Study on Influencing Factors of Population Changes in Dongguan City Based on Principal Component-Regression Analysis

Zhitao CUI
School of Computer and Information,
City College of Dongguan University of Technology
Dongguan , CHINA
136762576@qq.com

Shanle WAN
Dongguan Yeyuying Network Science Co.Ltd

Dongguan, CHINA
ben@datalab.run

Abstract**—**In this paper, 9 economic and social development indicators are selected from the *Statistical Yearbook 2019* of Dongguan. All indicator variables have a significant linear correlation with the permanent population variable. The similarity of variables is distinguished with the hierarchical clustering method. Variable groups are divided into 2 categories and then subjected to dimensionality reduction based on principal component analysis. Two principal components are extracted. Finally, two regression equations are set up according to the results of variable group clustering and principal component analysis, which have good test indicators.

Keywords- influencing factors; SPSS; Pearson correlation coefficient; hierarchical clustering; principal component analysis; regression analysis;

I. Introduction

Since the reform and opening up, Dongguan's economy has developed rapidly with its growth rate and order of magnitudes among the top of major economic cities in Guangdong Province. The demographic dividend is one of the most important factors [1-5]. However, China's natural population growth has entered a low level state since the 1990s. By 2000, the population aged 65 and over accounted for 7% in China, marking its entry into an aging society [6]. The decline in the growth rate of the total supply of labor force combined with the aging population structure will continue to affect the innovation momentum and the potential growth rate of the medium and long-term economy [7]. The formulation of policies conducive to increasing population supply and slowing down the aging process is an urgent task facing all major cities in China.

The importance of labor supply to the manufacturing industry is self-evident as Dongguan is an international manufacturing city. It must work out relevant policies conducive to promoting the supply of labor force and slowing down the aging of population in order to maintain the continuous increase in social innovation momentum and the continuous and stable development of the social economy. To play a positive role in finding an effective policy focus [8], this paper attempts to identify the factors with an important impact on population growth from the economic and social development indicators of Dongguan.

II. Correlation analysis of data indicators

The 9 data indicators herein are selected from the *Statistical Yearbook 2019* of Dongguan City, as shown in Table 1.

Table 1 Interpretation of data indicators

Variables	Variable interpretation
x_1	Regional Gross Domestic Product (GDP), in 100 million
x_2	Annual fixed asset investment
x_3	Number of large-scale industrial enterprises
x_4	Highway mileage
x_5	Number of teachers in regular institutions of higher learning
x_6	Number of teachers in other types of schools
x_7	Number of students in regular institutions of higher learning
x_8	Registered population
x_9	Natural population growth rate
y	Permanent population

SPSS25.0 software is used to calculate the Pearson correlation coefficients between the permanent population and the 9 data variables [9] and analyze influencing factors quantitatively [10], as shown in Table 2.

Table 2 Pearson correlation coefficients between permanent population and 9 data variables and their significant test results

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Pearson correlation coefficient	.831**	.833**	.821**	.881**	.748**	.849**	.696**	.872**	-.488**
sig	.000	.000	.000	.000	.000	.000	.000	.000	.007

There is a significant linear correlation between the permanent population and the 9 data variables at the significance level. The number of teachers and students in regular institutions of higher learning and the natural population growth rate are moderately related to the permanent population, while the other 6 variables are highly linearly related to the latter.

III. Distingguishing the similarity of variable groups

R-type clustering is made in the cluster analysis, and variables are classified based on their similarity. Before clustering, the data must be standardized to overcome the influence of dimensions [11]:

$$x_{ij}^{*}=\frac{x_{ij}-\mu_j}{\sigma_j} \qquad (1)$$

$$\mu_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij} \qquad (2)$$

$$\sigma_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\mu_j)^2} \qquad (3)$$

The clustering dendrogram is shown in Figure 1.

