Spatial simulation of population in Shandong Province based on night-time imagery and land cover data

Keyi Yang[1,*], Yunling Li[1], and Yang Liu[1]

[1] Shandong University of Science and Technology, Qingdao, Shandong, 266590, China

Abstract. Population spatial data can more truly express the actual distribution characteristics of the population, and provide data support for the regional environment and population development. Use Shandong Province as the research area, township-level census data, revised DMSP/OLS night-time data, and Globaland30 land cover data as data sources, partitions based on population agglomeration, and uses a stepwise regression method to build a population data spatial model. Use the model to simulate population density with a resolution of 100m. The experimental results show: Stepwise regression model good precision, the average relative error was 23.56%, and Root Mean Square Error, Mean Absolute Error are better than the other two public datasets. The simulation results are better than the two public datasets.

1 Introduction

The current population data is mainly obtained through the census or sampling survey, which is summarized level by level by administrative division units, which cannot reflect the spatial distribution of the population. The spatialized population data can provide key parameters for regional economic and environmental changes research [1].

As a hot issue, the research of population spatialization has developed rapidly, and many scholars have carried out research. Land use data can provide population distribution weight based on its own land type, but it cannot reflect the difference in population distribution between the same land type [2]. The night-time imagery has been proved to have a very high correlation with population distribution [3][4]. The combined use of the two effectively alleviated the problem of indistinguishable differences in population distribution of the same land type, then was later recognized by many scholars [5][6][7] analyze and verify and use in population spatialization modeling. Taking Shandong Province as an example, this article attempts to simulate the spatial distribution of population in a simpler and more computationally efficient way to provide basis and services for regional multi-faceted planning.

2 Area and Data

2.1 Study Area

Shandong Province is located on the east coast of China, as of the end of 2010, the results of the sixth census showed that the total population of Shandong Province was 9,572,719 million [8], accounting for about 7% of the total population of the country.

2.2 Data and Pre-treatment

All data use Asia North Albers Equal Area Conic. The regional scope is the vector boundary of the administrative division of Shandong Province. Raster data are resampled to 100m resolution. The data source and data preprocessing are as follows:

- The land cover data comes from the 2010 version of the global land cover data GlobeLand30 [9], which is reclassified into six types of land types: Farmland, Woodland, Grassland, Waters, Built-up area, and Unused land.

- The DMSP/OLS night-time lights data were downloaded from the National Geophysical Data Center (NGDC), which is the Stable Lights Data of non-radiative calibration at night in 2010.This data had Digital Number (DN) values ranging from 0 to 63 and the spatial resolution is about 850m. According to the DMSP/OLS night light data correction method proposed by ZhuoLi et al. [13], the EVI data is used to correct the night-time data to obtain the Enhanced Vegetation Index Adjusted Night-time Light Index (EANTLI).

- The township census data comes from Tabulation on the 2010 population census of the People's Republic of China by Township [8]. According to official information, some township boundaries were adjusted and population data were matched. In the end, 1908 township units were obtained.

- The MODIS-EVI data is extracted from the Level 3 grid data product (MOD13Q1) of the NASA Data Center. The data time is July and August 2010, and the spatial resolution is about 250 m.

* Corresponding author: nuoeryky@163.com

- The township boundary data comes from the township cadastral boundary in the administrative division of TianDitu (http://www.sdmap.gov.cn/), and the overall current situation is 2015.

- The 2010 WorldPop dataset with 100m resolution [11] and China's population spatial distribution kilometer grid dataset [12].

3 Methodology

3.1 Mapping Population

In this paper, referring to the classification standard of population agglomeration in China [14] and research needs, the study area is divided into 5 subareas (Table 1) to achieve the purpose of reducing regional differences and improving simulation accuracy.

$$JDD_i = \frac{(P_i/P_n)\times 100\%}{(A_i/A_n)\times 100\%} = \frac{P_i/A_i}{P_n/A_n} \quad (1)$$

Where JJD_i is the population concentration of i township; P_i is the population of i township(person); A_i is the land area of i township (km^2); A_n is the land area of the province (km^2); P_n is the total population(people) of the province.

Table1. Subareas of Shandong Province.

Subareas		Agglomeration	Township (number)
Sparse area	A	<0.5	258
Mean area	B	0.5-2	1287
Dense area	C_1	2-4	126
	C_2	4-15	126
	C_3	>15	111

Using GIS software, the number of lighting pixels and lighting values of the six types of land use types are counted according to the township, expressed as the area of lighted area NL, the area of non-lighted area NU, and the total light value LE. With the population data of the township census as the dependent variable, the area of lighted area NL, the area of non-lighted area NU, and the total value of light LE of each land use type are used as independent variables, and stepwise regression modeling is carried out. Keep the independent variables that are positive and significant at the 0.05 confidence level. And set the constant of the regression model to zero to satisfy the objective fact that there is no land and no population. The final model is:

$$P_i = \sum_{j=1}^{M}\left(A_j \times NL_{ij} + B_j \times NU_{ij} + C_j \times LE_{ij}\right)_{ij} \quad (2)$$

$$P_{ijk} = \sum_{j=1}^{M}\left(A_j \times NL_{ijk} + B_j \times NU_{ijk} + C_j \times LE_{ijk}\right)_{ijk} \quad (3)$$

In the two formulas, P_i is the population; *NL* stands for number of lit pixels and *NU* for the number of unlit pixels; *LE* denotes light emission; i is the index of different townships, j of various land cover types and k of different pixels; A, B and C are coefficients for each land cover type; and M is the count of different land covers.

$$P'_{ijk} = P_{ijk} \times \frac{\overline{P_i}}{P_i} \quad (4)$$

Where: P'_{ijk} is the final population of the grid; P_i, $\overline{P_i}$ are the simulated population and statistical population of the i-th township.

