims to analyze the impact of natural, climatic, and socio-economic factors on population distribution.

Table 1 provides a brief description of the 18 variables highly correlated with population distribution considered in this study. These variables include fundamental independent elements (e.g., topography, precipitation) as well as associated elements that interact with other variables (e.g., road network density and nighttime light index). The independent elements are more intuitive in the model, making it easier to understand their contribution to population density predictions, as they directly reflect the characteristics of the natural environment or socio-economic conditions.

Table 1. Factors affecting population distribution.

Factor	Impact Factor	Descriptive
Economic factors (A)	Nighttime light index (A1)	The nighttime lighting index is widely used to characterize the development of cities and towns as an indicator that can assess the level of economic development, population density, and other urban development [29].
Climate (B)	Annual precipitation (B1) Annual temperature (B2)	Humans cannot survive without adequate temperatures and precipitation, and studies in larger regions have demonstrated the positive effects of precipitation and temperature on population distribution. The positive effect of precipitation and temperature on population distribution has been demonstrated in studies conducted in larger regions.
Vegetation cover (C)	NDVI (C)	In Inner Mongolia, where the ecosystem is sensitive and the vegetation is susceptible to human activities, NDVI, as an important indicator characterizing the growth of vegetation on the surface, has been widely used in the monitoring of vegetation dynamics. NDVI, as an important indicator characterizing the growth status of surface vegetation, is widely used in monitoring vegetation dynamics.
Rivers (D)	River network density (D1) Distance from river (D2)	Areas with many rivers not only have fertile soil and flat terrain, but also provide sufficient water for agricultural production, and rivers can effectively influence the density of population distribution. Rivers can effectively influence the density of population distribution.
Roads (E)	Road network density (E1) Distance from road (E2)	The road network, which is the backbone of the city, has a high correlation between road density and population and employment density. The road density is highly correlated with population and employment densities.

Table 1. *Cont.*

Factor	Impact Factor	Descriptive
Topography (F)	Slope (F)	As the most important factor among many geographic environmental factors, topography has a close relationship with the spatial distribution of population, and in the correlation analysis of topographic factors and population distribution, it is found that slope is negatively correlated with the population size and population density. In the analysis of the correlation between topographic factors and population distribution, it is found that slope is negatively correlated with population size and density [30].
Land use (G)	Cropland index (G1) Forest index (G2) Grassland index (G3) Shrub index (G4) Wetland index (G5) Water column index (G6) Man-made land surface index (G7) Bare ground index (G8)	Land, as a valuable asset for human production and life, affects human daily life, and land use data are not only the basis for ecological conservation research, land resource management, and regional sustainable development, but also a direct manifestation of human activities affecting nature.

However, associated elements may exhibit multicollinearity issues, resulting in strong interrelations when used together, which necessitates a deeper analysis to clarify their contributions to population forecasting. To address the multicollinearity among the variables, all the relevant factors were initially included in the analysis, followed by a Variance Inflation Factor (VIF) analysis to eliminate variables with strong multicollinearity. This approach helps to select key variables that are highly correlated with population forecasting and significantly improves the predictive accuracy of the model.

2.3. Basic Datasets and Data Preprocessing

Based on the original data format, the dataset is categorized into spatial and tabular data. The spatial data uniformly adopts the Albers equal-area conic projection coordinate system, with vector data granularity set at the township level and a spatial resolution of 1 km for raster data. Leveraging ArcGIS's connection function for public fields, such as place names or standardized codes, facilitates the association between attribute tables and spatial vector data. Further conversion from vector to raster allows for the transformation of various attribute data into spatial raster data.

Administrative division data at the township level were obtained from the Inner Mongolia Autonomous Region Science and Technology Information Institute (https://www.11467.com/huhehaote/co/4036.htm, accessed on 9 August 2019). Road data were derived from the 2020 National Urban Road Dataset (https://download.csdn.net/download/weixin_42153420/85474133, accessed on 9 August 2020), incorporating first- through fourth-level roads. River and water system data were extracted from China's rivers and lakes dataset (data.tpdc.ac.cn, accessed on 9 August 2021), which amalgamates first-, third-, fourth-, and fifth-level rivers. These datasets were then combined for further analysis.

The population data used in this study were sourced from the China 2020 Census Information by Townships, Towns, and Streets (Inner Mongolia Autonomous Region), published by the National Bureau of Statistics (https://www.stats.gov.cn/sj/, accessed on 11 May 2021). It is important to note that due to political district adjustments and data updates, the number of township units recorded in the population census table (1027) does not align with the number of township units in the administrative division spatial data mentioned earlier (1020). To address this discrepancy, the authors cross-referenced various datasets, including announcements related to administrative division adjustments. Generally, the list of townships in the census forms served as the primary reference point. Attribute verification and regional adjustments were carried out on the administrative

division spatial data to ensure alignment with the demographic forms. This meticulous process resulted in the creation of a spatialized dataset comprising 993 township units.

To assess the river and road factors, the distances from roads and rivers were initially determined using the Near Neighbor Analysis tool in ArcGIS 10.4 software. Subsequently, the river and road network densities were calculated by creating fishnet grids and applying the following formula:

$$R(D) = \frac{L}{A} \quad (1)$$

where R(D) denotes the river network density or road network density in km/km^2, respectively. R or D represent the river or road length in kilometers, and A is the area of the fishnet grid within the statistical area in km^2.

Economic data, including GDP, total industrial output value, and agricultural, forestry, animal husbandry, and fishery output values, were sourced from the Inner Mongolia Statistical Yearbook (2021). These statistical survey data are aggregated at the flag (county, district, and county-level city) level. To align with the previously mentioned population data, the authors employed the principle of "proximity" to allocate economic data from the flag-county level to each township. This approach facilitated the acquisition of GDP, gross industrial output value, and agricultural, forestry, animal husbandry, and fishery output value data at the township level, thereby establishing correlations with the respective township units.

