Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 1, 2023

Population Spatialization based on Random Forest Model and Multi-source Geospatial big data

Xusong Zhang *, Maria Rosario Rodavia

Graduate School, Angeles University Foundation, Angeles 2009, Philippines
* **Corresponding author**: Xusong Zhang (Email: zhang.xusong@auf.edu.ph)

Abstract: Population spatialization research is an important approach to achieve fine-grained management of urban space and coordinated development of rural resources and the environment. By converting administrative-level population data into a finer grid scale, it allows for in-depth analysis of the spatial distribution characteristics of population density and geographic heterogeneity within a region. Currently, in China, a population census is conducted every ten years, with the township as the smallest statistical unit. However, due to advancements in computer science and geography, the level of precision in data can no longer meet the requirements of modern geographical research. Population spatialization, based on national population statistics, utilizes techniques such as multi-source data fusion and data mining to decompose large-scale population data into corresponding grid-based data, enabling more accurate spatial representation of national population statistics and facilitating the understanding of population distribution patterns. This study used administrative boundary data for 88 counties in Guizhou Province in 2021, county-level population data from the 2021 China County Statistical Yearbook, and diverse geospatial data from Guizhou in 2017. Nine spatial variables that impact the spatial distribution of the study area's population, such as points of interest and nighttime light indices, were extracted. A random forest method was used to construct a population spatialization model and simulate population distribution.

Keywords: Random Forest Algorithm; Population Simulation Distribution; Multi-source Geospatial Big Data; Points of Interest (POI).

1. Introduction

Since the reform and opening up, China has undergone a transformation from a planned economy to a market economy, which has also been accompanied by an unprecedented urbanization process [4]. The accelerated development of urbanization has brought about economic development and social progress, but it has also brought many challenges to cities. Firstly, the rapid growth of urbanization has led to a tense supply-demand relationship between urban population and resources such as housing, education, and healthcare, exacerbating the contradiction between urban population and land. In addition [5], the loss of rural population has led to a shortage of rural labor, exacerbating the aging phenomenon of rural population, and leading to issues such as farmland conversion, abandonment, and abandonment. Secondly, the rapid development of urbanization may also lead to a series of serious environmental problems. For example, despite rapid population growth in cities, problems such as water scarcity, land scarcity, and inadequate sanitation facilities still exist, further exacerbating the ecological pressure on the urban environment. In addition, urban population has heterogeneity and complex spatial patterns in space, which poses difficulties in scientifically understanding the contradictions between people and land within cities and solving urban problems. Population data is an important basic data for studying fields such as socio-economic and human geography [6]. However, traditional population statistics in China are conducted every ten years, and the smallest administrative unit for statistics is townships. Therefore, the annual resident population statistics are recorded in the form of yearbooks, making it difficult to express the spatial attributes of the data. This situation results in a long data update cycle and difficulty in obtaining real-time data. With the development of geospatial technology, many basic analytical data have spatial attributes. The low spatiotemporal resolution of traditional population data makes it unsuitable for spatial analysis, which poses scientific difficulties in understanding the natural and social coupling mechanisms of sustainable development at multiple scales. With the development of geospatial technology, many basic analytical data have spatial attributes. The low spatiotemporal resolution of traditional population data makes it unsuitable for spatial analysis, which poses scientific difficulties in understanding the natural and social coupling mechanisms of sustainable development at multiple scales.

