ATLANTIS PRESS

*Advances in Computer Science Research, volume 77*

8th International Conference on Management and Computer Science (ICMCS 2018)

# Research on Population Distribution Model Based on Real Estate Big Data

Jie Dong[1, a *], Gui Li[2,b] and Liming Du[3,c]

[1] Shenyang Jianzhu University. No.9, Hunnan East Road, Hunnan New District, Shenyang City, Liaoning, P.R.China

[a]623236930@qq.com, [b]1309104987@qq.com, [C]3876187@qq.com

**Keywords:** Real estate data; Population distribution; Matrix model

**Abstract.** Population distribution is one of the important factors that affects social economic vitality, infrastructure construction, public service allocation, transportation, resources and so on. Obtaining high density urban population distribution could lay a foundation for urban population management, adjustment and planning, and optimize people's living environment. Although there are many models of population distribution, few studies considered on the distribution of population in the region. Finding a good model to calculate regional population distribution could provide effective theoretical support for solving practical problems. Based on the data of urban real estate big data integration platform, the population distribution model studied in this paper. Through the results of statistical analysis of questionnaire, selected some important factors of population distribution, and carried on the in-depth analysis to these influencing factors. The weight of each factor in the population distribution model is calculated by constructing an improved matrix model. By comparing and analyzing the results of the research and the actual situation, the improved algorithm is put forward at the same time. The experiment shows that the study of population distribution model based on real estate big data is a powerful supplement to the traditional statistical model.

## Introduction

With the development of urbanization, a large number of floating population is pouring into the city. The number of urban population has increased sharply, such as the deterioration of natural environment, urban traffic congestion, the lag of service facilities in densely populated areas and so on. The historical practice shows that predictable and correct urban population distribution planning is closely related to the people's life, is a major event related to the long-term development of the region and the vast economic hinterland, and is also the basis for building a harmonious society.

Population distribution is the result of natural, social, economic and political factors. As the population economy continues to increase, the pressure on the environment and resources becomes greater and greater. Thus, if a high density of urban population distribution can be obtained, it can effectively lay the foundation for urban population management, adjustment and planning. Thus optimize people's living environment. At present, the urban planning of our country usually only pays attention to the population forecast, but not to the distribution of the population in the region. Population information is usually obtained through population census, population sampling, etc. Although this method can get better statistical results, the method mainly depends on artificial means and manual operation, the statistical period is too long. Lead to excessive time and energy, in terms of human, material and financial waste [1].

There are two kinds of representative studies on population distribution. One is the study on the evolution characteristics of population spatial distribution. Clark simulated the density of urban population distribution in 1951. For the first time, he put forward the model of population density distance attenuation, and considered that with the development of cities, The highest point of population density will move from the city center to the surrounding area, from urbanization to suburban urbanization, then to reverse urbanization, and finally to complete the evolution of re-urbanization [2]. With the continuous development of the metropolitan area, the evolution of population spatial distribution also shows the corresponding characteristics. Zhang Dan, Sun Tieshan and Li Guoping (2012) used the regional density function to analyze the spatial structure of the

employed population in the Beijing-Tianjin-Hebei metropolitan area[3]. Zhang Yaojun and Zhang Zhen (2014) studied the spatial pattern of population in Beijing-Tianjin-Hebei region from 2000 to 2010 by using spatial data exploratory analysis method. The results show that with the increase of Beijing's ability to absorb the population, The population agglomeration of Beijing, Tianjin and Hebei continues to increase, and the population in Beijing region increases by a large margin, resulting in the adverse effects of the city [4]. In order to solve this problem, Li Guoping, Luo Yan (2016), combined with functional positioning, divided the study area into four types of population functional zoning [5]. Different regions formulate different policies to achieve the effect of balanced population development.

The other is the study on the characteristics of population migration and agglomeration: Du Peng, Zhang Wenjuan (2010) defined the ladder flow of population. The floating population constantly changed their living standard by their own efforts and external factors, and presented the phenomenon of upward mobility [6]. Since then, du Peng and Zhang Aviation (2011) have conducted an empirical study on the trapezoidal mobility of floating population in China using the data of the floating population Survey in 2009. The results show that there are three types of ladder mobility in geography, occupation and family [7].

Although there are many models to study population distribution, there are few studies considering population distribution in local areas. In this paper, the research of population distribution model is one of the main features of this paper, which is based on the urban real estate big data integrated platform generated by data capture technology.