Meteorological data were obtained from the National Tibetan Plateau Scientific Data Center (https://doi.org/10.11888/Hydro.tpdc.270302, accessed on 16 November 2019), with precipitation data sourced from the China 1 km resolution month-by-month dataset spanning 1901 to 2022. Specifically, precipitation data were extracted from China's monthly precipitation dataset at a 1 km resolution over the same period [31]. Additionally, temperature data were derived from China's month-by-month mean temperature dataset at a 1 km resolution, covering the period from 1901 to 2022 [32]. For this study, the 12-month precipitation and air temperature data for 2020 were averaged to calculate the annual values. Subsequently, the spatial aggregation of average precipitation totals and air temperatures within each township was conducted based on the established administrative division boundaries.

The Normalized Difference Vegetation Index (NDVI) data for 2020 were obtained from the MOD13A3 dataset available on the NASA website (https://ladsweb.modaps.eosdis.nasa.gov/, accessed on 9 August 2020), with a spatial resolution of 1 km [33]. This dataset, acquired by NASA's MODIS (Moderate Resolution Imaging Spectroradiometer) on Landsat, underwent multiple processing and calibration stages to ensure data quality and accuracy. For this study, image processing, projection, and format conversion were performed using MET software (1.3.0). Subsequently, mean NDVI values within each township were calculated using administrative division boundaries as reference points.

The Digital Elevation Model (DEM) data were sourced from the SRTM dataset available on the Geospatial Data Cloud website (https://www.gscloud.cn, accessed on 9 August 2003), with a spatial resolution of 90 m. Using ArcGIS, the Slope and Aspect tools were employed to derive slope and slope direction data. The Focal Statistics tool was then applied to calculate the maximum and minimum values of the DEM within a 4-square-kilometer spatial domain. These maximum and minimum DEM values were used to compute terrain relief parameters, defined as the difference between the maximum and minimum DEM values. Additionally, the mean topographic relief within each township was calculated based on the administrative division data.

Nighttime stabilized light intensity data for 2020 were acquired from the NOAA website's NPP/VIIRS Annual VNL V2 dataset (https://eogdata.mines.edu/products/vnl/, accessed on 9 August 2012) at a resolution of 1 km. The averages of nighttime light intensity within individual townships were then calculated using the established administrative division boundaries as reference points.

Land cover data were sourced from the Globeland30 (2020) global land cover product, available at a spatial resolution of 30 m via http://www.globallandcover.com, accessed on 9 August 2020). This product classifies land use into 10 categories: cropland, woodland, grassland, shrubland, wetland, water bodies, tundra, man-made surfaces, bare ground, glaciers, and permanent snow. Third-party evaluations indicate an overall accuracy of 83.5%, providing essential data for global land cover analysis. Using ArcGIS, the researchers calculated the area and proportion of each land type within a 1×1 km fishnet grid, thereby establishing indices for each land type (e.g., the cultivated land index). These calculations were linked to the corresponding fishnet grid. Subsequently, the fishnet grid data (vector) was converted to raster data, yielding spatial data (raster) such as a cropland index, forest land index, grassland index, shrubland index, wetland index, water body index, man-made surface index, and bare land index at a 1 km resolution. Building upon this foundation, the mean value of each land index within each township was determined based on the administrative division data.

2.4. Methodology

2.4.1. Multiple Linear Regression and Various Machine Learning Models

In this study, multiple linear regression and various machine learning models were employed to develop population density prediction models at the township scale. Multiple linear regression is a classical statistical method used to explore linear relationships between multiple independent variables and population density. It quantifies the linear impact of independent variables on population density through a linear equation. However, given the complexity and nonlinearity of population distribution, relying solely on linear regression may not adequately capture the intricate relationships among variables.

To enhance prediction accuracy, several machine learning models were also applied, including Random Forest (RF), XGBoost, and LightGBM. These machine learning algorithms learn complex nonlinear relationships between variables and population density in a data-driven manner. Random Forest improves prediction accuracy by constructing multiple decision trees and averaging their results, offering strong robustness and noise resistance. XGBoost and LightGBM are optimized algorithms based on Gradient Boosting Decision Trees (GBDTs) that iteratively optimize model parameters, enhancing the model's predictive power. Compared to linear regression, these machine learning models are better suited to handle complex nonlinear relationships, especially in the context of multi-source data.

2.4.2. Cross-Validation

Cross-validation is a widely used model evaluation method designed to prevent overfitting and enhance the generalization ability of a model by dividing the dataset into multiple subsets for repeated training and validation. In this study, K-fold cross-validation was employed to assess the performance of various prediction models. In K-fold cross-validation, the dataset is randomly divided into K subsets of approximately equal size. For each iteration, one subset is used as the validation set, while the remaining K-1 subsets are used as the training set. This process is repeated K times, with a different subset serving as the validation set in each iteration. The results from all the K validation rounds are then averaged to provide an overall evaluation of the model. This approach effectively maximizes data usage, minimizes biases introduced by data partitioning, and offers a more stable assessment of model performance.

In this study, K was set to 10, meaning the dataset was divided into 10 subsets, resulting in 10 rounds of training and validation. During each round, 90% of the data was used for training, and 10% was used for validating the model's predictions. This method provides a more reliable estimation of the model's performance on unseen data and ensures robust generalization capability.

2.4.3. Accuracy Metrics for Population Modeling

To evaluate the performance of the population density prediction models, this study employed several accuracy metrics. The commonly used evaluation criteria include Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Mean Squared Error (MSE) measures the average of the squared differences between predicted and actual values, reflecting the overall level of prediction error. Mean Absolute Error (MAE) represents the average of the absolute differences between predicted and actual values, providing an intuitive measure of prediction accuracy. The Coefficient of Determination (R^2) assesses the model's ability to explain the variability of the dependent variable, with an R^2 value closer to 1 indicating a better fit. By using these accuracy metrics in combination, a more comprehensive evaluation of the predictive performance of different models can be conducted, allowing for the selection of the optimal prediction model.