2. Research on Population Simulation Based on Random Forest Algorithm

2.1. Random Forest

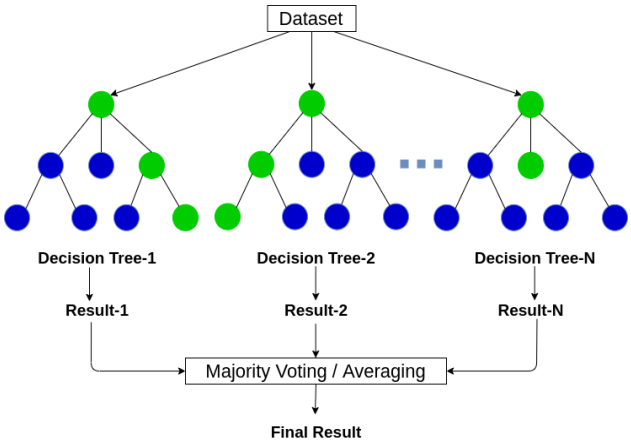

Figure 1. Working process of random forest

Random forest is the construction of multiple decision trees

based on decision trees by randomly selecting several features from samples. Random sample selection makes the model stochastic and improves the problem of overfitting caused by the inability to control the depth of the tree in the decision tree. The main core idea is to select the best classified tree from multiple decision trees constructed as the output result. The working process of a random forest is shown in the Figure [7]. (Figure 1.)

2.2. Overview of the Study Area

This study selected Guizhou Province, China as the research area. Guizhou Province is a provincial-level administrative region in China, located in the hinterland of the southwestern inland region. It is a transportation hub and an important component of the Yangtze River Economic Belt. Guizhou is located between 24°37′～29°13′. north latitude and 103°36′～109°35′east longitude, adjacent to provinces such as Sichuan, Chongqing, Hunan, Guangxi, and Yunnan. The landforms of Guizhou are mainly composed of plateaus, mountains, hills, and basins, with 92.5% of the area being mountainous and hilly. With a total area of about 176000 square kilometers, Guizhou has a subtropical monsoon climate and two major water systems, namely, the Yangtze River and the Pearl River. As of March 31, 2021, Guizhou Province has a total of 6 prefecture level cities, 3 autonomous prefectures, and 10 county-level cities, 50 counties, 11 autonomous counties, 1 special zone, 16 districts, and other county-level administrative regions. In addition, there are a total of 1509 towns, 122 townships, 193 ethnic townships, 362 streets, etc. As of 2021, the permanent population of Guizhou Province is approximately 38.52 million. (Figure 2.)

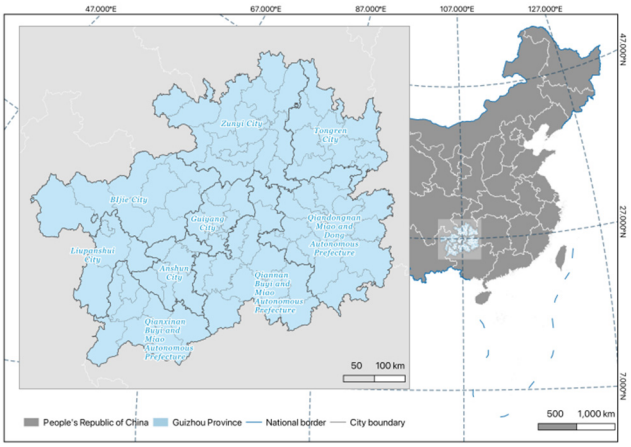

Figure 2. Guizhou Map Area

2.3. Data Preprocessing

The original data of this article is vector, raster, and various text data. Due to the fact that these data belong to different resolutions and coordinate systems, it is necessary to preprocess these data. Utilize ArcGIS to process vector and raster data with spatial attributes, use Excel 2016 and Origin 8 to process chart data, and use Python language to support calling the MaxMinNormalization function for normalization processing. Data preprocessing mainly includes the unification of coordinate system and resolution, as well as the grid processing of multivariate geospatial data [8,9].

The specific steps are as follows:

(1) The unified projection coordinate system in this article is set to WGS 1984 UTM ZONE 49N in Arcgis, which is the horizontal Mercator projection, and the data is matched with the boundaries of administrative divisions.

(2) The research scope of this article is mainly focused on Guangxi Zhuang Autonomous Region, and the Spatial Analyst tool in Arcgis10.2 is used to extract raster data by mask. Vector data is mainly cropped using the analysis tool in Arcgis10.2 to extract and analyze the crop analysis.

3. Experimentation

Random forest regression model is a method that can identify the nonlinear relationship between variables and predicted values, and accurately describe the complex relationship between multi-scale regression variables. This section intends to establish a multiple linear regression model for 9 population factors and permanent population at the township level [10,11], and on this basis, construct a population spatialization model based on 9 population factors. (Figure 3.)