## Population Distribution Model Impact Factor Analysis

In order to determine which influencing factors are more important when people choose their residence, a questionnaire is designed to analyze the reliability and disinfection of the questionnaire data. The results of the influencing factors shown in the table below are obtained.

Table 1 descriptive analysis of influencing factors results variables

| Item | average | standard deviation |
|---|---|---|
| traffic convenience | 2.04 | 1.07 |
| distance from downtown | 2.60 | 1.22 |
| distance from place of work | 2.33 | 1.18 |
| **e**ducational support | 2.35 | 1.22 |
| **p**rice of a house | 2.48 | 1.22 |
| **u**nit cost required for travel activities | 2.56 | 1.16 |
| **h**ouse cost performance radio | 2.33 | 1.18 |
| housing area and household structure | 2.39 | 1.14 |
| **h**ousing lighting and ventilation | 2.18 | 1.15 |
| housing water, Electricity, gas, heating, | 2.25 | 1.18 |
| soundproof and privacy of housing | 2.38 | 1.17 |
| area activity place | 2.71 | 1.22 |
| **h**ospital convenience | 2.45 | 1.16 |
| distance from bus station or subway station | 2.28 | 1.08 |
| shopping convenience | 2.23 | 1.10 |
| sanitary and noise environment | 2.37 | 1.14 |
| **g**reen level | 2.34 | 1.12 |
| **r**esidential area Security | 2.42 | 1.21 |
| **r**esidential property Management level | 2.50 | 1.17 |

As showed in Table 1, In order to facilitate observation and study, the above factors were ranked according to the average value. According to the results of the ranking, "the place of activity in the community" and "distance from the center of the city" were ranked 1st and 2nd, respectively. Synthesizing above analysis, according to the questionnaire, we can know that the main factors influencing people's choice of residence are as follows: housing price, management level of residential property, housing area and structure, The degree of transportation convenience and the construction of service facilities around the district.

## Study on Population Distribution Model

**Data Preprocessing.** Due to the need to use the large data platform of real estate to study the population distribution model, the information captured by the original database is more complete if it is used for the purchase of the house, but if the calculation of the population model could not meet the needs of the system, we need to carry out the process of further pretreatment. The original information should be changed to the following structure: Category 1: column 11, column 12,... Column 1n;

For the area around the bus or subway line, it is divided into semicolons.

**Weight Analysis and Calculation of Different Service Facilities.** Service facilities need to use exponential distance attenuation function to calculate service facilities. This paper considers three kinds of service facilities: school, hospital and shopping mall.

$$S = \alpha_1 e^{-\varepsilon_1 D_1} + \alpha_2 e^{-\varepsilon_2 D_2} + \alpha_3 e^{-\varepsilon_3 D_3}. \quad (1)$$

In Eq1 , $D$ denotes the path distance from one service facility to that point; $\varepsilon$ denotes the spatial attenuation coefficient of the service facility, which is set to 1; $\alpha_i$ denotes the weight coefficient of the service facility.

In order to measure the influence factors of each kind of different service facilities objectively and impartially, we still adopted the form of questionnaire, using a 5 grade scoring system, according to the proportion chosen by the investigators. The formula for calculating the weight coefficient of each service facility is as follows Eq2 :

$$\alpha_i = \sum_{j=1}^{5} (m_j * p_{ij}) \quad (2)$$

In Eq2, $\alpha_i$ represent No. $i$ service facility, $j$ stand for the standard of rating, $m_j$ stand for the score of No. $j$, $p_{ij}$is the probability of No. $i$ service facility to the No. $j$ rating.

**Path Distance Calculation.** We calculate the path distance between the community and the service facilities by the Baidu Map API functions, the path distance of the same kind of service facilities is taken as the average.

**Calculation of the Traffic Convenience Degree.** According to the house source database data acquisition. Among them, the traffic convenience is calculated through the number of public transport lines around the district.

**Solving the Coefficient Value of Influence Factors.** In order to obtain the evaluation value of the influencing factors, the population data obtained from the survey are divided into two groups, and the test sets are used to calculate the influencing factor coefficients. The target set is used to test the error between the result of the model and the actual investigation. Use the first group of population data and the weights of the various factors that have been calculated.