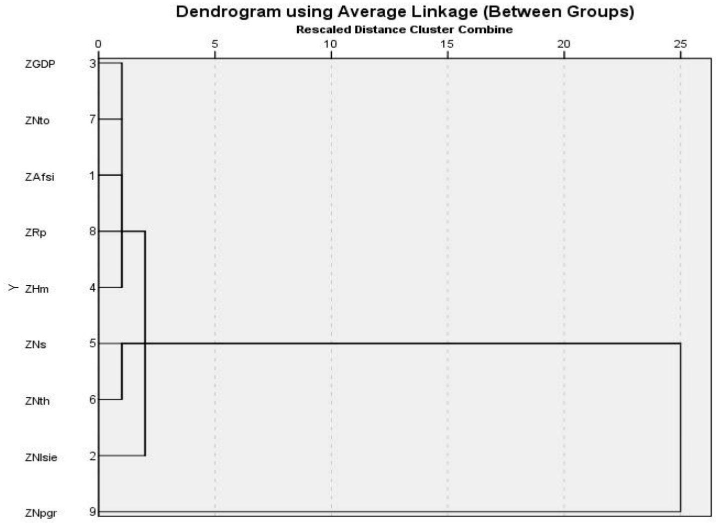


Figure 1 Cluster analysis of 9 data variables

It is appropriate to divide variable groups into 3 categories by the cutting level of clustering [12] according to the clustering dendrogram.

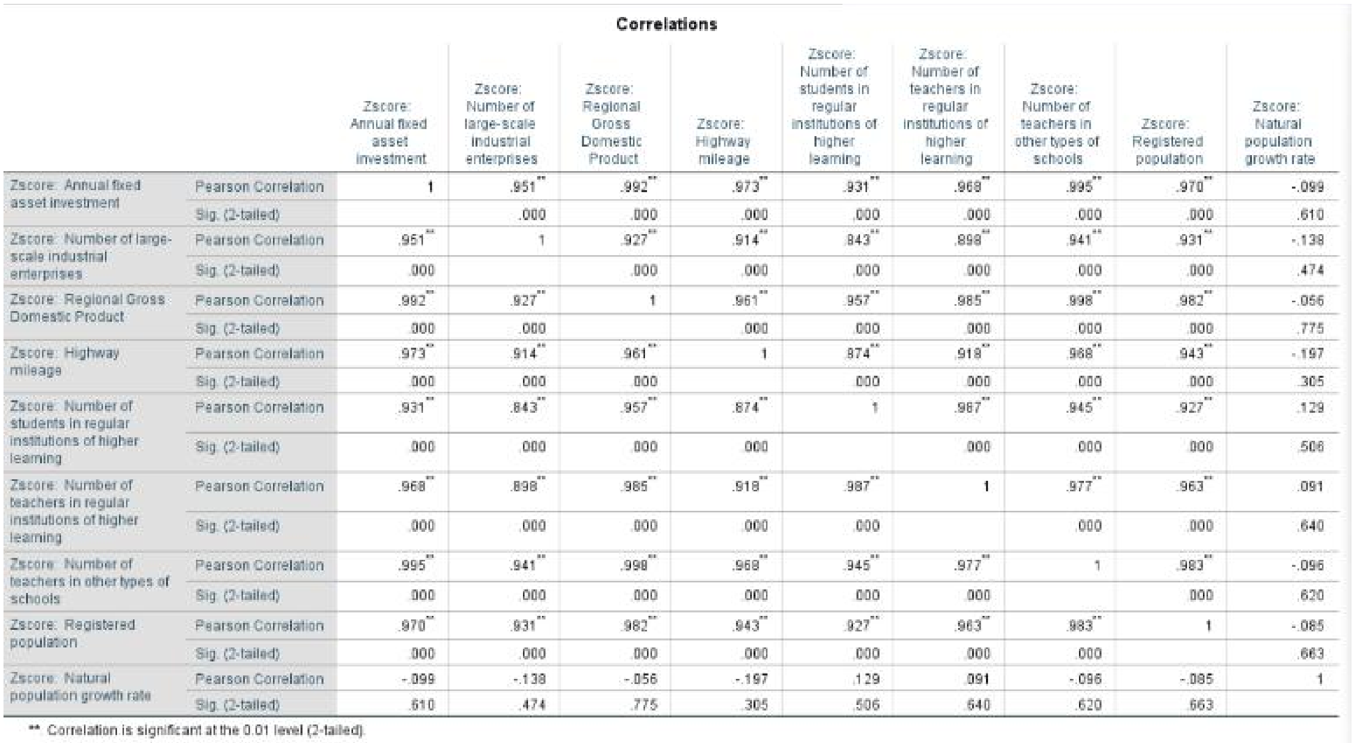
Category I: Regional GDP, annual fixed asset investment, highway mileage, number of teachers in other types of schools and registered population; the variables in this category mainly involve economic development;

Category II: Number of large-scale industrial enterprises, number of teachers in regular institutions of higher learning and number of students in regular institutions of higher learning; large-scale industrial enterprises and institutions of higher learning are similar in the agglomeration of floating population.

