A prediction model for population change using ARIMA model based on feature extraction

Wei Li¹, Zhongyu Su and Pengcheng Guo
School of Science, Dalian Ocean University, 116023 Dalian, China

¹Email: wwli@dlou.edu.cn

Abstract. Population is a social entity with complex contents and a variety of social relations. It includes gender, age and natural composition, as well as a variety of social composition, social relations, economic composition and economic relations. It is difficult to analyze the population change in large number and high dimension over the years in China. A prediction model for population change using ARIMA model based on feature extraction is proposed. It provides the reduction of population change indicators and predicts the population change in the future. Principal component analysis is firstly used to transform this problem in high-dimensional space into low-dimensional space, making the problem simple and intuitive. These less comprehensive indexes obtained are not related to each other and provide most information of the original indexes. The data of the national population change from 1979 to 2015 is effectively reduced to one principal component. On the basis of the analysis results, a prediction model using ARIMA model is established to predict the population change in the next few years.

1. Introduction
In the era of big data, researchers are faced with huge challenges in processing massive data and unstructured data [1, 2]. Population is a social entity with complex contents and diverse social relations. The large population and large increase in population have put enormous pressure on resources, environment and economic development. The sustainable development of China's social economy largely depends on the solution of the population problem, including the problem of quantity [3]. It is difficult to analyze population changes in the large number and high dimension over the years in China without effective feature extraction. As it is difficult to construct prediction model for the population data with 8 indicators in this paper, the principal component analysis method is used to effectively simplify several indicators reflecting the national population change into one principal component. On the basis of principal component analysis, the ARIMA (Auto Regressive Integrated Moving Average) prediction model is established to predict the population change in the next few years.

This paper consists of the following four parts. Section 1 is the introduction including data background and research significance. The second part is the theoretical introduction of the combined method, including principal component analysis and ARIMA model. In section 3, a prediction model for population change using ARIMA model based on feature extraction is established. Section 4 is the summary.

2. Method
Based on the evaluation results, the ARIMA model is established to predict the population change in the next few years. Firstly, principal component analysis is used to reduce the dimension of data indicators
reflecting population change. On the basis of principal component data, ARIMA prediction model is established.

2.1. Principal component analysis
Principal component analysis converts multiple indicators into several comprehensive indexes by using the idea of dimension reduction and minimizing the loss of information [4-6]. Usually called main component, the comprehensive indicators of the transformation to generate each principal component are a linear combination of the original variables and unrelated, and have better properties than the original variables. So only using a few principal components without loss of too much information, can solve the complex high-dimensional problems [7-8].

The algorithm steps are as follows:
(1) select initial analysis variables according to the research questions;
(2) according to the characteristics of initial variables, determine whether the principal component is determined by covariance matrix or correlation matrix;
(3) find the eigenvalues and eigenvectors of covariance matrix or correlation matrix;
(4) get the expression of principal components and determine the number of principal components, and select the principal components;
(5) analyze and deeply study the data with the combination of principal components.

2.2. ARIMA model
ARIMA model, the full name is autoregressive moving average model, which is a time series prediction method. It is a model for fitting stationary series, which is convenient for analyzing the structure and inherent properties of data, optimal prediction and controlling in the sense of minimum variance [9-10].

The ARIMA model is very simple, requiring only endogenous variables and no exogenous variables [11]. Compared with the BP neural network, the ARIMA model is more convenient to operate, while the BP neural network is relatively complex. As long as the ARIMA model based on SPSS is given data, it can quickly calculate the only certain result without repeated operations to make the result converge [12]. Compared with GM (1,1) model, the fitting effect of ARIMA model is better [13]. ARIMA model is a kind of prediction model based on traditional mathematical theories such as calculus and mathematical statistics. It is also one of the most mature prediction methods for time series data. Its advantage is that it can put the influencing factors of various infectious diseases into time variables, quantitatively express them by comprehensively considering the trend change, cycle change and random interference of time series, and obtain a more perfect model through repeated modifications.

