Research on Spatialization Optimization of Demographic Data based on Reclassification

Wei Zhanxiang, Hou Yunxian

School of Economics and Management, China Agricultural University, Beijing 100083, China

Keywords: population spatialization; nuclear density; reclassification; regression analysis; Fangshan District

Abstract: The optimization research on the spatialization of population data makes it easy to grasp the spatial distribution of population and improve the government management level and service quality. First, the population data and the spatial data are collected to generate a spatial distribution map of the population. Secondly, the nuclear density is extracted by reclassification to generate simulation data of the population distribution. Then, the simulation data of the spatial distribution of the population and the actual regression analysis are corrected. Finally, verify the population distribution of emergency service facilities in Fangshan District of Beijing. The method solves the functional relationship between population nuclear density distribution and actual distribution, and solves the problem of insufficient basic data in population spatialization.

1. Introduction

The accelerated population movement has exerted tremendous pressure on society, resources and environment. Mastering the accurate information of population spatial distribution plays an important role in improving government management level [1]. The spatial distribution of population is the core issue of population geography research [10], and it is an important foundation for the study of human-land relationship [11]. Spatial data modeling of population data, mainly divided into regional density, multiple regression, and multi-factor mixing ideas, etc. [17, 18], Dong Nan and others (2015) proposed that partition modeling can improve accuracy, in addition, model control in partitions facilitate the acquisition of demographic data and speed up the quantification process.

2. Question raised

At present, the population data mainly comes from the statistical data of the administrative department, collected from planar area (administrative division) and statistical data. Due to the low spatial resolution of the data (in units of counties), the population of the study area is inconsistent with its affiliated space unit and the spatial distribution of population aggregate data is hard to reflect the actual situation [2-8]. The existing population spatialization models and methods require high accuracy of population impact factors and geographic data, and it cannot solve the problem of insufficient basic data. Modeling pays too much attention to the number of factors and quantification, which results in information redundancy and complex processes. It is not easy to work out the actual distribution on population spatial distribution, and it is not convenient for quantitative analysis. Studies have shown that residential population as an important factor affecting the spatialization of population it is seldom applied to population spatial analysis [9], especially under situation of obscure boundary. At present, when dealing with the spatialization of population, the general method of “administrative zone with average unit“ is adopted. Through nuclear density analysis combined with demographic data, regression analysis is used to obtain population spatialization optimization data, and the population distribution of emergency shelters in Fangshan District of Beijing is verified.

DOI: 10.25236/iwmecs.2019.024

3. Methods and Models

The spatialization of population data is to use scientific and reasonable indicators of population distribution, to construct mathematical models, to expand the demographic data of the original administrative area into a certain size of geographic grid, and to realize the spatial transformation of population information. Deichmann [14] systematically analyzed the population spatialization methods and summarized to the methods of surface interpolation and surface modeling. Surface modeling is more accurate and suiTable for different geographic levels and becomes the main measurement method. Grid is the basic unit of population spatial information, which determines the accuracy of population spatial data, and is divided into spatialization based on data sources and population statistics [17]. The surface modeling of quantitative analysis of population spatialization consists of three steps: determining standard grid, geographic data entry, and population data gridding [16, 19]. The simulation of population density is the core step of population spatialization.

With the application of econometrics in geography research, geographers applied mathematical statistics methods to the study of population spatial distribution, and created a nuclear density model to explain the distribution characteristics of population density attenuation law [12]. Similar models include Sherat model, "negative exponential model", Schmid model and weight population distribution model, etc., these are mainly used in macroscopic research [13]. The traditional population density analysis method requires a large amount of geographic information and auxiliary data. The calculation process is complex, and the parameters need to be adjusted repeatedly which is not applicable for scenario analysis with simple data while complex in geography, and the simulation results cannot be used for quantitative analysis.

In this paper, the partition density method is used to generate the population distribution by the surface interpolation method [20]. Based on statistical data and adopted the method of GIS reclassification to carry out the regression analysis on the simulated population, and the result is quadratic spatial distribution. The model is solved into the following four steps.

3.1 Data collection and reduction

Combine the residential information (geographical location) with the administrative division layer, and recheck the geographic information through ArcGIS 10.2, then record the population data of the residential area. The unit of population distribution is grid, which can effectively explain the appropriate information of source data and population spatial data, which is conducive [21].