Category III: Natural population growth rate.

In order to have a definite understanding of the clustering results, SPSS25.0 is used to give the Pearson correlation coefficients between the 9 data variables, as shown in Table 3.

Table 3 Pearson correlation coefficients between variable groups



Correlations

		Zscore: Annual fixed asset investment	Zscore: Number of large-scale industrial enterprises	Zscore: Regional Gross Domestic Product	Zscore: Highway mileage	Zscore: Number of students in regular institutions of higher learning	Zscore: Number of teachers in regular institutions of higher learning	Zscore: Number of teachers in other types of schools	Zscore: Registered population	Zscore: Natural population growth rate
Zscore: Annual fixed asset investment	Pearson Correlation	1	.951**	.992**	.973**	.931**	.968**	.995**	.970**	-.099
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.610
Zscore: Number of large-scale industrial enterprises	Pearson Correlation	.951**	1	.927**	.914**	.843**	.898**	.941**	.931**	-.138
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.474
Zscore: Regional Gross Domestic Product	Pearson Correlation	.992**	.927**	1	.961**	.957**	.985**	.998**	.982**	-.056
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.775
Zscore: Highway mileage	Pearson Correlation	.973**	.914**	.961**	1	.874**	.918**	.968**	.943**	-.197
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000	.305
Zscore: Number of students in regular institutions of higher learning	Pearson Correlation	.931**	.843**	.957**	.874**	1	.987**	.945**	.927**	.129
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.506
Zscore: Number of teachers in regular institutions of higher learning	Pearson Correlation	.968**	.898**	.985**	.918**	.987**	1	.977**	.963**	.091
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.640
Zscore: Number of teachers in other types of schools	Pearson Correlation	.995**	.941**	.998**	.968**	.945**	.977**	1	.983**	-.096
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000	.620
Zscore: Registered population	Pearson Correlation	.970**	.931**	.982**	.943**	.927**	.963**	.983**	1	-.085
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000		.663
Zscore: Natural population growth rate	Pearson Correlation	-.099	-.138	-.056	-.197	.129	.091	-.096	-.085	1
	Sig. (2-tailed)	.610	.474	.775	.305	.506	.640	.620	.663	

**. Correlation is significant at the 0.01 level (2-tailed).

The variable groups can be divided into 2 categories by their Pearson correlation coefficients;

Category I: Natural population growth rate;

Category II: Economic and social development indicators represented by GDP;

IV. Dimensionality reduction of variable groups

In principal component analysis, the original variables are transformed into a few new variables. The sum of variances of the new variables is as close as possible to that of the original ones. Therefore, the number of variables is decreased, achieving the effect of dimensionality reduction of the data set [13]. Information omission and overlap are taken into account in the process.

The analysis results are shown in Table 4.

Table 4 Kmo test and Bartlett test

KMO value	Approximate chi-square	Degree of freedom	Sig
0.837	738.779	36	0.000

The KMO test value is 0.837 and Bartlett test probability $sig < 0.05$. The data set is appropriate for principal component analysis.

Table 5 Variance explanation of principal component analysis

Principal component factors	Eigenvalues	Variance %	Cumulative variance%
Component 1	7.658	85.084	85.084
Component 2	1.092	12.129	97.213

Two principal components are extracted and the cumulative contribution rate of their variances is 97.213% .