The model formula is as follows:
\[
y_t = \mu + \phi \times y_{t-1} + \ldots + \phi \times y_{t-p} + \ldots + \theta_q \times e_{t-q}
\]

where \( \phi \) is AR coefficient, \( \theta \) is MA coefficient.

The algorithm steps are as follows:
(1) obtain data;
(2) plot the data and observe whether it is a stationary time series; For non-stationary time series, it is necessary to carry out advanced d-order difference operation and transform it into stationary time series;
(3) after the second step, the stationary time series has been obtained. The autocorrelation coefficient ACF and partial autocorrelation coefficient PACF should be obtained for the stationary time series respectively. By analyzing the autocorrelation graph and partial autocorrelation graph, the optimal hierarchy \( p \) and order \( q \) can be obtained;
(4) get the ARIMA model from \( d, p \) and \( q \) above, and then test the model [14-16].

The flow chart of ARIMA model is shown in Figure 1:
3. Results
Firstly, principal component analysis is used to reduce the dimension of national population change data. Based on principal component analysis, ARIMA prediction model is established.

3.1. Principal component analysis
This part is the data of national annual population change. It is from “China statistical summary” 2018 [17]. Population change data are shown in Table 1:

| Year | Population | The increase in the national population | Male | Female | ... | Increase of urban population | Rural depopulation |
|------|------------|----------------------------------------|------|--------|-----|----------------------------|------------------|
| 1979 | 97542      | 1283                                   | 50192| 47350  | ... | 1250                       | -33              |
| 1980 | 98705      | 1163                                   | 50785| 47920  | ... | 645                        | -518             |
| 1981 | 100072     | 1367                                   | 51519| 48553  | ... | 1031                       | -336             |
| 1983 | 103008     | 1345                                   | 53152| 49856  | ... | 1309                       | -273             |
| ...  | ...        | ...                                    | ...  | ...    |     | ...                        |                  |
| 2013 | 136072     | 668                                    | 69728| 66344  | ... | 1929                       | 1261             |
| 2014 | 136782     | 710                                    | 70079| 66703  | ... | 1805                       | 1095             |
| 2015 | 137462     | 680                                    | 70414| 67048  |     | 2200                       |                  |

3.1.1. Extraction factor. The method for solving the factor solution is the principal component analysis method, as shown in Table 2:

| Composition | Initial eigenvalue | Extract the sum of squares of the load |
|-------------|--------------------|---------------------------------------|
|             | Aggregate | Percentage variance | Cumulative percentage | Aggregate | Percentage Variance | Cumulative percentage |
| Population  | 6.198     | 77.481               | 77.481                  | 6.198     | 77.481               | 77.481                |
| The increase in the national population | .750 | 9.370               | 86.851                  | .546 | 6.827               | 93.678                |
| Male        | .375     | 4.688               | 98.366                  | .375     | 4.688               | 100.000               |
| Female      | .128     | 1.602               | 99.968                  | .128     | 1.602               | 100.000               |
| Increase of urban population | .003 | .032               | 100.000                  | .003 | .032               | 100.000               |
| Rural depopulation | 1.513E-6 | 1.891E-5               |                            | 1.513E-6 | 1.891E-5               |                            |
3.1.2. Screen plot. As shown in Figure 2, when the eigenvalue of the second principal component arrives at a steep inflection point, the eigenvalue of the first principal component is obviously different from that of the others. From the second principal component, the eigenvalue is almost flat. According to Table 3, the eigenvalue of the first principal component contains 77.481% of all information and the majority of all information. Therefore the number of principal components is determined as one. In this way, only one principal component can represent most information of all the original data indicators [18].