3. Results Analysis

3.1. Spatial Distribution Pattern of Population

From the spatial distribution map of the population density in Inner Mongolia (Figure 2), it can be observed that the pattern of population density roughly forms an inverted "S" shape. Specifically, the boundary is defined by a line connecting "Ulubutie (C1)—Tule Maodu (C2)—Honggeergaole (C3)—Bayinbaolige (C4)—Jia'ergale Saihan (C5)", indicating a spatial distribution trend from the southeast to the northwest. Population density in the eastern and southern parts of this boundary is significantly higher compared to the western and northern regions. This boundary aligns closely with the 300 mm isohyet and the Daxing'anling (M1)–Yinshan mountains (M2)–Langshan Mountains (M3). Areas with higher population density are primarily concentrated in the Hetao region, through which the Yellow River flows in central Inner Mongolia (extending from south of the Daqingshan Mountains (M2) to the north bank of the Yellow River), and the Xiliao River basin in southeastern Inner Mongolia (including Tongliao and Chifeng, located south of the Daxing'anling (M1)). Conversely, regions with lower population density are mainly situated in grasslands and deserts below the 300 mm isohyet.

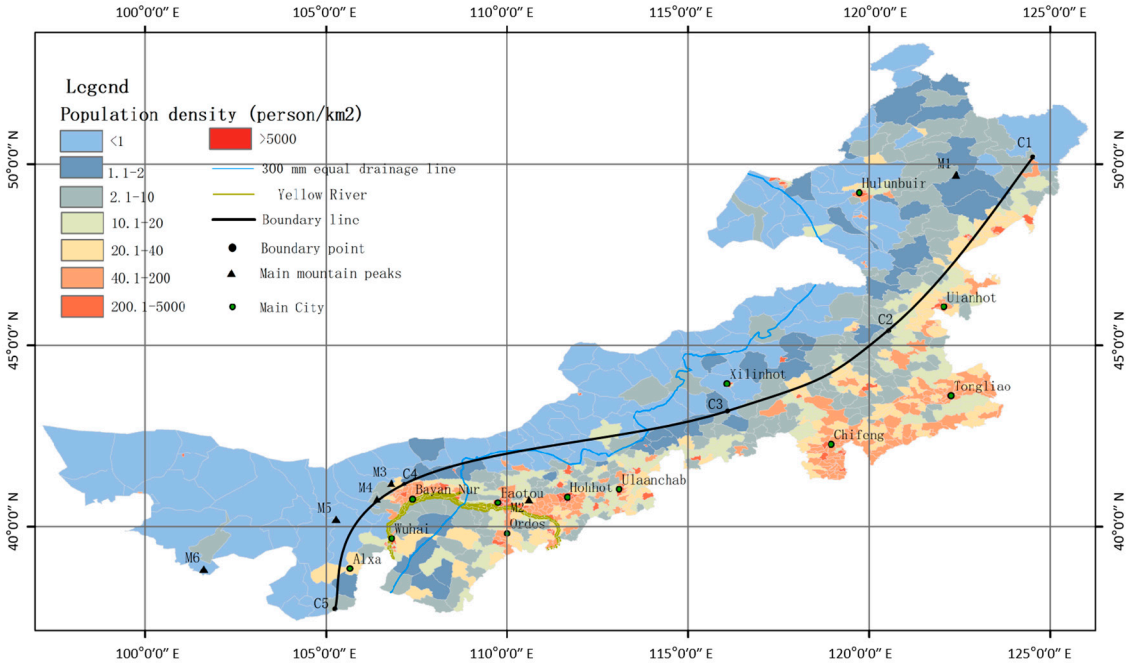

Figure 2. Population density distribution at the township level in Inner Mongolia in 2020. C1: Ulubutie; C2: Tule Maodu; C3: Honggeergaole; C4: Bayinbaolige; C5: Jia'ergale Saihan. M1: Daxing'anling Mountains; M2: Daqingshan Mountains, M3: Yinshan Mountains; M4: Lanshan Mountains, M5: Yabulai Mountains, and M6: Longshou Mountains.

The average population density of the autonomous region is 21 people/km^2. The maximum population density is found in Chifeng City (Zhenxing region) in the southeastern part of the region, with a density of 36,631.48 people/km^2, while the minimum is in Eerguna City (Enhehada region) in the northeastern part, with a density of 0.002335 people/km^2. There is a severe mismatch between population distribution and land area (Figure 3). Areas with a population density of less than 100 people/km^2 account for 97.6% of the total land area but only constitute 36.4% of the autonomous region's total population. Conversely, areas with a population density greater than 200 people/km^2 contain 54.3% of the total population, yet they occupy only 1.03% of the total land area.

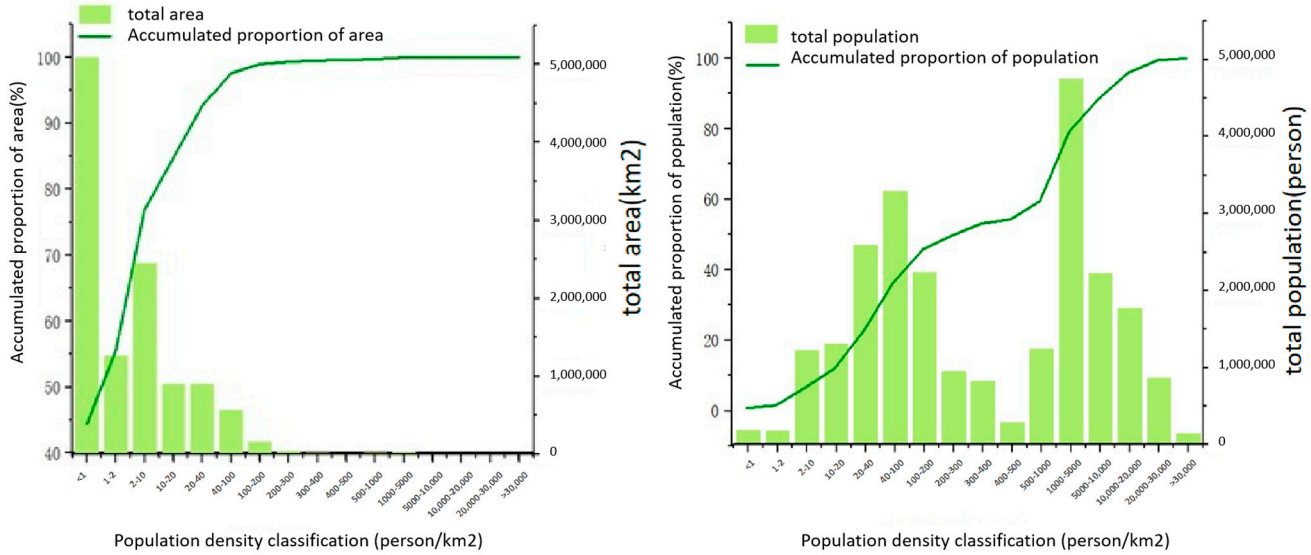

Figure 3. Line chart of total population and area of townships in Inner Mongolia with different population density classifications.

