Study on the Demographic Analysis of Countries and the Construction of Population Health Index

Heng Zhao
Beijing Forestry University, Beijing, China
math_zh@163.com

Abstract. The data in this paper are from the POPULATION Data Bureau of the United States, Statistics Bureau of the European Union, CEIC global database, etc. SPSS software was used to carry out statistical analysis on the population data of about 230 countries and regions in the world from 2002 to 2012. Through the hierarchical cluster analysis, the population data of the Middle East countries are clustered, and it is found that the population distribution of the Middle East countries is relatively uniform. Multiple linear regression analysis was used to explore the determinants of the total American population, and it was found that the population aged 15-64 and the birth rate were significantly positively driven to the total American population. Finally, the population health index was constructed by factor analysis, and the supporting factors were life expectancy health factor and population quantity factor respectively.

Keywords: SPSS; Multiple Linear Regression Analysis; Hierarchical Cluster Analysis; Factor Analysis; Principal Component Analysis; Population Health Index.

1. Introduction

Because the population change of a country has a great impact on the society, economy and policy, the number of population has always been the focus of researchers from all walks of life. In the 21st century, the world is experiencing unprecedented population changes, which include the rapid expansion of population as well as other population trends, such as the regional imbalance of population structure, the impact of modernization and urbanization, the aging of population, etc.

For the age structure of the population, many scholars at home and abroad have used different methods to analyze and study: Chen Baoping used the grey clustering theory to establish the evaluation model of the population age structure, so as to divide the different regions of China into three grades according to the excellent, the medium and the poor; Bai Xuemei et al. Evaluated the age structure of the population by the improved fuzzy comprehensive evaluation model, which can not only describe the degree and ranking of the population, but also once again Zhang et al. Explored the impact of population age structure on rural environment in China, taking biomass cogeneration capacity as an example.

Similarly, many scholars and professional institutions have devoted themselves to the research on the development trend of world population, which is the next hot spot of population study hours. Guo ran et al. Made a deep explanation on the theory and reality of the world population development trend and population transformation, and found that the world population development is still unbalanced, the total population is increasing but the growth rate is decreasing; Hou Yankun combined with data analysis method, based on Leslie model, constructed the world population growth prediction model, which has more mathematical significance; Ren Qiang analyzed it from the perspective of life expectancy evolution The development track of world population.

In recent years, the growth of population is more and more rapid, and the population structure and population development trend around the world are extremely unbalanced. Therefore, how to control population growth and what policies to promote population health have gradually become the next research focus. Based on this, this paper analyzes the population data of about 230 countries and regions in the world from 2002 to 2012. The data are from the U.S. population information bureau, Eurostat, and CEIC global database.
2. Hierarchical Cluster Analysis based on the Population of Countries in Different Regions of the World

2.1 Descriptive Statistics of Global Population by Country, 2009-2012

Due to some limitations, real-time population data cannot be collected, only the annual population data of about 230 countries in the world from 2002 to 2012 are obtained. In order to reserve space, only the global population data from 2009 to 2012 are used for descriptive statistics, in which few countries fail to collect population information due to unknown reasons, and such data are treated as missing values.

The results of descriptive statistics on the population of all continents and countries from 2009 to 2012 are shown in Table 1.

| Year | Region     | Valid sample (State) | Average population | Standard error | Median     | Variance      |
|------|------------|----------------------|--------------------|----------------|------------|--------------|
| 2012 | Africa     | 53                   | 20400643.74        | 3982526.557    | 11451273  | 8.406E+14    |
|      | Asia       | 34                   | 114836534          | 52316304.30    | 18559712  | 9.306E+16    |
|      | Europe     | 48                   | 16946989.08        | 4115635.387    | 5410775   | 8.130E+14    |
|      | Oceania    | 15                   | 2486388.67         | 1536896.966    | 247262    | 3.543E+13    |
|      | Americas   | 44                   | 21754323.50        | 8608382.805    | 3523399   | 3.261E+15    |
|      | Middle East| 13                   | 17025695.92        | 5795025.566    | 7910500   | 4.366E+14    |

The results show that the average population of every continent has an obvious growth trend from 2009 to 2012. Among them, the population is mainly distributed in Asia, America and Africa; the average population in Asia is the largest, reaching 114836534 in 2012, with the standard error of 52316304.30; from the perspective of variance, the average population variance of Oceania countries is small from 2009 to 2012, and the variance value in 2012 is 3.543e + 13, which shows that the population growth of Oceania countries tends to be a comparison. However, the average population variance of Asian countries is large, which may be caused by the uneven population and unbalanced growth of each country.