![Screen plot](image)

**Figure 2.** Principal component screen plot.

3.1.3. Score of each component. The scoring matrix of each component is calculated with SPSS, and the output is shown in Table 3:

| Composition | Composition 1 |
|-------------|--------------|
| Population | .095 | The urban population | 0.157 |
| The increase in the national population | -0.149 | The rural population | -0.128 |
| Male | 0.153 | Increase of urban population | 0.140 |
| Female | 0.152 | Rural depopulation | 0.152 |

| Year | Factor 1 | Composition 1 | Year | Factor 1 | Composition 1 |
|------|---------|--------------|------|---------|--------------|
| 1979 | -1.20476 | -3.00 | ... | ... | ... |
| 1980 | -1.35720 | -3.38 | 2011 | 1.29156 | 3.22 |
| 1981 | -1.26744 | -3.16 | 2012 | 1.34098 | 3.34 |
| 1982 | -1.21890 | -3.03 | 2013 | 1.32733 | 3.30 |
| 1983 | -1.26316 | -3.14 | 2014 | 1.30759 | 3.26 |
| 1984 | -0.80666 | -2.01 | 2015 | 1.55424 | 3.87 |

**Table 3.** Factor score coefficient matrix.

The score of each principal component on each index is the weight. The expression of the principal component is as follows:

First principal component:

\[ PRIN1 = 0.095 \times X_1 - 0.149 \times X_2 + 0.153 \times X_3 + 0.152 \times X_4 + \ldots + 0.140 \times X_7 + 0.152 \times X_8 \]

3.1.4. Score of principal components of each variable is calculated. The score of principal component 1 is obtained by multiplying the value of factor 1 by the arithmetic square root of each variance shown in Table 4:

**Table 4.** Score of principal components of each variable.
3.2. ARIMA model result

ARIMA model, the difference auto-regressive integrated moving average model, also known as the auto-regressive integrated moving average model (mobile can also be referred to as the sliding), is one of time series prediction method. By principal component analysis, the data of national population changes over the years are effectively reduced to one principal component. Based on the above result, the ARIMA model is established to predict population changes in the next few years [19-20].

In this paper, the difference algorithm is applied to obtain the stationary sequence. The significance of stationary is to make the fundamental characteristics of time series unchangeable with time. The original sequence is shown in Figure 3. The stationary sequence with third-order difference is shown in Figure 4:

![Figure 3. The original sequence.](image1)

![Figure 4. Stationary series.](image2)

The model is identified as ARIMA (p,3,q). Through adjusting the sequence autocorrelation function diagram and partial correlation function diagram, it is identified as ARIMA (2, 3, 4), as Figure 5:

![Figure 5. Graph of autocorrelation function (left) and graph of partial autocorrelation function (right).](image3)

According to SPSS analysis results, R squared reaches 0.948, which is well fitted, as shown in Table 5:

| model                                    | Number of predictors | Model fitting statistics | Ljung-Box Q(18) | Number of outliers |
|------------------------------------------|----------------------|--------------------------|-----------------|-------------------|
| Household consumption level - model 1    | 1                    | .882                     | .948            | 9.567             | .654              | 0                 |
Figure 6 shows the autocorrelation function graph (ACF) and partial autocorrelation function graph (PACF) after third-order difference. Where the autocorrelation function is the function formed by the auto-correlation sequence with the lag period k as the variable, the partial autocorrelation function is the function formed by a certain coefficient in the sequence and the corresponding k. It can be seen from the Figure that both ACF and PACF are stable.

\[
\Delta X_t = 0.206 -1.491 \times \Delta X_{t-1} -0.918 \times \Delta X_{t-2} + \epsilon_t + 0.618 \times \epsilon_{t-1} + 0.714 \times \epsilon_{t-2} \\
+ 0.621 \times \epsilon_{t-3} -0.988 \times \epsilon_{t-4}
\]

where \(\Delta X_t\) represents the stationary sequence after the difference.

According to the known national population change time series data in recent years, the ARIMA model is established to predict the population change trend in the next few years, as shown in Figure 7.