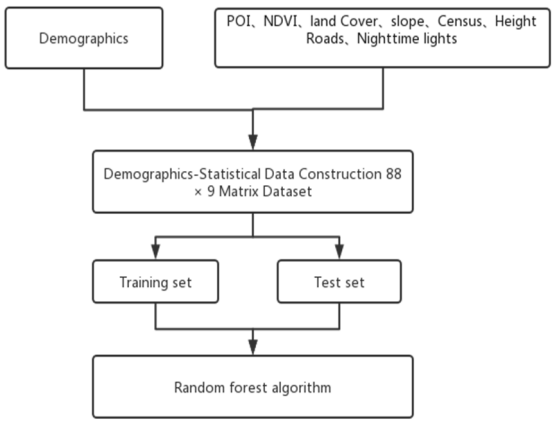

Figure 3. Population spatialization process based on Random Forest regression model

The random forest regression method mainly simulates population distribution in the following four steps:

(1) Training data preparation: This study takes county-level data of each indicator as the independent variable, extracts the average values of each factor at the township level, and constructs a matrix size of 88 × A dataset of 9 is input as an independent variable into the random forest model for training. Constructing 88 based on the population density of districts and counties × A dataset of 1 is input as the dependent variable into the random forest model for training. Call the scikit learn library to split the training and testing sets of samples according to the 7:3 random allocation principle, that is, 70% of the sample data is trained, and 30% of the sample data is used as the testing set for debugging to find the best parameters [12].

(2) Sample and Growth: Based on the Bootstrap method, extract b training samples from n sample source datasets and construct b regression trees [13]. The remaining samples were validated as packaged in bags (OOB).

(3) OBB accuracy verification: Machine learning improves model accuracy by setting parameters to adjust the model structure, so adjusting the model to the optimal parameters is extremely important for the results of random forests [14]. This article uses the GridSearchCV class provided in the Scikit-Lear library in Python language to find the optimal parameters for the Random Forest model. After extensive debugging, the accuracy of the RF model OD is 94.56%. Then, save the model training parameters.

Table 1. Main parameters of Random Forest

Parameter	Parameter value	The role of parameters
n_estimators	155	Number of decision trees
min_samples_leaf	1	Number of small sample leaf nodes
min_samples_split	5	Minimum sample size for internal node partitioning
max_depth	7	Maximum depth of decision tree
max_features	auto	Number of subsets of randomly selected feature sets
bootstrap	True	Parameters of control sampling techniques

4. Conclusion

The commonly used accuracy verification method for random forests is to evaluate the fit of statistical models by comparing the predicted values of 30% of the predicted data with the initial values. Unlike the precision analysis commonly used in accompanying forests, the precision analysis method of population spatialization mainly involves summarizing the population density of the subdivision grid onto the corresponding minimum administrative unit, and comparing the population on the minimum administrative unit generated by the experimental summary. This article obtains the population density of each grid by substituting grid feature data into a trained random forest model, and then summarizes the population data on each grid unit to the corresponding townships to compare the population from the sixth population census of Guizhou Province in 2010 to output the fitting degree. The majority of the predicted data values in this experiment are distributed near the trend line, and the goodness of fit R^2=0.876 is close to 1, indicating that the prediction accuracy of this experiment is relatively high. At the same time, it should be noted that in areas with a population of less than 300000, most of the predicted values are higher than the actual statistical population. However, in areas with a population above 30W on the horizontal axis, the predicted values of the population are relatively low. This indicates that the predicted values of rural and township areas are higher, and the predicted values of built-up areas and nearby cities are lower. This result is a completely opposite result compared to other studies, and further analysis and discussion are needed. Through the statistical data in the segmented statistical table of population spatialization errors in Guizhou Province, we can see that 63.1% of the population prediction results are within the error range of -10% to 10%, and 27.3% are within the error range of -20% to 20%. Most of the prediction results are within a relatively reasonable range.