3.2. Evaluation of Multiple Models in Population Density Prediction

This study employed four models to predict population density, including linear regression, XGBoost, Random Forest, and LightGBM. The performance of the models was evaluated using three key metrics: R^2, MBE (Mean Bias Error), and MAE (Mean Absolute Error). The results indicated that the ensemble models exhibited stronger predictive capabilities compared to linear regression.

The R2 value of the linear regression model is 0.8236, indicating that the model explains approximately 82% of the variance in the data (Figure 4). However, the Mean Absolute Error (MAE) of the linear regression model is 1043.55, reflecting a significant prediction error. The Mean Bias Error (MBE) is −157.95, indicating a general tendency for underestimation and a noticeable prediction bias. This result highlights the limitations of linear regression in capturing the nonlinear and complex relationships inherent in population density data, leading to suboptimal performance of the test set.

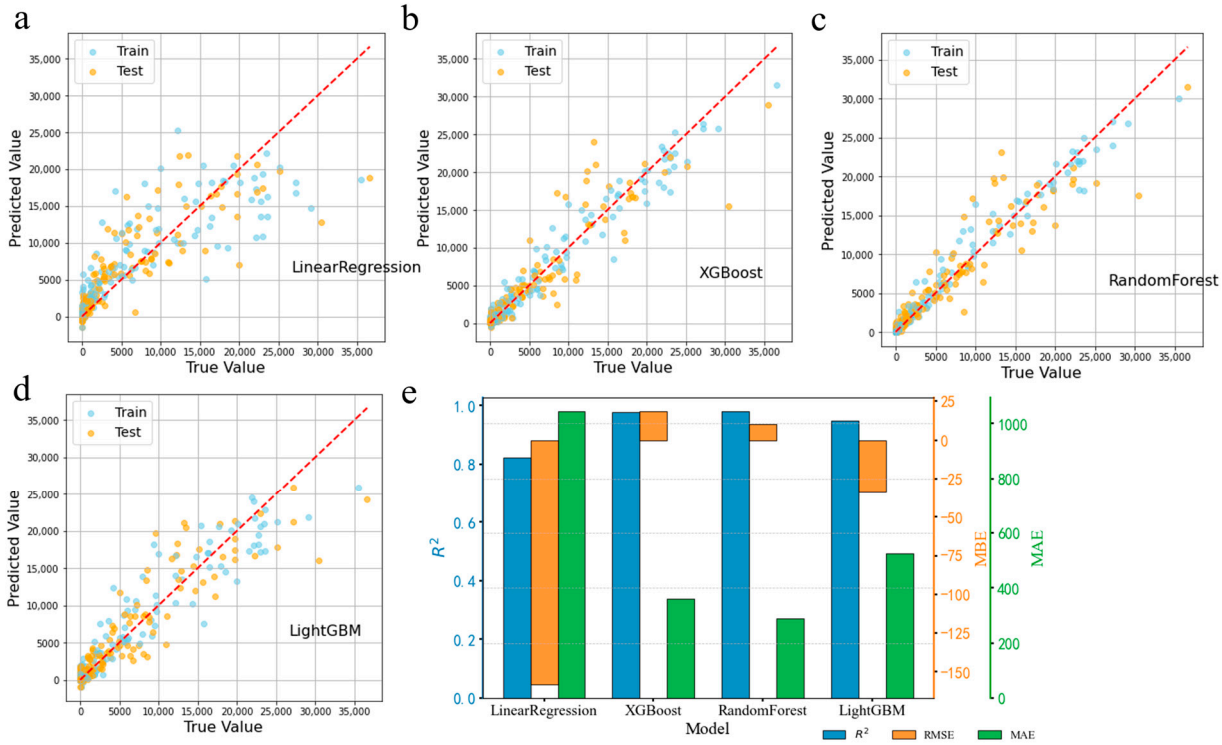

Figure 4. Comparison of population density prediction results of different regression models on the training and testing sets. (**a**–**d**) show the scatter plots of population predictions for training and testing models under different models, with the red line representing the one-to-one fit line. (**e**) presents the specific accuracy values of R^2, RMSE, and MAE for the different models.

In contrast, the XGBoost model significantly improved predictive accuracy, achieving an R^2 of 0.9758, explaining over 97% of the variance. This suggests that the model fits the training data well and maintains strong generalization capability in the test set. The MAE for XGBoost is 358.91, substantially lower than that of linear regression, confirming enhanced predictive precision. The MBE for XGBoost is 18.72, close to zero, indicating minimal systematic bias in the predictions.

The Random Forest model performed similarly to XGBoost, with an R2 of 0.9787, the highest among the four models, indicating the strongest predictive capability for population density. Its MAE is 287.23, the lowest of all the models, demonstrating that Random Forest excels in minimizing average prediction error. The MBE is 10.08, further substantiating the model's accuracy with minimal bias.

LightGBM achieved an R^2 of 0.9478, slightly lower than XGBoost and Random Forest but still explaining approximately 95% of the data variance. Its MAE is 526.22, somewhat higher than XGBoost and Random Forest, yet significantly better than linear regression. The MBE is −34.01, indicating a slight overall underestimation, though with relatively low bias.