2.2 Descriptive Statistics based on the Number of People in Different Parts of the World

![Figure 1. The radar map of the population distribution of African countries in 2012](image)

This section will classify the population data of nearly 230 countries and regions in the world, and first draw the population distribution map of each continent and country in 2012 according to the data. Due to the large amount of data, only the radar map of population distribution of African countries in 2012 (see Figure 1) and the fan chart of population distribution of Oceania countries in 2012 are given.
here. It can be seen intuitively from the figure that the population distribution among the countries of each continent is not uniform, even polarized; from the total population data of 2012, Asia still has the largest population, reaching 3904442155, and Oceania is the least, with only 37295830. In 2012, the global population eventually totaled 7014951860 (excluding a small number of missing values).

3. Research on the Influencing Factors of Population in Different Regions and Countries

3.1 Basic Methods and Steps of Multiple Linear Regression

The general form of multiple linear regression model is

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \]  \hspace{1cm} (1)

Among them, \( X_1, X_2, \cdots, X_p \) represents the influencing factors, which are usually controllable or given in advance, and are called explanatory variables or independent variables; \( y \) is the research object, that is, the prediction target, which is called the explanatory variable or dependent variable; \( \varepsilon \) represents the sum of the influences of various random factors on \( y \), which is called random error term, and obeys normal distribution, that is, \( \varepsilon \sim N(0, \sigma^2) \), and \( \beta_i \) is called the regression coefficient of multiple linear regression model; \( P \) is the number of explanatory variables in multiple linear regression model.

If you give a set of observations \( \{(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) : i = 1, 2, \ldots, n\} \) The overall regression model can also be written:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n \]

Or

\[ Y_i = X_i \beta + \varepsilon_i, \quad i = 1, 2, \ldots, n \]

Among them,

\[ X_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip}), \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \]

The matrix representation of the multivariate linear statistical model can be obtained as follows

\[ Y = X \beta + \varepsilon \]  \hspace{1cm} (2)

Among them,

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix} =
\begin{bmatrix}
X_{11} & \cdots & X_{1p} \\
X_{21} & \cdots & X_{2p} \\
\vdots \\
X_{n1} & \cdots & X_{np}
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p
\end{bmatrix}
+ 
\begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]

In general, the least square method is used to estimate the regression model, that is, to minimize the objective function

\[
\frac{1}{n} \| Y - X \beta \|_2^2
\]

Results of parameter estimation
\[ \hat{\beta} = (X^TX)^{-1}X^TY \]

At this point, the estimated regression equation is

\[ \hat{Y} = X\hat{\beta} + \varepsilon \]

Whether the preliminary estimated regression model can objectively reveal the relationship between various factors in the economic phenomenon studied, whether it conforms to the objective law among variables, whether the introduction of the influencing factors is effective, whether there is a linear correlation between variables, and whether the model can be applied should be determined by model test.

### 3.2 Regression Analysis of Population Data from Countries in the Americas

This paper will take the population data of the Americas in 2012 as an example to conduct multiple regression analysis and establish regression equations. The variables in the regression model include: total population (y), 0-14 age group (x1), 15-64 age group (x2), population over 65 years old (x3), urban population (x4), birth rate (x5), neonatal mortality rate (x6), male life expectancy (x7), female life expectancy (x8)

1. Standardized processing of sample data

   Due to the different dimensions of the selected sample data, in order to avoid the error caused by the dimension, it is necessary to standardize the sample data first

   \[ Z_i = \frac{X_i - \mu_i}{\sigma_i} \quad (i = 1, 2, 3, \ldots, 8) \]  

2. Regression analysis

   SPSS software was used to conduct stepwise regression analysis on the total population (y), standardized population of 0-14 age group (z1), 15-64 age group (z2), population over 65 years old (Z3), urban population (Z4), birth rate (Z5), neonatal mortality rate (Z6), male life expectancy (z7), female life expectancy (z8). Finally, the selected variables are population (z2) and birth rate (Z5) of 15-64 age group. Excluding the other six variables, the regression results are shown in Table 2.