In the process of model construction, the independent variables and corresponding model coefficients required for modeling can be obtained through equation (2); the model coefficients can be substituted into equation (3) to obtain the simulated population on the pixel scale; use the township boundary Statistical formula (4) simulates the population at the pixel scale and compares it with the census population data to adjust the population on each pixel so that the final simulated population distribution data is consistent with the census data.

3.2 Accuracy Assessment

The simulated population and public data sets are counted by towns. The Relative Error (RE) between the simulated data and the census data is calculated; the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) of the simulated results and the two public data sets are calculated. The accuracy of the model is evaluated through these three errors, and the relevant formula is as follows:

$$RE = \frac{P'_i - P_i}{P_i} \times 100\% \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P'_i - P_i)^2}{N}} \quad (6)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P'_i - P_i| \quad (7)$$

Where P', P is the simulated and census population; i is the index of different townships; N is the count of townships.

4 Results

4.1 Population Density

The model results (Table 2) and the correlation analysis between land cover types and population (Table 3) are shown in the table.

Take the Shandong Province township division map as the base map and combine the census population data to form a population density distribution map with townships as the unit (Figure 1a). According to the above model, the population distribution simulation is carried out, and the simulation distribution map of the township population in 2010 is obtained (Figure 1b). The figure shows that the population density map is restricted to administrative divisions and cannot reflect the spatiality of population distribution. The spatialized population

distribution data effectively avoids the restrictions of administrative boundaries, and the population distribution is more realistic: The population density changes within each township are better reflected, and the population is mainly concentrated on the Built-up area; The population is mainly concentrated in urban cents, which is significantly higher than the population density in rural areas, and gradually decreases to the surrounding area; The northern coastal areas are mostly affected by low-lying terrain, and most of them are saline-alkali land. The central and southern areas of Shandong and the peninsula are mostly affected by the terrain, mainly mountains and hills, which have a greater impact on population distribution and lead to sparse population distribution. To sum up, the population space simulation has a good effect on the detail description, which is difficult to show in the population density map drawn by the administrative boundary.

Table2. Population model in each subarea.

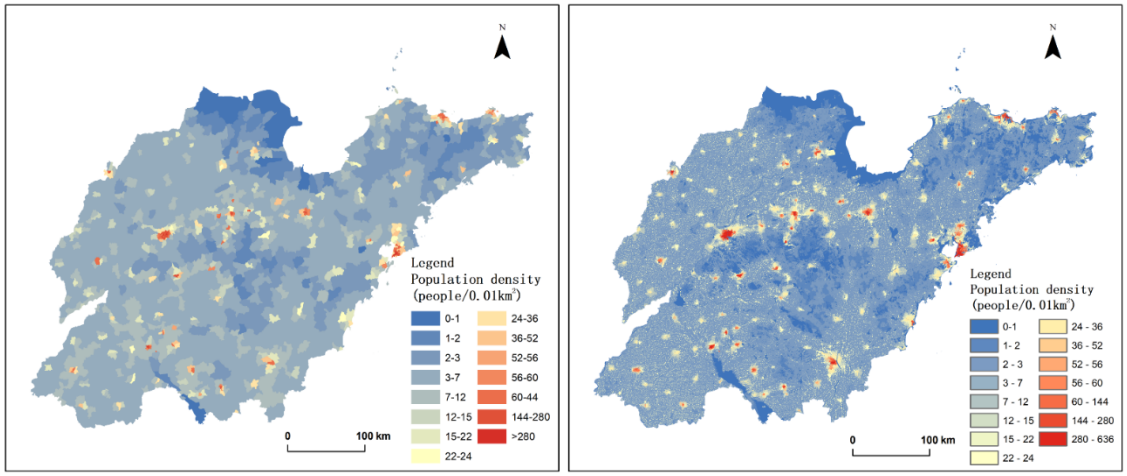
Subarea	Model	R^2
A	$0.048 \times LE_b + 3.053 \times NL_f + 2.027 \times NU_f$	0.915
B	$23.290 \times NL_b + 3.036 \times NL_f + 4.842 \times NU_f + 0.017 \times LE_f$	0.941
C_1	$24.034 \times NL_b + 12.133 \times NL_f + 0.025 \times LE_f$	0.959
C_2	$46.414 \times NL_b + 0.022 \times LE_b + 13.774 \times NL_f$	0.924
C_3	$115.087 \times NL_b + 0.055 \times LE_b$	0.920

b stands for Built-up area; *f* stands for Farmland.