3.3. Feature Importance Analysis Based on Interpretable Models

The best-performing Random Forest model was further employed to predict population density, and SHAP values were utilized to analyze the contribution of different features to the model's outputs. The figure displays the ranking of feature importance for each variable and their specific impacts on population density predictions. SHAP values quantitatively reveal the relationships between each variable and population density. The nighttime light index emerges as the most significant variable in population density predictions, with SHAP value analysis indicating that its importance far exceeds that of other features. Regions with higher nighttime light indices are typically closely associated with higher population densities. The broad distribution range of SHAP values for the

nighttime light index signifies its substantial impact on predicting population density. This index reflects levels of urbanization and economic development, which are highly correlated with population concentration. Therefore, its contribution to the model is prominent, establishing it as a key indicator for explaining population distribution.

Road network density ranks second in feature importance, exhibiting a strong correlation with population density. Areas with higher road network density often coincide with higher population densities, indicating a robust relationship between the two. This relationship may be influenced by the level of infrastructure development and regional attractiveness. However, due to potential interactive relationships between road network density and population density, further exploration with additional data is necessary to clarify these association mechanisms.

The cultivated land index is the third most significant variable contributing to population density predictions, with the SHAP values revealing a complex relationship with population density. The distribution of the SHAP values indicates that the influence of the cultivated land index varies across regions with differing population densities. In some low-density areas, a high cultivated land index may exhibit a negative correlation, whereas in areas with higher agricultural intensification, a positive correlation with population density may be observed.

In contrast, variables such as the water body index and river network density exhibit lower feature importance, with the SHAP values indicating a weak correlation with population density. Their relatively small SHAP value ranges suggest limited overall contributions to the model's predictions.

Through the quantitative analysis of the SHAP values, the results clearly demonstrate that the nighttime light index, road network density, and cultivated land index are the most important features for predicting population density. The correlations of these features with population density significantly enhance the model's predictive capability. Specifically, the nighttime light index has the largest SHAP value range, indicating its status as the most critical explanatory variable, while road network density and cultivated land index also play significant roles, though their effects are relatively more stable and region-specific.

The six most important variables were selected to illustrate their nonlinear relationships with population density predictions. Figure 5 depicts the fluctuations of the SHAP values for these variables as they change, highlighting the contributions of each variable to the model's predictions and their impact on population density forecasting.

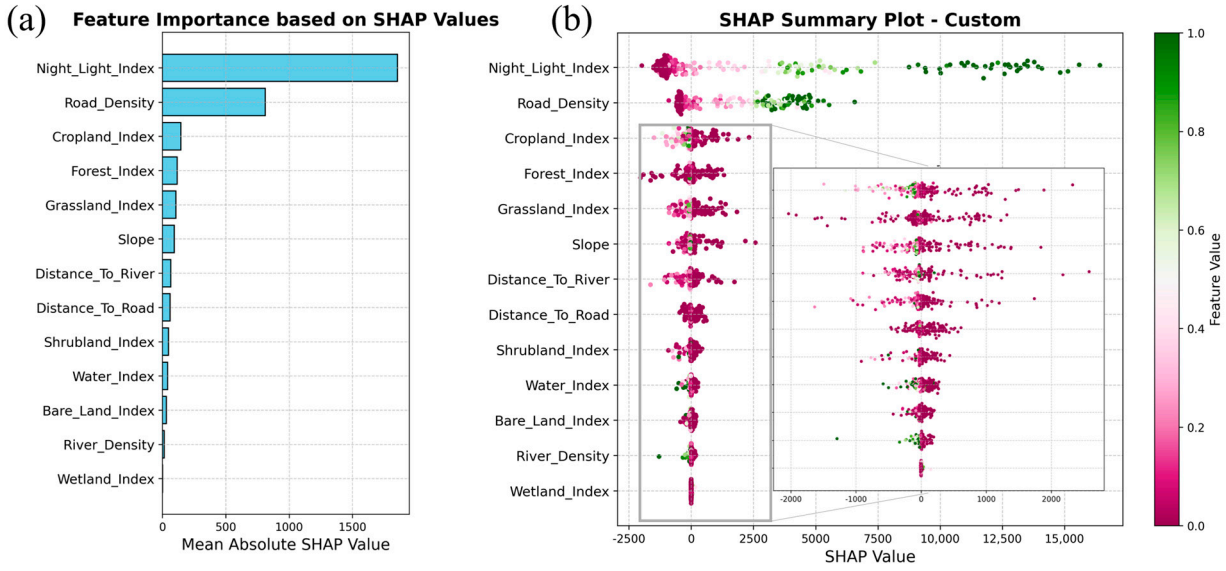

Figure 5. Feature importance analysis and SHAP value distribution of Random Forest model. (**a**) shows the feature importance values for the model predictions; Figure (**b**) displays the distribution of SHAP values for each feature at each data point.

The nighttime light index shows a clear positive correlation with population density (Figure 6). As the nighttime light index increases, the SHAP values exhibit a continuous upward trend, indicating a strong linear relationship between this variable and population density predictions. Throughout the entire range of the index, the SHAP values for the nighttime light index consistently increase without any apparent breakpoints, suggesting a persistent positive influence on population density within the model, making it one of the most important predictors.

