Spatialization of population based on Xgboost with multi-source data

Yinxing Li¹, Minghao Liu²

¹ College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
² ESpatial Information Research Center, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
*Corresponding author’s e-mail: liumh@cqupt.edu.cn

Abstract. Aiming at the problem that demographic data cannot visually and clearly show the true distribution of population and cannot be combined with other environmental resource spatial data for analysis. This paper takes Chongqing as an example, selects nighttime light data etc. as variable factors affecting population distribution. Using the Xgboost model to build a regression model on the county level, and generates the population data of 100m in Chongqing in 2010. The accuracy of the population spatialization results and three public data sets were compared on the township scale. Finally, based on the importance of the variable factors of the Xgboost model, the influencing factors of the spatial distribution of Chongqing's population were explored. The results show that the root mean square error in this paper is significantly better than the other three population data sets, the absolute value error is significantly better than the GPW data set and the Chinese kilometer grid data set, and slightly better than the WorldPop data set. Through the analysis of the importance of variable factors, it is found that the distance from construction land is the most important indicator, and the nighttime light data, residential area and POI data all play an important role in population distribution of Chongqing.

1. Introduction
The population is a complex social entity that integrates multiple social relationships. On June 17, 2019, the United Nations Department of Economic and Social Affairs reported in the World Population Outlook 2019: Discovery Summary that the global population is expected to increase by another 2 billion people in the next 30 years, from 7.7 billion in 2019 to 2050 Of 9.7 billion, China is still the world’s most populous country. All social activities, social relations, social phenomena and social problems are related to the population development process. Therefore, accurate population spatial distribution has important application value and scientific significance for the planning and decision-making, disaster assessment, resource allocation and other aspects of government departments at all levels.

The spatial distribution of population refers to the geographical distribution of the population at a certain point in time. It is the core issue of the study of population geography and the important foundation of the study of man-land relationship. Census data is the main source of demographic data and has certain limitations in the application of earth sciences. First of all, the census data only provides a population count value for each census unit; therefore, it cannot clarify the spatial population distribution within each census unit, nor can it reflect the changes within the population[1]. In addition, the unit of census data is sometimes inconsistent with the unit of socio-economic variables, and is
inconsistent with the regional system of natural variables (such as the layer of remote sensing images). Due to the lack of clear and detailed geographic reference, the combination of demographic data and geographic reference environmental data is difficult, this problem hinders interdisciplinary research. So the spatialization of demographic data is a very important task.

In the past few decades, people have developed a variety of methods to downscale the census data of irregular administrative units to a gridded population distribution map on a fine scale. The earliest was mainly based on the negative exponential model of mathematical functions[2] and the simple area weighting method[3]. Since these models did not consider the influence of spatial population distribution, spatial interpolation methods were subsequently proposed, among which are surface interpolation, point interpolation and geostatistical methods. Later, as technology was updated, more and more auxiliary data were added to the model. Dasymetric Mapping was developed to combine auxiliary data to improve the details of the gridded population map. Wright (1936) used the United States. Geological Survey Bureau topographic map, estimate the population density of different types of residential areas, for the first time the area density mapping technology is introduced into the study of population spatial distribution characteristics; the famous WorldPop population distribution project developed a semi-automatic Dasymetric method based on random forest regression to generate grids Population map[4]; Many scholars use random forest models based on the idea of partition density and combine POI data to spatialize the population to improve the accuracy and details of the grid population[5-9]; Gervasoni[10] use the cnn model to build an end-to-end population mapping; Li[11] combine night light and land use to generate city night light index; Wei[12] proposed a new urban-scale demographic model method. Although many scholars have made a lot of explorations on different data sources, different scales, and different simulation methods, most of them are mesoscale studies. Most of the spatial resolution is 1km, which is difficult to meet the current requirements for refined urban management. Less explored the impact of spatial variables on population distribution.