Table 6 Factor loading matrix in principal component analysis

Component matrix

Data indicators	Components	
	1	2
Regional GDP (RMB 100 million)	.997	.014
Annual fixed asset investment (RMB 100 million)	.995	-.037

Number of large-scale industrial enterprises (Nos.)	.947	-.094
Highway mileage (km)	.967	-.144
Number of teachers in regular institutions of higher learning over the years (Nos.)	.982	.165
Number of teachers in other types of schools (Nos.)	.998	-.029
Number of students in regular institutions of higher learning (0,000)	.952	.213
Registered population (0,000)	.984	-.019
Natural growth rate	-.067	.993

Note: Extraction method: principal component

Table 7 Principal component coefficients

	$prin1$	$prin2$
x_1^*	0.3603	0.0134
x_2^*	0.3596	-0.0354
x_3^*	0.3422	-0.0900
x_4^*	0.3494	-0.1378
x_5^*	0.3549	0.1579
x_6^*	0.3606	-0.0278
x_7^*	0.3440	0.2038
x_8^*	0.3556	-0.0182
x_9^*	-0.0242	0.9503

In the first principal component, all variables except the natural population growth rate have higher values and equivalent orders of magnitudes, indicating that the corresponding variables have equivalently great influence on the first principal component. Economic development has brought the demand for population, while the agglomeration of population has promoted economic growth [14-16]. Population growth is both the result and cause of economic development [17].

In the second principal component, the natural population growth rate has the greatest impact.

Calculation formula of principal components:

$$prin1 = 0.3603\times x_1^* + 0.3596\times x_2^* + 0.3422\times x_3^* + 0.3494\times x_4^* + 0.3549\times x_5^* + 0.3606\times x_6^* + 0.3440\times x_7^* + 0.3556\times x_8^* - 0.042\times x_9^* \qquad (4)$$

$$prin2 = 0.0134\times x_1^* \text{-} 0.0354\times x_2^* \text{-} 0.09\times x_3^* \text{-} 0.1378\times x_4^* + 0.1579\times x_5^* \text{-} 0.0278\times x_6^* + 0.2308\times x_7^* - 0.0182\times x_8^* + 0.9503\times x_9^* \qquad (5)$$

x_i^* is the standardized variable of x_i $(i=1...9)$.

V. Regression analysis

Equations used in the regression analysis describe the correlation between dependent variables and explanatory variables. The process includes the least squares estimation of coefficients of regression equations, the significance test of equations and the significance test of regression coefficients.

A regression model of the standardized permanent population variable, standardized regional GDP variable and standardized natural population growth rate is established with the input method in the SPSS regression analysis according to the results of the cluster analysis of variable groups:

Table 8 Test of regression coefficients between permanent population variable and regional GDP and natural population growth rate in the standardized data

Model	Coefficient	t	Sig	VIF
Regional GDP	0.807	12.444	0.000	1.003
Natural population	-0.443	-6.835	0.000	1.003

growth rate

$R = 0.942$ and $R^2 = 0.887$. R2 is adjusted to 0.878. Probability of significance test of the model $sig = 0.000$.

Regression equation:

$$y^* = 0.807 \times x_1^* - 0.443 \times x_9^* \qquad （6）$$

A linear regression equation is established for the standardized permanent population variable and the 2 principal components:

Table 9 Test of regression coefficients between standardized permanent population variable and 2 principal components

Model	Coefficient	t	Sig	VIF
$prin1$	0.303	13.452	0.000	1.000
$prin2$	-0.419	-7.014	0.000	1.000

$R = 0.946$ and $R^2 = 0.895$. R2 is adjusted to 0.887. Probability of significance test of the model $sig = 0.000$.

Regression equation:

$$y^* = 0.303 \times prin1 - 0.419 \times prin2 \qquad （7）$$

Substitute equations (4) and (5) into equation (7):

$$y^* = 0.1036 \times x_1^* + 0.1238 \times x_2^* + 0.1414 \times x_3^* + 0.1636 \times x_4^* + 0.0414 \times x_5^* + 0.1209 \times x_6^* + 0.0188 \times x_7^* + 0.1154 \times x_8^* - 0.4055 \times x_9^* \qquad （8）$$

It can be concluded that the 9 variables are sequenced below based on the influence on the permanent population variable:

$$x_9^* > x_4^* > x_3^* > x_2^* > x_6^* > x_8^* > x_1^* > x_5^* > x_7^*$$

（9）

VI. Discussion of variables in the regression model

The regression model in this paper is explained as below based on existing literature research results. The labor-intensive enterprises represented by the "three-plus-one" trading-mix and "three kinds of foreign-funded enterprises" gathering in Dongguan in the process of reform and opening-up have greatly improved the demand for labor force and contributed to the inflow of floating population [18]. Regional income differences represented by regional GDP and the number of large-scale enterprises are two manifestations of China's population migration mechanism [18]. From the perspective of urban-rural dual economic structure, the agglomeration of production factors and the advantages of production methods in urban areas have facilitated the flow of labor from the low-productivity agricultural sector to the high-productivity industrial sector [19], forming population migration.