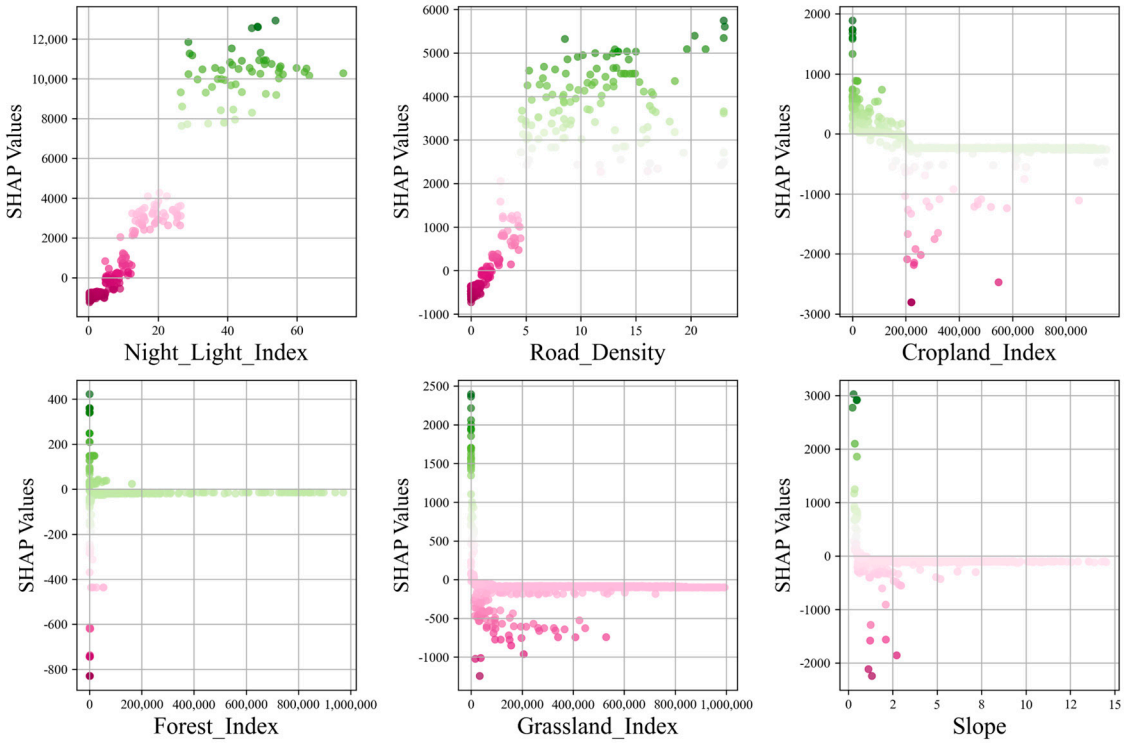

Figure 6. Nonlinear relationship between key variables and population density prediction.

In contrast, road network density demonstrates a different trend. In the lower range of road network density, the SHAP values show a linear increase, indicating a strong correlation with population density. However, once road network density reaches a certain threshold (approximately around 10), the SHAP values stabilize and no longer exhibit significant increases. This suggests that in areas with low road network density, the variable contributes significantly to predicting population density, but its marginal effect diminishes once a certain level is reached. This phenomenon indicates that in regions with well-developed infrastructure and high road network density, further increases do not substantially alter population distribution.

The cultivated land index exhibits a complex nonlinear relationship. Within the lower range of the index, the SHAP values are relatively high, indicating a strong contribution to population density predictions. However, as the cultivated land index increases, the SHAP values decline sharply and tend to stabilize, suggesting that beyond a certain extent, additional cultivated land has a limited impact on population density. This trend may be due to the fact that regions with extensive cultivated land are often sparsely populated agricultural areas, where further expansion of cultivated land does not enhance population density.

The trends for the forest index and grassland index are like that of the cultivated land index. At lower values, the SHAP values are relatively high, indicating that these variables contribute to population density predictions in certain areas. However, as the areas of forest and grassland increase, the SHAP values decline and stabilize, reflecting a diminished influence on population density. This suggests that forests and grasslands are primarily

distributed in remote or protected zones, leading to a lower correlation with population distribution.

The SHAP values for slope also exhibit nonlinear characteristics. In flatter terrain (lower slope range), the SHAP values are higher, indicating an association with higher population density. As the slope increases, the SHAP values decline sharply, suggesting a negative impact on population density as terrain becomes steeper. Once the slope exceeds a certain value (approximately around 5), the SHAP values stabilize, indicating that the marginal effect of terrain on population density diminishes in areas with extreme slopes.

In summary, the nighttime light index and road network density are the most influential variables for predicting population density. The nighttime light index maintains a strong positive correlation across its entire range, while the impact of road network density stabilizes after a certain threshold is reached. The cultivated land index, forest index, grassland index, and slope exhibit clear nonlinear relationships, with significant contributions to population density predictions in their lower ranges and diminishing marginal effects at higher values. These nonlinear relationships highlight the complex connections between these variables and population density, underscoring the critical roles of different factors and their tipping point effects within the predictive model.

4. Discussion

4.1. Performance of Machine Learning Models

The three machine learning algorithms employed in this study—Random Forest, XGBoost, and LightGBM—demonstrated significant advantages in predicting population density at the township scale in Inner Mongolia, particularly given the region's vast size and sparse population distribution. Inner Mongolia is characterized by a large geographical span and notable regional differences, with population density exhibiting strong spatial heterogeneity. Traditional multiple linear regression methods, while providing some predictive capability, are constrained by their assumption of linear relationships between variables. This limitation makes them less effective at capturing complex nonlinearities and interactions, especially in large-scale, low-density areas.

In contrast, machine learning models, particularly Random Forest, XGBoost, and LightGBM, excel at handling complex data within a diverse environment like Inner Mongolia. The Random Forest model demonstrated strong performance in managing high-dimensional data and capturing complex nonlinear relationships between variables, especially in remote areas with sparse and noisy data. Its robust noise tolerance and stability help reduce overfitting and enhance accuracy in large-scale datasets.

XGBoost and LightGBM further improved model performance by gradient boosting algorithms, making them particularly well-suited to the diverse spatial characteristics of Inner Mongolia. The region exhibits significant differences, with economically developed and densely populated eastern and southern parts, contrasting with sparsely populated western and northern areas due to factors such as resource scarcity and limited infrastructure. By employing multiple iterations to adjust weights, XGBoost and LightGBM effectively fit complex spatial heterogeneity, capturing variations in population density across different regions. For instance, in the economically active and infrastructure-rich eastern areas, these models can better illustrate the relationship between economic activity and population density. Conversely, in the remote western and northern regions, the models accurately predict population distribution despite limited sample data.

Overall, all three machine learning models significantly enhanced the prediction accuracy of population density at the township scale in Inner Mongolia, particularly in regions with substantial spatial heterogeneity and a low population density. By comparison, traditional statistical models struggle to adapt to complex spatial distributions and nonlinear relationships, making them less effective in such environments.

4.2. Relationships Between Key Factors and Population Distribution

This study employed machine learning models to predict population density at the township level in Inner Mongolia, uncovering the relationships between multiple variables and population distribution. These variables do not operate independently; rather, they interact in complex ways that collectively shape the spatial distribution patterns of the population.