This paper attempts to use the Xgboost model to explore the relationship between the spatial auxiliary data such as night light data, residential data, land use data, elevation data, topography data, POI data and the population distribution of Chongqing, and to achieve a 100m resolution of Chongqing The population is rasterized and the importance of variables obtained through the Xgboost model is used to analyze the spatial data affecting the population distribution of Chongqing.

2. Research area and data processing

2.1. Introduction to the study area
Chongqing is located in the southwestern inland China, in the upper reaches of the Yangtze River, bordering Hubei, Hunan, Guizhou, and Sichuan. The main stream of the Yangtze River runs through the entire territory from west to east. The jurisdiction is 470 kilometers long from east to west, 450 kilometers wide from north to south, and covers an area of 82,400 square kilometers. The terrain of Chongqing descends gradually from north to south to the Yangtze River valley. The northwest and central parts are dominated by hills and low mountains. The southeast has two major mountain ranges, Daba Mountain and Wuling Mountain, so it is known as the "mountain city". Chongqing is the largest industrial and commercial city in Southwest China, an important modern manufacturing base of the country, and an important transportation hub in Southwest China.

2.2. Data source and data processing
The spatial distribution of the population is the result of the combined effects of natural factors and socio-economic factors in the region. For cities, the main factors affecting their population distribution are socioeconomic factors, such as night light data, POI data, Twitter data[13], taxi GPS[14], etc.; for rural areas, The main factors affecting its population distribution are natural factors, such as elevation data, NDVI data[11], etc.
The main data used in this study are night light data, land cover data, POI data, settlement data, roads, rivers, demographic data, and administrative boundary data. POI data is crawled through Amap API, the last crawling time (March 2020), and the details of the number of various POIs.

The above heterogeneous data have been processed consistently (unified research scope, unified spatial coordinate system is GCS_WGS_1984, unified grid resolution is 100m), and a unified spatial database is established. Data source and details (see Table 1)

| Data                      | Year | Type   | Data Source      |
|---------------------------|------|--------|------------------|
| County-level demographic data | 2010 | Excel  | China Census     |
| Township demographic data  | 2010 | Excel  | China Census     |
| Administrative boundaries  | 2010 | Shapefile | OpenStreetMap |
| River                     | 2010 | Shapefile | OpenStreetMap |
| Road                      | 2010 | Shapefile | OpenStreetMap |
| Land use                  | 2010 | Raster  | GlobeLand30      |
| DEM                       | 2011 | Raster  | ASTER GDEM       |
| Night-time light          | 2010 | Raster  | WorldPop         |
| Residential area          | 2010 | Raster  | WorldPop         |
| POI                       | /    | Excel   | Amap             |

3. Research method and processes

3.1. The technological process of population spatialization based on Xgboost

The main steps to achieve population spatialization based on Xgboost: 1) Data filtering and processing: preprocess the collected geospatial data and social perception data to generate auxiliary data (including projection transformation, cropping vector data and raster data, To match the spatial scope of the census data and other operations); generate gridded covariate data from auxiliary data; 2) upscale aggregation of raster data: aggregate gridded covariate data to county-level administrative units, Generate covariates on the census unit scale (aggregation of independent variable data); combine census data and county-level administrative unit data to obtain population density data of county-level administrative units as the dependent variable of the model (aggregate dependent variable data); 3) XGboost model construction and training: input covariates and county-level administrative unit population density data into the XGboost model, train the model and tune it; substitute grid-scale feature data into the trained model to obtain the population spatial distribution weight layer; 4) Dasymetric Mapping: Combine census data and use the method of district density mapping to obtain population spatialization data; 5) Accuracy test: evaluate the accuracy of population spatialization data through a set of accuracy evaluation indicators (Figure 1).
3.2. Xgboost

Xgboost extreme gradient boosting tree is an end-to-end boosting tree system proposed by Chen Tianqi in 2016. It is improved through the gradient boosting decision tree algorithm (gradient boosting decision tree, GBDT). It is an optimized distributed gradient boosting library with high efficiency, Advantages such as flexibility and portability\textsuperscript{15}.