The abscissa 1-41 represents 1979-2019. The ordinate represents the values of a principal component extracted from the population change data in the previous years. The black vertical line represents the prediction part. The blue bold line represents the predicted value. UCL and LCL represent control on-line and control off-line respectively.

From the trend of the curve, it can be seen that the population shows an upward trend in the previous years, which is quite obvious. With the change of time, the upward trend gradually levels off. The population shows a downward trend in the next few years.
4. Summary
In this paper, a prediction model for population change using ARIMA model based on feature extraction is provided to predict the population change in the next few years. Since eight indicators reflect the changes of the national population, it is difficult to build the ARIMA model to predict the changes in the next few years. Therefore, the principal component analysis is first used to effectively reduce the national annual population change data into a major component. A principal component is extracted to build the ARIMA model to predict the national population change in the next few years.

According to the survey conducted by the national statistics bureau [21], the population of China in 2016 and 2017 is known to be 1382.71 million and 13390.08 million respectively, showing an upward trend compared with that in 2015, which is consistent with the predicted curve in this paper. According to the future population development trend prediction of 2016-2030 China national population development plan, due to the decrease of the number of women of childbearing age and the increase of death rate caused by the aging population, the potential energy of population growth weakens and finally shows a downward trend, which is basically consistent with the prediction results in this paper. In the future, the relations between every indicator and prediction results will be studied deeply.

Acknowledgments
This research is financially supported by the Natural Science Foundation of Liaoning Province of China (20170540103).

References
[1] Wang D, Shen H, Truong Y 2016 Neurocomputing 190 25-34
[2] Pan Q 2017 The financial times 672(9) 19-20
[3] Zeng Y 2004 China population analysis [M]. Peking University press 9(1)
[4] He X Q 2015 Multivariate statistical analysis (fourth edition) [M]. Renmin university of China press 3 113-127
[5] Song W H, Zhang Q 2014 Journal of huangshan college 16(05) 20-22
[6] San W P, Zeng X Q 2016 Signal reconstruction and bearing fault identification based on deep belief network[J]. Electronic Design Engineering 24(4) 67-71
[7] Su W H 2000 Study on the theory and method of multi-index comprehensive evaluation [D]. Xiamen: xiamen university
[8] Guo Y J 2002 Comprehensive evaluation theory and method [M]. Beijing: science press, 111-117
[9] Box G E P, Jenkins G M, Reinsel G C 1994 Journal of Marketing Research 14(2) 353-412
[10] Fildes R A 1976 The analysis of time series: Theory and practice: by C. Chatfield, Chapman and Hall, London(1975) [J]. Long Range Planning 9(6) 113-114
[11] Yang C J 2018 Stock price forecast based on time series model[J]. West Leather 40(12) 98-99
[12] Chen Q Z 2014 Friends of Accounting 34 25-28
[13] Zheng L, Liu D J, Xu X 2015 Practical Preventive Medicine 03 365-368
[14] Peng Y 2014 Introduction of ARIMA model [J]. Electronics World 10 259
[15] Fang K H, Zhao L 2018 Journal Of Sichuan University Of Arts And Science 28(05) 7-11
[16] Li G Q, Huang L Y, Tan F Z 2019 Application of differential integrated moving average autoregressive model in surveillance of influenza-like cases in hospital[J]. Journal of capital medical university 40(02) 286-291
[17] Summary of Chinese statistics 2018 http://www.yearbookchina.com/naviBooklist-YCDES-0.html, 2018
[18] USORO A E, MOFFAT I U 2015 Principal component analysis of nigeria value of major imports [J]. Am J Econ 5(5) 508-512
[19] Pan X F, Peng X X 2016 Time series analysis [M]. Beijing: Tsinghua university press 40-62
[20] Xiao Z H, Guo M Y 2012 Time series analysis and SAS application [M] 44-65
[21] National Bureau of Statistics. http://www.stats.gov.cn/tjsj/zxfb/201802/t20180228_1585631.html