VII. Conclusion

In this paper, principal component analysis is used to rank the influence degree of many independent variables which are linearly related to the dependent variables. In view of the dimensionality reduction idea of principal component analysis, this paper uses the cluster method of independent variables. Through the inter group variable correlation analysis, the representative variables are taken from the unrelated variable group for regression analysis, and the dimensionality reduction effect similar to that of principal component analysis can also be achieved. The degree of explanation of independent variables to dependent variables of the two methods is almost equal.

VIII. Acknowledgement

Supported by the young teacher development fund of City College of Dongguan University of Technology (2019QJY008Z).

References

[1] Cai Fang. Future demographic dividend - the development of sources of China's economic growth [J]. Chinese Journal of Population Science, 2009, 2 (1): 4-12.

[2] Zhao Weihua. Law of urban population growth in China and its enlightenment [J]. Journal of the Party School of the Central Committee of the C.P.C, 2016, 20(3): 80-85.

[3] Chen Youhua. Demographic dividend and China's economic growth [J]. Journal of Jiangsu Administration Institute, 2008(4): 60-65.

[4] Wang Xiaoqin, Wang Hongmei." Demographic dividend" effect and China's economic growth [J]. Economist, 2007(1): 104-110.

[5] Zhong Shuiying, Li Kui. Summary of research on the relationship between demographic dividend and economic growth [J]. Population & Economics, 2009(2): 57-61.

[6] Yan Yueping, Huang Meixuan, Zheng Yiran. Research on changes and trend of China’s population age structure [J]. Dongyue Tribune, 2021, 42(1): 148-163.

[7] Li Wenxing, Zhang Zhengpeng. Forecast of Guangdong's population trend under the "two-child" policy [J]. Journal of Guangzhou University (Social Science Edition), 2018,17(10):89-97.

[8] Pan Wenxuan. Focus and path selection of the improvement of structural tax reduction policy [J]. Taxation and Economy, 2012(4): 67-71.

[9] Deng Weibin, Zhou Yumin et al. Practical course of SPSS23 statistical analysis [M]. Beijing: Publishing House of Electronics Industry, 2017.

[10] Wang You. Transformation of migration pattern of floating population in Hebei Province and analysis of influencing factors [D]. Hebei Normal University, 2020.

[11] Zhao Jing, Dan Qi. Mathematical modeling and experiments. 4th edition [M]. Beijing: Higher Education Press, 2014.

[12] Wang Xiao'an. Cutting level determination of fuzzy graph clustering [J]. Acta Botanica Boreali - Occidentalia Sinica, 1998, 18(3): 445-449.

[13] He Xiaoqun. Multivariate statistical analysis [M]. Beijing: China Renmin University Press, 2012.

[14] Liu Zhijia, Huang Heqing. Analysis of spatial-temporal evolution characteristics of the interaction between the expansion of construction land and economic and demographic changes in the Pearl River Delta region [J]. Resources Science, 2015, 37(7): 1394-1402.

[15] Wang Zhiyong. Population agglomeration and regional economic growth - A test of Williamson hypothesis [J]. Social Sciences in Nanjing, 2018(3):60-69.

[16] Gao Jian, Wu Peilin. Impact of urban population size on urban economic growth [J]. Urban Problems, 2016(6): 4-13.

[17] Smith. A. The wealth of Nations[M]. New York: Bantam Classics,2003:35-40.

[18] Gao Guoli, Ji Renjun. Research on population migration in the process of regional economic development - Taking the Pearl River Delta region in Guangdong Province as an example [J]. Economic Geography, 1995, 15(2): 76-82.

[19] Kong Weijun. On the influence of rural floating population on urban-rural dual economic structure [J]. Guangxi Social Sciences, 2001(1):129-132.