Suppose that $R$ is the total population of one region, $n$ is the number of cells in this region, $P = \{p_1, p_2, p_3, \ldots, p_n\}$ is a population vector, $p_i$ is the population of the cell $i$, $n$ is the number of cells, $C$ Stand for the coefficient vector of the influence factor, Any element $c_{ij}$represents the weight of the influence factor $j$ of the cell $i$, $i \in [1, n]$, and any element $m_{ij}$ represents the weight of the influence factor $j$ of the cell $i$, in the influence factor matrix $M$. formulr as follows:

$$P = X * M * R / k \quad (3)$$

**Population Quantity Estimation in Residential Area.** Because it is difficult to obtain the exact sample population data in a large area in a short period of time, we selected Hunnan District of

Shenyang city as the research object to carry on the related algorithm research. Based on the real estate data of China, we collected a amount of houses data. From the statistics, there are 250 residential districts in this area, with a total population of 500000 people. Through the data preprocessing and according to Eq3, the population is calculated, and the following table 2 is part of the data.

Table 2 Comparison between the studied and the calculated

| Cell Number | Research Population | Population Calculation | Error Rate |
|---|---|---|---|
| 1 | 3900 | 4056 | 4% |
| 2 | 11620 | 12899 | 11% |
| 3 | 5280 | 5544 | 5% |
| 4 | 5365 | 5472 | 2% |
| 5 | 1115 | 1193 | 7% |
| 6 | 5178 | 5489 | 6% |
| 7 | 7310 | 7674 | 5% |
| 8 | 5520 | 5597 | 1% |
| 9 | 11230 | 12334 | 7% |
| 10 | 1756 | 1719 | 2% |

It is acceptable to use matrix model to calculate the population number and the error of population quantity is less than 10%, which shows that it is feasible to use matrix model to solve the population distribution.

## Optimization and Improvement of the Population Distribution Model

**Factor Selection is not Enough.** In the reality of life, there are many factors that affect the population distribution, from environmental factors, social factors to policy factors, involving all aspects, and the relationship between the various factors is complex, there may be a situation of mutual influence. However, due to the feasibility of the study, this paper can only analyze and calculate from six factors, and can not take all factors into account, so the calculation of the results will produce errors. In the follow-up study, try to select more factors to analyze, so that the error is minimized.

Because of the majority of the respondents who own a house or do not have a house, the price is chosen as a factor affecting the distribution of the population. However, if you own two or more homes, the price may no longer be a factor in your choice of residence, and the results will change.

Table 3 Comparison of two algorithms

| Cell number | Research Population | Calculation before Optimization | Error Rate before Optimization | Calculation after Optimization | Error Rate after Optimization |
|---|---|---|---|---|---|
| B1 | 5520 | 5597 | 1.48% | 5559 | 0.70% |
| B2 | 11230 | 12334 | 9.83% | 11606 | 3.35% |
| B3 | 1756 | 1719 | 2.10% | 1780 | 1.40% |

**Improvements**. Based on the above reasons, we made an improvement of the algorithm. According to the unit price of the house, we set up different ranges, each range separately calculated the weight of each influencing factor, and applied to the estimation of the population of the community in the range of different selling price. The accuracy of the improved algorithm for cell population prediction is improved obviously,as showed in table 3.

**Summary**

The application of data mining technology based on web is more and more extensive. This subject is based on the data mining technology to carry out the further application of real estate big data grab, using real estate data and sample population data of residential area. Carry on the research of population distribution in the whole region, evaluate the research results and improve the method. It is found that it is feasible to use the improved matrix model to solve the population distribution. Although there are certain errors, but also can be accepted and improved. The study of population distribution model based on real estate big data is a powerful supplement to the traditional statistical model.

**Acknowledgements**

This thesis is funded by the project- Research and Application of Urban Real Estate big data Integrated platform and Value-added Service system ,which is from Ministry of Housing and Urban-Rural Development, the project number is 2017-K8-038.

**References**

[1]M. Li: Journal of natural science, Harbin normal university.vol.33(2017)No.2,p.97. (In Chinese)
[2]R. Geary: The Incorporated Statistician. (1954)No.5,p.115. (In Chinese)
[3]D.Zhang, T.S. Sun, G.P. Sun: Geographical Study. (2012)No.5,:p.900. (In Chinese)
[4]Y.J. Zhang, Z.Zhang: Northwest population. (2013)No.6,p.52. (In Chinese)
[5]G.P.Li, Y.Luo: Hebei Academic Journal. (2016)No.1,p.131.(In Chinese)
[6]P.Du, W.J. Zhang: Population journal. (2010)No.3,p.25. (In Chinese)
[7]P.Du, H.K.Zhang: Population journal. (2011)No.4,p.14. (In Chinese)