3.2 Nuclear density analysis

Population spatial density was analyzed by ArcGIS 10.2 kernel density analysis, which is a nonparametric density estimation method. The population data distribution is estimated by interpolation function, and the population probability density value can be estimated by bandwidth without assuming prior density. The kernel density estimation principle is to extract X1…Xn, representing mutually independent samples, from the population of distribution function f according to Silverman's definition of kernel density function [22]. Then the estimated value f (x) of the function f at some point x is solvable, and the estimated kernel density of point x is:

$$\mathrm{f}(x)=\frac{1}{\mathrm{n\,h^{d}}}\sum_{\mathrm{i=1}}^{\mathrm{n}}K\left(\frac{\mathrm{x}-X_{i}}{\mathrm{h}}\right) \tag{1}$$

As shown in the above formula, K () at this time represents the kernel function, h is the bandwidth, and x is the radius of the circle; N is the number of points in the bandwidth; D stands for data dimension. (x-Xi) is the distance from the center X to Xi. When the geographic eigenvalue or event is the point corresponding to the two-dimensional data, and when the parameter of d is 2, the kernel density function can be expressed in the two-dimensional space as:

$$k(u)=\begin{cases}\frac{3}{\pi}(1-u^{T}u), & u^{T}u<0\\ 0, & \text{Other}\end{cases} \tag{2}$$

If any point on the research area R is represented by S, $S_1 ... S_N$ is the observed value corresponding to N points. f (s) is used to represent the intensity value at point S, and the further estimated value is denoted as f~(s). Substitute the function into f~(s) to obtain the estimated point density at point S:

$$f^{\sim}(x)=\frac{3}{n\pi\ h^{2}}\sum_{i=1}^{n}\left[1-{d_{ij}^{2}}\Big/{h^{2}}\right] \tag{3}$$

In the above formula, h is the bandwidth of the kernel density function, d_{ij} is the distance from the point S, at which the density is estimated, to the event S_i, the density function has a small effect on the interpolation value, while the bandwidth has a large effect on the density interpolation result [23].

3.3 Data extraction based on reclassification

The spatial population distribution can be obtained directly through the kernel density analysis, but the kernel density analysis tool can only be used to represent the density of pixel point elements within the grid unit, and the density of each output grid pixel is shown as the sum of the values of all the core surfaces of the grid pixel. In general, during kernel density processing, pixel density calculated is multiplied by the corresponding factor and written into the output grid to obtain the graphic simulation effect, which cannot be directly involved in vector calculation, which can’t be solved by ArcGIS software. In this paper, the pixel value is converted by the reclassification tool, the pixel value is decomposed into different groups, and the pixel value in the input grid is converted into vector data, so that the data can be statistically analyzed. 2.4 Population space optimization based on regression

Regression analysis is a common method for data correlation analysis in statistics. It is mainly used to analyze whether there is a correlation between two or more variables and the weight of them, and to test the change rule between the observation variable and the expected variable through mathematical model [25, 26].In this paper, geographical features are fuzzy processed by linear regression method. The idea is to use the simulated distribution data of settlements to conduct regression processing on the reclassified vector data, obtain the functional relationship of the actual data, and calculate the actual population of grid units [27]. The steps are as follows:

(1) Calculation of kernel density. The kernel density analysis of ArcGIS was used to simulate the distribution of residential areas. Through the debugging of different pixel units and bandwidth, the parameters of 75 m grid and 2000 m bandwidth were finally determined as the best value.

(2) Reclassify vector processing. Reclassification is generally divided into 5-15 groups. In general, the more groups, the more accurate the data and the greater the workload. Therefore, researchers can select the corresponding classification criteria according to the target requirements.

(3) Simulation data optimization. Regression analysis was performed on the results of population density, and the simulated data was optimized. The processing steps are as follows: ①Integer data; ②Population distribution data;③Classify and count regional population;④count population;⑤ Determinate regression coefficient and actual population data in grid cells. The regression coefficient optimization formula is as follows:

$$\theta=\frac{\sum_{i=1}^{n}\alpha_{i}}{\sum_{j=1}^{m}d_{j}\times d_{m}} \tag{4}$$

Among them, θ represents the population adjustment coefficient of the grid, n represents the number of towns and villages, and α represents the population of towns and villages;d_m stands for the reclassification group, d_J stands for the median of the groups, and m stands for the number of groups.