| Variable                | Regression coefficient | t value | p value |
|-------------------------|------------------------|---------|---------|
| Constant term           | 28875070.576           | -2.645  | 0.013   |
| Population aged 15-64   | 1.002                  | 345.179 | 0.000   |
| Birth rate              | 0.009                  | 3.239   | 0.003   |
| F value (P value)       | 63319.322 (0.000)      | R²      | 0.999   |

From Table 2, it can be seen that the explanatory variables selected by stepwise regression are significant, and the overall equation is significant, and the fitting effect is also good. Finally, the regression equation is obtained as follows:

\[ y = 28875070 \cdot 0.576 + 1.002 z_2 + 0.009 z_5 \]  

Among them, the regression coefficients of the two explanatory variables are positive, indicating that the population and birth rate of 15-64 age group have a significant positive impact on the total population; while the regression coefficient of z5 is far greater than that of z5, which means that the contribution of population of 15-64 age group to the total population is far greater than the birth rate.

\[ y = 28875070 \cdot 0.576 + 1.002 z_2 + 0.009 z_5 \]  

245
4. Construction of National Population Health Index in Different Regions

4.1 Presentation of Population Health Indicators

Population health index is a set of indicators that can comprehensively diagnose and objectively compare the health status of population development in various countries, and give early warning to some population development status. As early as 2005, the International Monetary Fund clearly mentioned in the world economic outlook that world economic imbalance is an important risk of world economic growth.

This paper will take each country in the world as the unit, select the typical index reflecting the population characteristics to construct the population health index. Specifically, the birth rate, neonatal mortality, male life expectancy and other indicators are selected to form a comprehensive index to measure the population health level of each country through factor analysis, and then calculate the population health index. The advantage of the establishment of population health index is that it can promote the governments of all countries in the world to pay more attention to their own population development and control the population in a balanced and coordinated way. After effectively controlling the population, it will greatly promote the sustainable and coordinated development of economy, society and resources, so as to promote the development of various countries.

4.2 Take Asia as an Example to Construct Asian Population Health Index Model

This paper will take the population data of Asian countries in 2012 as an example to construct the health index model. The original variables include: urban population (x1), birth rate (x2), neonatal mortality (x3), male life expectancy (x4), female life expectancy (x5), and total population (x6). Firstly, whether the data is suitable for factor analysis is tested. The value of kmo test is 0.815 > 0.6, indicating that the data is suitable for factor analysis and the statistics are significant. Using SPSS software and principal component analysis method, all variance is introduced into the matrix, and the least variable is used to explain the variance as much as possible. After that, the orthogonal factor is rotated by the variance maximization method to change the coordinate axis position and redistribute the variance proportion explained by the factor, so that the factor structure is simpler and easier to explain. Finally, two common factors were selected, and the cumulative contribution rate was 82.974%, indicating that factor analysis was successful. According to the rotated factor matrix (see Table 3), the factor load of life expectancy of men and women on the first common factor is relatively large, while the factor load of total population on the second common factor is relatively large. The meaning of each factor should be clear and easy to explain, so the first factor is named as the life span health factor, and the second factor is named the population quantity factor.

Table 3. The rotated factor loading matrix

| SELECTED POPULATION INDICATORS       | COMMON FACTORS 1 | COMMON FACTORS 2 |
|--------------------------------------|------------------|------------------|
| URBAN POPULATION                     | 0.796            | -0.1             |
| BIRTH RATE                           | -0.853           | -0.221           |
| NEONATAL MORTALITY                   | -0.914           | 0.017            |
| MALE LIFE EXPECTANCY                 | 0.904            | 0.051            |
| FEMALE LIFE EXPECTANCY               | 0.956            | -0.075           |
| TOTAL POPULATION                     | 0.001            | 0.992            |