Firstly, the strong correlation between the nighttime light index and population density suggests that this measure effectively reflects levels of urbanization and the distribution of economic activities within a region [34–36]. While nighttime lights indicate economically active areas, the relationship between economic activities and population density is not unidirectional. Regions with dense economic activities often exhibit high population densities, while high population density can also drive local economic development [37–39]. Future research could incorporate more detailed data on regional industrial distribution [40,41] and public service facilities [42,43] to further elucidate the complex relationship between nighttime lights and population distribution.

The impact of road network density tends to stabilize after reaching a certain level, indicating that once infrastructure is sufficiently developed, changes in population distribution are influenced by other factors, such as accessibility to public services and the allocation of infrastructural resources. This suggests that in regions with high road network density, the interaction between infrastructure and population distribution is complex. Future research could explore the mutual influences of socio-economic factors and population density to further optimize infrastructure planning [40]. Regarding land use types, farmland, forests, and other land categories exhibit nonlinear effects on population density. The interaction between different land types and population distribution varies under different contexts [44]. This variation may be attributed to differences in production efficiency, land policies, and economic development models across regions engaged in agriculture, forestry, and other activities [45].

4.3. Limitations

Although this study successfully revealed the relationships between multiple factors and population density at the township scale in Inner Mongolia, certain limitations and uncertainties remain. First, despite the use of multi-source data, the data quality and spatial coverage in remote areas of Inner Mongolia may be inadequate, potentially impacting the accuracy of predictions in regions with low population density. Future research could improve data collection by incorporating new data sources, such as high-resolution remote sensing data and social media data, to further enhance model adaptability and precision.

Second, this study primarily identified correlations between variables without delving into the causal relationships among these factors. Future research could consider incorporating causal inference models to explore the underlying mechanisms linking population distribution with factors such as the economy and infrastructure. This approach would provide stronger theoretical support for regional planning and policy-making.

Additionally, due to the significant regional differences across Inner Mongolia, a single model may struggle to adapt effectively to the characteristics of all areas. Future studies could adopt a regional modeling approach, adjusting models based on the geographic and socio-economic characteristics of different areas, thereby further improving prediction accuracy and generalizability.

5. Conclusions

This study developed a population density prediction model at the township scale in Inner Mongolia using advanced machine learning models, including Random Forest, XGBoost, and LightGBM, combined with multi-source data such as nighttime lights and transportation networks. The experimental results demonstrate that these machine learning models have significant advantages in handling complex nonlinear relationships and

multivariable interactions, thereby effectively improving the accuracy of population density predictions.

By incorporating explainable machine learning techniques, we not only enhanced the transparency of the model but also uncovered the nonlinear relationships between various variables and population density, providing new perspectives and methodologies for future research on population distribution. Despite certain limitations, such as the influence of data quality and spatial scale, this study offers strong technical support for population density prediction in small-scale regions and demonstrates high practical value.

Future research could integrate additional auxiliary data, such as social media and climate change indicators, and improve data preprocessing and model optimization techniques to further enhance prediction accuracy and model robustness. Additionally, the methods developed in this study have strong potential for broader regional applications, particularly for similar population density predictions in areas with analogous geographical conditions.

Author Contributions: Chenxi Cui: Writing—original draft, Software, Methodology, Visualization, and Reviewing and Editing. Hao Li: Methodology, Visualization, and Reviewing and Editing. Yuhai Bao: Validation, Supervision, and Funding Acquisition. Yunfeng Hu: Reviewing and Editing, Conceptualization, Supervision, and Funding Acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (42371304, 42261144746); 2023 Young Scientific and Technological Talent Development Program (NJYT23017); Natural Science Foundation of Inner Mongolia Autonomous Region, China (2022QN04002); Research Startup Project for High-Level Talent Introduced by Inner Mongolia Normal University (2021JYRC004); Special Fund for Basic Research Business of Inner Mongolia Normal University (2022JBQN098); and the Key Project of Innovation LREIS (KPI011).

Data Availability Statement: Data will be made available on request.

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

References

1. Dong, N.; Yang, X.; Cai, H. Research progress and perspective on the spatialization of population data. *J. Geo-Inf. Sci* **2016**, *18*, 1295–1304.
2. 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]
3. 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]
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. Cheng, F.; Zhao, G. Fine-scale simulation of population distribution based on zoning strategy and machine learning. *Sci. Surv. Mapp.* **2020**, *45*, 165–173.
6. 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]
7. Zhu, J. The model of population urbanization in urban land spatial planning based on multi-source data fusion: A case study of Yangzhou city. *J. Nat. Resour.* **2019**, *34*, 2087–2102. [CrossRef]
8. Yao, Y.; Liu, X.; Li, X.; Zhang, J.; Liang, Z.; Mai, K.; Zhang, Y. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1220–1244. [CrossRef]
9. Fukuda, K. Interpolation and forecasting of population census data. *J. Popul. Res.* **2010**, *27*, 1–13. [CrossRef]
10. Liao, S.-b.; Li, Z.-h. Study on spatialization of population census data based on relationship between population distribution and land use—Taking Tibet as an example. *J. Nat. Resour.* **2003**, *18*, 659–665.
11. Yunfeng, H.; Guanhua, Z.; Qianli, Z. Spatial Distribution of Population Data Based on Nighttime Light and LUC Data in the Sichuan-Chongqing Region. *J. Geo-Inf. Sci.* **2018**, *20*, 68–78.
12. Jiang, D.; Yang, X.-H.; Wang, N.-B.; Liu, H.-H. Study on Spatial Distribution of Population Based on Remote Sensing and GIS. *Adv. Earth Sci.* **2002**, *17*, 734.
13. Tian, Y.; Yue, T.; Zhu, L.; Clinton, N. Modeling population density using land cover data. *Ecol. Model.* **2005**, *189*, 72–88. [CrossRef]
14. Cheng, L.; Wang, L.; Feng, R.; Yan, J. Remote sensing and social sensing data fusion for fine-resolution population mapping with a multimodel neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5973–5987. [CrossRef]