(4) Extract the study area. Based on the classification, the group study was carried out, and the range of research objects was extracted -- emergency public shelter in Fangshan district, Beijing.

(5) Calculate the regional population. By the product of pixel number and pixel value, the total population covered by the research object can be obtained, and the corresponding formula is as follows:

$$c_1 = \sum_{i=1}^{n} p_i \times p_{m\,i} \tag{5}$$

Where, C_1 represents the total population of the research object region, p_i represents the classified area of the research object, $p_{m\ i}$ represents the cell represents the population, and n represents the number of core density reclassification groups.

4. Case analysis

This paper selects the population data of villages and towns in Fangshan District of Beijing for spatial analysis to verify the effectiveness of the method in this paper, and applies it to calculate the coverage population of emergency shelters in Fangshan District.

4.1 Emergency shelter in Fangshan District

Located in the southwest of Beijing, Fangshan District covers a total area of 2,019 square kilometers and has a registered population of nearly 800,000 by the end of 2018. From 2009 to 2014, a total of 15 emergency shelters have been constructed, including Fangshan District Stadium, Yan Village Cultural Square and Changyang Group Park. In case of war, earthquake, flood, fire and other natural disasters and emergencies, it is estimated that 215,000 urban residents can be resettled for disaster prevention and risk aversion.

4.2 Emergency service facilities covered population

Geographic spatial data were obtained from GIS database, and population data were obtained from the Statistics Bureau of Fangshan District.

Firstly, geographical data and demographic data of all 459 villages in the administrative region of Fangshan District were selected to generate core density distribution through ArcGIS 10.2. Subsequently, the population in the coverage area of emergency service facilities was calculated according to the population spatial optimization steps mentioned above, and the results in Table 1 were obtained. It can be seen from Table 1 that the actual population of the area covered by facilities is 40,660, accounting for about 6.84% of the total statistical population.

Table 1 Spatial optimization of population density in Fangshan District and estimation of population covered by facilities

Serial number	Reclassified cell volume	Nuclear density unit population	Nuclear population	The actual representative population of the pixel	The amount of pixels represented by the facility coverage area	Facility coverage area actual population
1	194280	74	14376720	0.416830522	460	192
2	56644	258	14614152	1.453273983	808	1174
3	59249	499	29565251	2.810789604	2806	7887
4	40551	745	30210495	4.196469448	2963	12434
5	16432	1017	16711344	5.72860326	3312	18973
Total			105477962			40660

Note: in this paper, core density simulation data were reclassified into 5 groups; then according to formula (4), the actual unit population value represented by pixel is derived; according to formula (5), the number of people covered by emergency shelters is calculated.

4.3 The population covered by emergency shelters in Fangshan District

ArcGIS 10.2 was used to show the coverage effect of 15 emergency shelters in Fangshan District (Figure 1). As can be seen from Figure 1, the distribution of emergency shelters in Fangshan District matches the population density, basically covering the level 3 and level 4 areas with high population density. Only some high-density areas of Dongfeng Street in the north are not covered; there are few areas with high density coverage in Zhangfang Town and Dashiwo Town in southwest area; covering high density area in the middle of Doudian Town and Shilou Town is still insufficient.