Table3. The correlation between land use and population.

subarea	Farmland	Woodland	Grassland	Waters	Built-up area	Unused land
A	0.720**	0.173**	0.157**	0.015	0.288**	0.080
B	0.598**	0.133**	0.154**	0.161**	0.727**	0.109**
C_1	0.798**	0.157	0.181*	0.288**	0.661**	0.114
C_2	0.628**	-0.026	0.216*	0.164	0.877*	-0.026
C_3	0.302**	0.212*	0.098	0.099	0.898**	0.055

**Significantly correlated at 0.01 level (two-sided);*Significantly correlated at the 0.05 level(two-sided).



(a) Population density map (b) Population simulation map

Figure 1. Comparison of population density map and distribution simulation map

4.2 Accuracy Assessment

Calculate the Relative Error (RE) between the simulated data and the census data according to formula (5); calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) between the simulated data and the public data set according to formula (6) (7).

The RE are shown in Table 4. The percentages of the five error intervals are 56.98%, 31.34%, 7.91%, 1.99%, 1.78%, and the average relative error is 23.56%. The RMSE and MAE are shown in Table 5. The RMSE and MAE of the simulated data in this paper are both smaller than the two public data sets. The comparison shows that the accuracy of the population spatial simulation data in this paper is better.

Table4. Relative error distribution interval statistics.

	Error Interval (Number of township)					Average relative error
	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	>0.8	
Simulation results	1087	598	151	38	34	23.56%

Table5. Error statistics.

	RMSE	MAE
Simulation data	14694	10031
WorldPop	23577	13329
Kilometer grid data	29790	18546

5 Conclusion

This paper draws the following conclusions during the process of spatial simulation of the population census at the township level in Shandong Province:

1) The correlation between various land types and population shows that land types such as cultivated land and man-made ground have strong indications for population distribution. Combining with night light data can better simulate population distribution. Spatial data can better reflect the differences in population distribution compared to administrative division statistics.

2) The R^2 of the stepwise regression models established in this paper is greater than 0.90. In the simulation data, there are 1685 towns and towns with a relative error of less than 0.4, accounting for 88.31%. The Root Mean Square Error and Mean Absolute Error of the simulated data are smaller than the two public datasets, which proves that the overall simulation effect is better.

3) The two kinds of population indicative data used in this study cover the whole world, the modeling method is simple and the calculation efficiency is high, it has good applicability and migration, and has certain reference value for generating large-scale population data sets in other regions.

References

1. D.R. LI, X. LI, An Overview on Data Mining of Nighttime Light Remote Sensing. Acta Geodaetica et Cartographica Sinica, 44(6): 591-601, (2015).
2. D.J. Briggs, J. Gulliver, D. Fecht, et al. Dasymetric modelling of small-area population distribution using land cover and light emissions data. Remote sensing of Environment, 108(4): 451-466, (2007).
3. C.D. Elvidge, K.E. Baugh, J.B. Dietz, et al. Radiance calibration of DMSP-OLS low-light imaging data of human settlements. Remote Sensing of Environment, 68(1): 77-88, (1999).
4. P. Sutton, D. Roberts, C. Elvidge, et al. Census from Heaven: An estimate of the global human population using night-time satellite imagery. International Journal of Remote Sensing, 22(16): 3061-3076, (2001).
5. C. Zeng, Y. Zhou, S. Wang, et al. Population spatialization in China based on night-time imagery and land use data. International Journal of Remote Sensing, 32(24): 9599-9620, (2011).
6. Q. GAO, K. ALIMUJIANG. Modeling the population spatial distribution of Tianshan north-slope urban agglomeration based on DMSP/OLS night lighting data. Northwest Population, 38(03):113-120, (2017).
7. M.M. WANG, J.L. WANG. Spatialization of township-level population based on nighttime light and land use data in Shandong province. Journal of Geo-information Science,21(5):699-709, (2019).
8. Population Census Office under the State Council. Tabulation on the 2010 population census of the People's Republic of China by Township. Beijing: China Statistics Press,(2012).
9. J. CHEN, J. CHEN, A.P. LIAO, et al. Concepts and key techniques for 30 m global land cover mapping. Acta Geodaetica et Cartographica Sinica,43(6):551-557, (2014).
10. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information.Version 4 DMSP-OLS nighttime lights time series.https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html, 2021-02-11.
11. WorldPOP. China population 2010. https://www.worldpop.org/geodata/summary?id=24916.2021-02-11.
12. X.L. XU, China's population spatial distribution kilometer grid dataset. Data Registration and Publishing System of the Resource and Environmental Science Data Center of the Chinese Academy of Sciences. http://www.resdc.cn/DOI.2017.
13. L. ZHUO, X.F. ZHANG, J. ZHENG, et al. An EVI-based method to reduce saturation of DMSP/OLS nighttime light data. Acta Geographica Sinica,70(8): 1339-1350,(2015).
14. R.W. LIU, Z.M. FENG, Y.Z. YANG, et al. Research on the spatial pattern of population agglomeration and dispersion in China. Progress in Geography, **29(10)**:1171-1177,(2010).