Then, we construct the population health index and calculate the population health index of each country in Asia. In this paper, the proportion of the variance explained by each common factor is used...
as the weight. The variance explained by the first common factor is 65.457%, and the variance explained by the second common factor is 17.517%. The weighted sum of the two common factors of each country in Asia is obtained, and then the population health index of Asian countries is obtained through standardized processing. (See Table 4)

| COUNTRY       | RANK | POPULATION HEALTH INDEX |
|---------------|------|-------------------------|
| CHINA         | 1    | 3.92132                 |
| INDIA         | 2    | 3.55753                 |
| INDONESIA     | 3    | 0.39915                 |
| PAKISTAN      | 4    | 0.18311                 |
| BANGLADESH    | 5    | 0.10505                 |
| JAPAN         | 6    | 0.01847                 |
| PHILIPPINES   | 7    | -0.07998                |
| VIET NAM      | 8    | -0.10530                |
| THAILAND      | 9    | -0.17546                |
| MYANMAR       | 10   | -0.22009                |
| KOREA         | 11   | -0.22900                |
| AFGHANISTAN   | 12   | -0.29339                |
| UZBEKISTAN    | 13   | -0.29355                |
| MALAYSIA      | 14   | -0.29526                |
| NEPAL         | 15   | -0.30089                |
| NORTH KOREA   | 16   | -0.30954                |
| SRI LANKA     | 17   | -0.32370                |
| KAZAKHSTAN    | 18   | -0.33498                |
| LAOS          | 19   | -0.34113                |
| KYRGYZSTAN    | 20   | -0.35890                |
| SINGAPORE     | 21   | -0.36300                |
| TURKMEENISTAN | 22   | -0.36735                |
| GEORGIA       | 23   | -0.37067                |
| ARMENIA       | 24   | -0.37161                |
| MONGOLIA      | 25   | -0.37205                |
| TIMOR-LESTE   | 26   | -0.37423                |
| BHUTAN        | 27   | -0.37909                |
| BRUNEI DARUSSALAM | 28 | -0.37964                |
| MALDIVES      | 29   | -0.38748                |

The results show that in Asia, China and India rank the top two in terms of population health index, which is mainly due to the advantage of the total population; on the whole, the population health index of East Asian countries and Southeast Asian countries is generally higher, while that of Western Asian countries is slightly lower, and that of other small countries is generally lower, which is also related to the comprehensive national strength of Asian countries to a certain extent.
5. Summary

This paper collects the population data of 230 countries in the world from the U.S. population information bureau, Eurostat data, CEIC global database, and makes hierarchical cluster analysis on the population characteristics of different countries in different regions; and carries out regression analysis on their population variables, and obtains the regression equations of population variables in different regions and countries; finally, it constructs a comprehensive population model by factor analysis Population health index of each country in the world.

The results show that the clustering results of population characteristics in different regions and countries are mainly determined by the total population of each country; the variables that have significant influence on the total population of a country are population of 15-64 age group and birth rate, which have a positive driving force on the total population, and the contribution of population in the 15-64 age group is greater; the factors that constitute the population health index are population life-span health factor and population size Factors, namely life expectancy and population size, mainly determine the health level of the population.

However, there are still some deficiencies in the analysis of this paper: it is one-sided to cluster analysis only through the variables of population quantity characteristics, ignoring the influence of population structure; the factors influencing population quantity and health level in various countries and regions are far more than the superficial characteristics of population, and economic, policy, cultural and other factors also have a great influence on population, which are not found in this paper in consideration of. The improvement of the above problems needs more in-depth study in the future.

References

[1] Chen Baoping. Evaluation model of population age structure based on grey clustering theory [J]. Computer and information technology, 2014, 22 (01): 1-3 + 63.
[2] Bai Xuemei, Zhao Songshan, Zhuang Wenwu. Comprehensive judgment of population age structure by fuzzy mathematics [J]. Population research, 1995 (04): 64-68.
[3] Yifei Zhang, Peng Wang. Effect of population age structure on capacity for biomass cogeneration: A case study of a rural area in China. Building and Environment. 2020(170): 106601.
[4] Guo ran, Wang Jun. world population development trend and population transformation: Theory and reality [J]. Population and society, 2019, 35 (03): 52-63.
[5] Hou Yankun. Population growth prediction and research based on Leslie model [J]. Science and technology innovation and application, 2014 (27): 35-36.
[6] Ren Qiang. Evolution of life expectancy of world population in recent 50 years [J]. Population research, 2007 (05): 75-81.
[7] Fang Kaitai. Cluster analysis (I) [J]. Practice and understanding of mathematics, 1978 (01): 66-80.
[8] Qi Mingzhu. Research on the construction of world population development health index [J]. Population and economy, 2010 (03): 7-12.
[9] International Monetary Fund. World economic outlook (establishment of the system in September 2005). Beijing: China finance press, 2006.