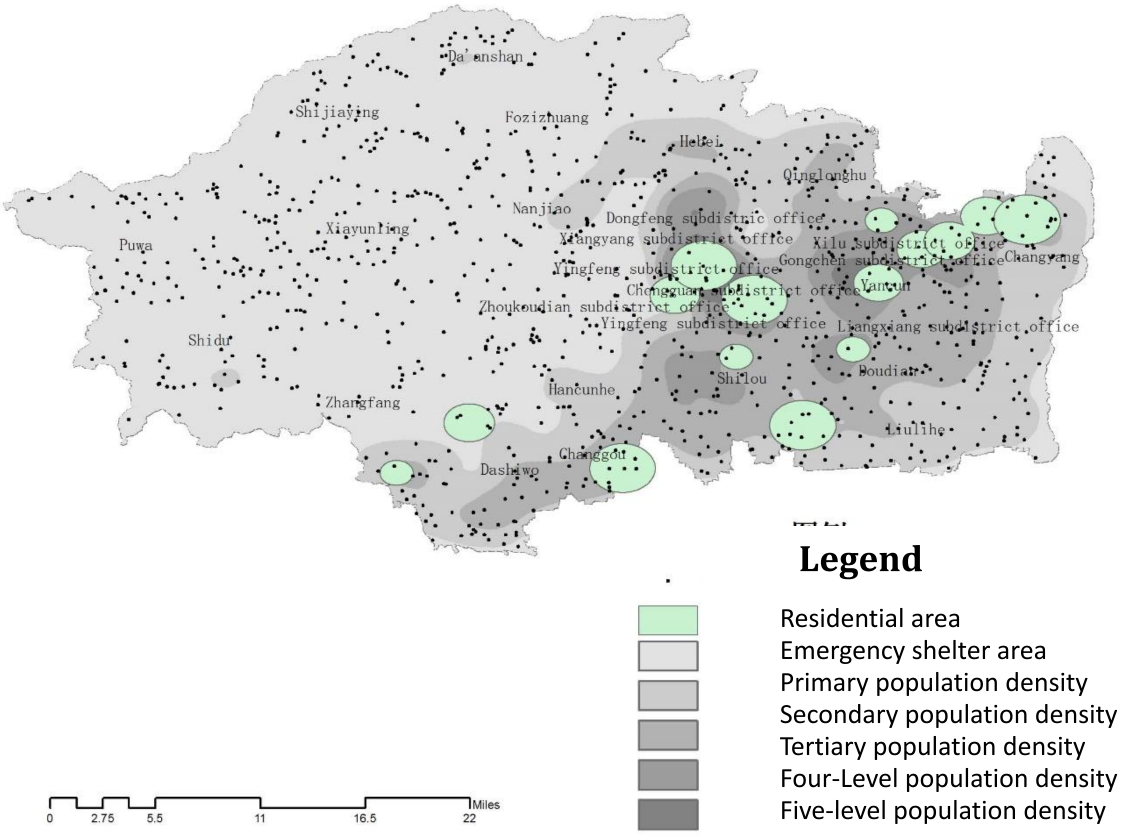
Da' anshan
Shijiaying
Fozizhuang
Hebei
Qinglonghu
Nanjiao
Dongfeng subdistric office
Xiayunling
Xiangyang subdistrict office
Xilu subdistrict office
Puwa
Yingfeng subdistrict office
Gongchen subdistrict office
Changyang
Chengguan subdistrict office
Yancun
Zhoukoudian subdistrict office
Yingfeng subdistrict office
Liangxiang subdistrict office
Shidu
Doudian
Shilou
Hancunhe
Zhangfang
Liulihe
Changgou
Dashiwo
Legend
Residential area
Emergency shelter area
Primary population density
Secondary population density
Tertiary population density
Four-Level population density
Five-level population density
0
2.75
5.5
11
16.5
22
Miles

Figure 1. Coverage effect of emergency shelter in Fangshan District

5. Conclusions

(1) The existing results of kernel density analysis cannot be directly used for vector calculation, and the reclassified population density optimization method is adopted to solve the functional relationship between kernel density and actual distribution, which can be applied to the quantitative analysis when boundary conditions are missing and basic data are insufficient.

(2) Through the empirical study on the population of emergency shelters in Fangshan district of Beijing, it is proved that the distribution of emergency service facilities in this area is reasonable and basically covers the areas with high population density, but the coverage in the southwest is low, the coverage in the northeast is insufficient, and there is an obvious gap in the central region.

(3) The method in this paper is simple and practical with strong universality, which can be applied to the spatialization of population data with different regional characteristics. Furthermore, the relationship between settlements and the area occupied by land should also be considered.

References

[1] Dyn N. Smooth Pycnophylactic Interpolation for Geographical Regions: Comment [J]. 1979, 74(367): 530-535.

[2] Martin D. Mapping population data from zone centroid locations [J]. Institute of British Geographers Transactions. 1989, 14(1): 90-97.

[3] Langford M, Unwin D J. Generating and mapping population density surfaces within a geographical information system [J]. Cartographic Journal. 1994, 31(1): 21.

[4] Yan Qingwu, Bian Zhengfu, Zhang Ping, et al. Spatialization of population density based on residential density [J]. Geography and Geographic Information Science. 2011, 27(5): 95-98.

[5] Ye Yu, Liu Gaohuan, Feng Xianfeng. Spatial expression and application of population data [J]. Journal of Earth Information Science. 2006, 8(2): 59-65.

[6] Dou kaili. Research on planning support methods for urban emergency shelters for disaster prevention [D]. Wuhan University, 2014.