The idea of the algorithm is to continuously add trees to fit the residuals of the last prediction, and the final result is the sum of the results of all trees. The model formula is expressed as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$  \hspace{1cm} (1)

Where $\mathcal{F} = \{ f(x) = w_{q(x)} \}(q: R^m \rightarrow T, w \in R^T)$ is the space of the CART tree, where $q$ represents the score of each tree structure mapping each sample to the corresponding leaf node, that is, $q$ represents the tree model, input a sample, and map the sample to the leaf node according to the model to output the predicted Score; $w_{q(x)}$ represents the set of scores of all leaf nodes of tree $q$; $T$ is the number of leaf nodes of tree $q$.

3.3. Dasymetric Mapping

Since the grid-scale population density obtained by the Xgboost model is predicted by the Xgboost model trained based on the county-level administrative unit scale covariates and population density, the actual population density requires the total population of each county-level administrative unit for total control. According to the grid-scale population density as the weight layer, the population of each county-level administrative unit is redistributed to each grid, the formula is as follows:

$$p = c_r \times w_p + w_r$$  \hspace{1cm} (2)

$P_p$ is the estimated population spatialized in a 100m grid, $C_r$ is the census data in each county-level administrative unit, $W_p$ is the population density of each 100m grid generated by the Xgboost model, and $W_r$ is the total population density within the administrative unit.

Through this formula and the weight layer generated by Xgboost, this paper disperses the 2010 census data of district and county administrative units into a 100m grid by district density mapping method to complete population spatialization.
4. Results and analysis

4.1. Analysis of spatialization results
Through XGboost model learning and district density mapping, a population density map with a spatial resolution of 100m in Chongqing in 2010 was obtained (Figure 2). The population density of Chongqing as a whole shows a trend of spreading from Yuzhong District to the surrounding area. The specific manifestation is that the population in the western part of Chongqing is relatively dense, while the population in the northeast and southeast is relatively sparse. The highest population density occurs in Yuzhong District. Mainly due to the relatively flat terrain in the central and western regions, the population is mainly concentrated in this area, while the northeast and southeast are mostly mountainous terrain with rugged terrain and large terrain undulations, resulting in less population distribution and lower density.

4.2. Accuracy inspection
Randomly select 170 township and street-level administrative units in Chongqing City, through the spatialization results of this article and three mainstream public grid population datasets at home and abroad[16] (WorldPop[4], GPW[17], China’s population spatial distribution kilometer grid dataset[18](CNPop)) The accuracy test is performed. The root mean square error(RMSE) and mean absolute value error(MAE) of the four population spatialization results in the census data of 160 township and street level administrative units are shown in Table 2. It can be seen from Table 2 that the RMSE of this paper in the study area is significantly better than the other three population data sets, and the MAE is significantly better than the GPW and the CNPop, and slightly better than the WorldPop. Table 3 is a grading statistical table constructed by the relative errors of 170 townships. The number of townships accurately estimated in this article are all higher than the other three data sets. The number of townships that are seriously overestimated is also less than the other three data sets, which is seriously underestimated. The number of townships is much smaller than the WorldPop and close to the other two datasets.

Table 2. Error comparison of different population spatialization.

| Data            | RMSE   | MAE    |
|-----------------|--------|--------|
| This paper      | 28629.99 | 10975.09 |

Figure 2. Population spatialization based on xgboost

Figure 3. Four data comparison. (a) Xgboost-based method, (b) Worldpop, (c) GPW, (d) CNPop

[Pie charts and maps showing population distribution]
Select the three data sets in 2010 to display the results of the comparison of population spatialization (Figure 3), and select the high population density of Shapingba District, the medium population density of Beibei District, and the low population density of Qianjiang District to perform the local population spatialization comparison results (Figure 4). The population density distribution trends of the four data sets are roughly the same. Compared with WorldPop, the population spatialization results of this article are more heterogeneous in population density in high and middle regions, and are closer to the real population distribution, which is reflected in the population differences in each region. The transition is more natural, but the performance is worse in areas with low population density, and overpopulated areas are evenly distributed to each grid. Compared with the GPW and CNPop, the spatial resolution of this paper is significantly higher. Under the same scale, the jagged edges are not obvious. The spatialization results can show more spatial details and pattern differences. At the urban scale, it can better meet the requirements of multiple applications.