Table 2. Estimation Error Statistics Table

Close to actual value	error range	number	Proportion
Close to actual value	(-10%, 10%)	61%	54
overestimate	(-30%, -10%)	1%	1
Mild overestimation	(10%, 30%)	17%	15
Mild underestimation	>30%	15%	13
underestimate	<30%	6%	5

Although there are still significant or minor errors in the predicted data of this experiment compared to the statistical data, statistical analysis shows that the total error value is within an acceptable range. By performing a dimensionality reduction process on a large range of population data, we have obtained grid data with a more accurate and semantically rich description of population density in the smallest administrative region, this indicates that the progressive forest model has better population spatialization results compared to traditional linear regression models.

References

[1] Smith, A., Martin, D. , & Cockings, S. . (2016). Spatio-temporal population modelling for enhanced assessment of urban exposure to flood risk. Applied Spatial Analysis & Policy, 9 (2), 145-163.

[2] Jia, P. , Qiu, Y. , & Gaughan, A. E. . (2014). A fine-scale spatial population distribution on the high-resolution gridded population surface and application in alachua county, florida. Applied Geography, 50, 99-107.

[3] Stevens F R , Gaughan A E , Linard C , et al. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data[J]. PLoS ONE, 2015, 10(2):e0107042.

[4] .Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data. Journal of Applied Remote Sensing, 9(1), C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[5] Liaw, A. , & Wiener, M. . (2002). Classification and regression by randomforest. R News, M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[6] Ramón Díaz-Uriarte, Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest[J]. Bmc Bioinformatics, 2006, 7(1):3.

[7] Assuncao, M. D. , Calheiros, R. N. , Bianchi, S. , Netto, M. A. S. , & Buyya, R. . (2013). Big data computing and clouds: challenges, solutions, and future directions. Journal of Parallel and Distributed Computing, 79-80, 3–15.

[8] Wenjie, Wu, Jianghao, Wang, Tianshi, & Dai. (2016). The geography of cultural ties and human mobility: big data in urban contexts. Annals of the American Association of Geographers.

[9] Miller, H. J. . (2015). Selected keynote speech abstracts: Data-driven geography: Can we build geographic knowledge from big data?. IEEE International Conference on Spatial Data Mining & Geographical Knowledge Services. IEEE.

[10] Elvidge, C. D., Zhizhin, M., & Zhizhen, L. (2017). Using random forests to estimate the distribution of impervious surface area based on remotely sensed imagery and ancillary data. Remote Sensing of Environment, 193, 149-163.

[11] Li, Y., & Chen, Y. (2021). Random Forest: A Literature Review. Journal of Big Data Science and Engineering, 6(3), 67-79.

[12] .Gamba, A., Conti, M., & Rossi, L. (2019). Random Forests for Time Series Forecasting: A Review of Applications and Ensemble Designs. Data Mining Knowl. Disc., 23(4), 1769-1797.

[13] Pérez-Lombard, L., Sánchez-Romero, J. M., & Pinto, J. (2018). Using Random Forest to Predict the Number of Visitors in Cultural Attractions. Tourism Management, 62, 45-56.

[14] Zhao, X., Wang, W., & Wu, F. (2021). Spatialization of Population Distribution Based on Remote Sensing Images and

GIS Data: A Case Study in China. Remote Sensing, 13(16), 3480.IEEE Criteria for Class IE Electric Systems (Standards style), IEEE Standard 308, 1969.

[15] Xu, Z., Zhang, X., & Wu, X. (2020). Integrating Remote Sensing Images and Geo-Social Data for Population Distribution Prediction: A Deep Learning Approach. ISPRS International Journal of Geo-Information, 8(5), 192.R. E. Haskell and C. T. Case, "Transient signal propagation in lossless isotropic plasmas (Report style)," USAF Cambridge Res. Lab., Cambridge, MA Rep. ARCRL-66-234 (II), 1994, vol. 2.

[16] Wang, Y., He, Q., & Wu, J. (2019). Predicting the Spatial Distribution of Population Using Remote Sensing Images and Machine Learning Algorithms: A Case Study in China. Applied Sciences, 9(17), 3867.

[17] Liu, Y., & Wang, Z. (2018). Integrating Remote Sensing Images and Mobile Phone Data for Population Distribution Prediction: A Random Forest Approach. Remote Sensing Letters, 9(9), 743-752.