15. Chun, J.; Zhang, X.; Huang, J.; Zhang, P. A gridding method of redistributing population based on POIs. *Geogr. Geo-Inf. Sci* **2018**, *34*, 124.
16. 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]
17. Li, L.; Zhang, Y.; Liu, L.; Wang, Z.; Zhang, H.; Li, S.; Ding, M. Mapping changing population distribution on the Qinghai–Tibet Plateau since 2000 with multi-temporal remote sensing and point-of-interest data. *Remote Sens.* **2020**, *12*, 4059. [CrossRef]
18. 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]
19. 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]
20. 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]
21. 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. Geo-Inf.* **2020**, *9*, 369. [CrossRef]
22. Freire, S.; Kemper, T.; Pesaresi, M.; Florczyk, A.; Syrris, V. Combining GHSL and GPW to improve global population mapping. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 2541–2543.
23. Sutton, P.C.; Elvidge, C.; Obremski, T. Building and evaluating models to estimate ambient population density. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 545–553. [CrossRef]
24. Ye, T.; Zhao, N.; Yang, X.; Ouyang, Z.; Liu, X.; Chen, Q.; Hu, K.; Yue, W.; Qi, J.; Li, Z. 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]
25. Wang, Y.; Huang, C.; Zhao, M.; Hou, J.; Zhang, Y.; Gu, J. Mapping the population density in mainland China using NPP/VIIRS and points-of-interest data based on a random forests model. *Remote Sens.* **2020**, *12*, 3645. [CrossRef]
26. Sinha, P.; Gaughan, A.E.; Stevens, F.R.; Nieves, J.J.; Sorichetta, A.; Tatem, A.J. Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Comput. Environ. Urban Syst.* **2019**, *75*, 132–145. [CrossRef]
27. Li, Y.; Liu, M. Spatialization of population based on Xgboost with multi-source data. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Zhangjiajie, China, 23–25 April 2021; p. 012083.
28. Şahinarslan, F.V.; Tekin, A.T.; Çebi, F. Machine Learning Algorithms to Forecast Population: Turkey Example. In Proceedings of the International Engineering and Technology Management Summit 2019, Istanbul, Turkey, 10–12 October 2019.
29. Zhong, Y.; Lin, A.; Zhou, Z.; Chen, F. Spatial pattern evolution and optimization of urban system in the Yangtze River economic belt, China, based on DMSP-OLS night light data. *Sustainability* **2018**, *10*, 3782. [CrossRef]
30. Cheng, D.; Li, X. Relationship between population distribution and topography of the Wujiang River Watershed in Guizhou province. *Geogr. Res.* **2020**, *39*, 1427–1438.
31. Peng, S.; Ding, Y.; Wen, Z.; Chen, Y.; Cao, Y.; Ren, J. Spatiotemporal change and trend analysis of potential evapotranspiration over the Loess Plateau of China during 2011–2100. *Agric. For. Meteorol.* **2017**, *233*, 183–194. [CrossRef]
32. Peng, S.; Gang, C.; Cao, Y.; Chen, Y. Assessment of climate change trends over the Loess Plateau in China from 1901 to 2100. *Int. J. Climatol.* **2018**, *38*, 2250–2264. [CrossRef]
33. Gao, J.; Shi, Y.; Zhang, H.; Chen, X.; Zhang, W.; Shen, W.; Xiao, T.; Zhang, Y. *China Regional 250 m Normalized Difference Vegetation Index Data Set (2000–2022)*; National Tibetan Plateau/Third Pole Environment Data Center: Beijing, China, 2023.
34. Bagan, H.; Yamagata, Y. Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience Remote Sens.* **2015**, *52*, 765–780. [CrossRef]
35. Sutton, P.; Roberts, D.; Elvidge, C.; Baugh, K. Census from Heaven: An estimate of the global human population using night-time satellite imagery. *Int. J. Remote Sens.* **2001**, *22*, 3061–3076. [CrossRef]
36. Henderson, M.; Yeh, E.T.; Gong, P.; Elvidge, C.; Baugh, K. Validation of urban boundaries derived from global night-time satellite imagery. *Int. J. Remote Sens.* **2003**, *24*, 595–609. [CrossRef]
37. Fee, K.D.; Hartley, D.A. Urban Growth and Decline: The Role of Population Density at the City Core. Economic Commentary. 2011. Available online: https://www.clevelandfed.org/publications/economic-commentary/ec-201127-urban-growth-and-decline-the-role-of-population-density-at-the-city-core (accessed on 18 November 2024). [CrossRef]
38. Greyling, T.; Rossouw, S. Non-economic quality of life and population density in South Africa. *Soc. Indic. Res.* **2017**, *134*, 1051–1075. [CrossRef]
39. Mutunga, A. Examining Effects of Changes in Population Density on Economic Growth in Kenya. Master's Thesis, University of Nairobi, Nairobi, Kenya, 2020.
40. Zeng, P.; Zong, C. Research on the relationship between population distribution pattern and urban industrial facility agglomeration in China. *Sci. Rep.* **2023**, *13*, 16225. [CrossRef]
41. Guan, X.; Wei, H.; Lu, S.; Su, H. Mismatch distribution of population and industry in China: Pattern, problems and driving factors. *Appl. Geogr.* **2018**, *97*, 61–74. [CrossRef]
42. Shi, Y.; Yang, J.; Shen, P. Revealing the correlation between population density and the spatial distribution of urban public service facilities with mobile phone data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 38. [CrossRef]

43. Ding, J. Study on the Spatial Distribution of Public Service Facilities in the Central Districts of Nanjing based on POI Data. *Trans. Econ. Bus. Manag. Res.* **2023**, *2*, 162–174.
44. Zhang, H.; Zhang, S.; Liu, Z. Evolution and influencing factors of China's rural population distribution patterns since 1990. *PLoS ONE* **2020**, *15*, e0233637. [CrossRef]
45. Xu, Z.; Ouyang, A. The factors influencing China's population distribution and spatial heterogeneity: A prefectural-level analysis using geographically weighted regression. *Appl. Spat. Anal. Policy* **2018**, *11*, 465–480. [CrossRef]