Through the comparison of the RMSE, MAE, relative error (RE) three evaluation indicators and the overall spatialization result with the local spatialization result, comprehensive analysis shows that the accuracy of the population spatialization result obtained by this research method is better than the other three data sets. The feasibility of the method in this paper is proved and the expected effect is achieved.

Table 3. Relative error level of different population spatialization.

| Data     | Seriously underestimated | -50%<RE<=-20% | -20%<RE<=20% | 20%<RE<=50% | RE>50% |
|----------|-------------------------|---------------|--------------|-------------|--------|
| This paper | 6          | 28            | 74           | 33          | 19     |
| WorldPop  | 26         | 60            | 33           | 19          | 22     |
| CNPop     | 1          | 15            | 66           | 35          | 43     |
| GPW       | 7          | 37            | 51           | 19          | 46     |

Figure 4. Population distribution maps within a subarea of chongqing in 2010.

Figure 5. Feature importance
4.3. Covariate contribution analysis

Based on the importance of the variable factor of the Xgboost model, the larger the value, the greater the effect of the variable (Figure 5). It can be seen from Figure 5 that the distance from construction land has the greatest impact on population distribution. Night lighting data, distance from green spaces, residential areas, shopping-related POIs, bank POIs and hospital POIs all play an important role in the population distribution of Chongqing. Since Chongqing was established as a municipality directly under the Central Government in 1997, the "Jiefangbei Center Shopping Plaza", one of the earliest commercial streets in China, has been built with Yuzhong District as the center. The vigorous development of commerce has promoted the rapid development of the city and attracted a large number of immigrants. Therefore, construction land continues to expand to meet the needs of rapid urban development, so the distance from construction land has a significant correlation with population distribution. The closer the area to construction land, the denser the population distribution. At the same time, the tourism industry has also been vigorously developed in recent years. The Jiefangbei area not only has a strong commercial atmosphere, but also has a large number of Internet celebrity tourist attractions such as Jiefangbei, Hongyadong, Chaotianmen, and Shiba Ladder. Passenger flow. Therefore, for the needs of development and management, infrastructure has been continuously improved, and lighting has been continuously strengthened. Therefore, night lighting is also an important factor affecting the population distribution of Chongqing. The brightness of the lights at night represents the intensity of human activities and is directly related to population distribution. With the continuous development of cities, the various needs of people continue to increase, and the supporting facilities in the construction area are constantly improving. Banks, hospitals, shopping places, etc. are often built in densely populated areas. Therefore, shopping-related places, banks and hospitals are all important to the population. Distribution has an important impact. Chongqing is known as a "mountain city". The general terrain is high in the southeast and northeast, and low in the middle and west. It gradually decreases from north to south to the Yangtze River valley. Therefore, the population is mainly concentrated in this area, while the population in the northeast and southeast It is relatively sparse, so the slope also has a certain influence on the population distribution.

5. Conclusion and discussion

This paper uses night light data, residential data, elevation data, POI data, road and river data, etc., combined with the Xgboost model and the district density mapping method, to spatialize the 2010 county-level demographic data of Chongqing City with a 100m grid. The root square error and the average absolute error verify the accuracy of the simulation results, and compare and evaluate them with the mainstream raster population data sets at home and abroad, and use the variables of the Xgboost model The importance of factors discussed the influencing factors of Chongqing's population distribution. The specific conclusions are as follows:

1. The root mean square error and average absolute error of this paper are lower than the WorldPop data set, GPW data set and China's kilometer population data set, which proves the effectiveness of this method. And when the spatialization results are compared with the GPW dataset and the Chinese kilometers population dataset, it can reflect more spatial distribution details and pattern differences of the population, and there is no obvious edge jagged under the same scale; compared with the WorldPop dataset, The spatialization result of this article transitions more naturally in the part of population differences, closer to the true population distribution, and can meet the needs of fine urbanization management.

2. The distance from construction land is the most important indicator of Chongqing's population distribution. Night lighting data, distance from green spaces, residential areas, shopping-related POIs, bank POIs and hospitals POIs all play an important role in Chongqing population distribution. Due to the special geographical environment of Chongqing, the slope is also an important factor affecting the population distribution of Chongqing.

Shortcomings and improvements:
The spatialization results in this paper are not sufficiently differentiated in rural areas. The main reason is that the night light data with large covariate contributions cannot capture the light intensity of the country well, and a large number of POI data are concentrated in the main urban area. POI in rural areas is not perfect, so the spatialization results of rural areas are relatively average, and the differentiation is not obvious. Therefore, follow-up research can consider grading modeling of the study area, while exploring the variable factors that affect the population distribution in rural areas.

References

[1] CHEN Y, ZHANG R, GE Y, et al. (2019) Downscaling Census Data for Gridded Population Mapping With Geographically Weighted Area-to-Point Regression Kriging[J]. IEEE Access, 7:149132-149141.

[2] CLARK C. (1951) Urban population densities[J]. Journal of the Royal Statistical Society, Series A (General) 114, no. 4 490-496.

[3] FLOWERDEW R, AND MICK GREEN. (1993) Developments in areal interpolation methods and GIS[J]. Geographic Information Systems, Spatial Modelling and Policy Evaluation. Springer, Berlin, Heidelberg, 73-84.

[4] STEVENS F R, GAUGHAN A E, LINARD C, et al. (2015) Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data[J]. Plos One, 10(2).

[5] BAKILLAH M, LIANG S, MOBASHERI A, et al. (2014) Fine-resolution population mapping using OpenStreetMap points-of-interest[J]. International Journal of Geographical Information Science, 28(9):1940-1963.

[6] YE T, ZHAO N, YANG X, et al. (2019) Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model[J]. Sci Total Environ, 658:936-946.

[7] YANG X, YE T, ZHAO N, et al. (2019) Population Mapping with Multisensor Remote Sensing Images and Point-Of-Interest Data[J]. Remote Sensing, 11(5).

[8] YAO Y, LIU X P, LI X, et al. (2017) Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data[J]. International Journal of Geographical Information Science, 31(6):1220-1244.

[9] ZHAO, LI, ZHANG, et al. (2019) Improving the Accuracy of Fine-Grained Population Mapping Using Population-Sensitive POIs[J]. Remote Sensing, 11(21).

[10] GERVAISONI L, FENET S, PERRIER R, et al. (2018) Convolutional Neural Networks for Disaggregated Population Mapping Using Open Data[C]. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 594-603.

[11] LI X, ZHOU W. (2018) Dasymetric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data[J]. Sci Total Environ, 643:1248-1256.

[12] WEI C Z, TAUBENBOCK H, BLASCHKE T. (2017) Measuring urban agglomeration using a city-scale dasymetric population map: A study in the Pearl River Delta, China[J]. Habitat International, 59:32-43.

[13] PATEL N N, STEVENS F R, HUANG Z J, et al. (2017) Improving Large Area Population Mapping Using Geotweet Densities[J]. Transactions in Gis, 21(2):317-331.

[14] YU B, LIAN T, HUANG Y, et al. (2018)Integration of nighttime light remote sensing images and taxi GPS tracking data for population surface enhancement[J]. International Journal of Geographical Information Science, 33(4):687-706.

[15] CHEN T, HE T, BENESTY M, et al. (2015) Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 1-4.

[16] LLOYD C T, SORICHETTA A, TATEM A J. (2017) High resolution global gridded data for use in population studies[J]. Scientific Data, 4.

[17] CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK - CIESIN
- COLUMBIA UNIVERSITY. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), 2018.

[18] Xu, X. L. (2017). China population spatial distribution kilometer grid dataset. http://www.resdc.cn/DOI. 2017.