The application of ARIMA model in forecasting population data

Jie Dai and Shuping Chen

College of Mathematics, Xiamen University of Technology, Xiamen 361024, China

1Email: Shupingchen@126.com

Abstract. The prediction of population data plays a positive role in adjusting population policy and promoting the development of social, economic and cultural undertakings. In this paper, the population data of Zhejiang Province from 1978 to 2016 are processed and analyzed by using the principle of time series analysis. In order to test the validity of the parameters selected by this model, the population data of Zhejiang Province in 2017 are verified in this paper. The experimental results show that the actual fitting effect of the model is good. Finally, with the help of SAS software, the population data of Zhejiang Province from 2018 to 2022 are analyzed and forecasted.

1. Introduction

As a country with a large population, the population problem has always been one of the important factors affecting China's economic and social development. As we all known, the population of a country has a significant influence on its political, economic and social development of the country, and the study of the social issues is also closely related to the number and structure of the population. The prediction of population data is conducive to grasping the future trend of economic development and better allocation of existing resources. Therefore, the prediction of population data can effectively control the growth and decline of the population. It can not only provide rich labor resources for all walks of life, but also prevent the negative effects of population aging and decay.

In recent years, many scholars at home and abroad have done a lot of research work on the prediction and analysis of population data. Based on the revised population data for 2006-2015, He et al. [1-3] predicted the population by establishing a time series model and a two-layer "small world" model, and obtained the total population from 2017 to 2030. They also discussed the impact of China's population on economic development by data envelopment analysis. In reference [4], two kinds of population prediction models were established by using ARIMA model based on population time series and exponential smoothing method, and the optimal models were also obtained. This model was used to estimate the population of China from 2006 to 2015. In addition, in order to study the application of various models in the prediction of the income gap between urban and rural residents in the country, a variety of prediction models, such as ARIMA model, grey model and quadratic polynomial were established in reference [5], and the accuracy of the models were also compared. In the same year, Tu et al. [6] used the ARIMA product seasonal model to predict and analyze the tourism demand of Guilin by making a difference in the number of tourists in Guilin, and carried out diagnostic tests on the model. It was found that the product season model has a good fitting effect on the number of tourists in Guilin.
In 2014, Han [7] analyzed and predicted the population of our country by using multiple linear regression model and time series model, the results indicated that China's population will continue to grow in the short term. In the same year, Mumbare et al. [8] carried out time series analysis of the average surviving children during terminal contraception. They preprocessed the data and tested the stationarity of each sequence, and obtained the non-seasonal ARIMA (p, d, q) model. In 2016, Chen et al. [9] predicted the population quantity and structure by time series and the death changes of the corresponding age groups, and analyzed the ARIMA model based on the static data of the sixth population census of Q City.

In reference [10], using the time series data from 2002 to 2016, the practicability of price prediction was demonstrated for the main crops in Karnataka. The results of ARIMA price forecasting fully proved the power of ARIMA model as a price forecasting tool. In reference [11], a linear stochastic model was established by using ARIMA model. The predicted data of the optimal ARIMA model was compared with the measured data to verify the effectiveness of the model, and the results were in good agreement with the actual data.

As a strong economic province, the literatures on the study of the population quantity and structure of Zhejiang Province is not many. Feng et al. [12, 13] used the autoregressive distribution lag model and equal-dimensional grey number supplement model to predict the population data of Zhejiang Province. In this paper, the ARIMA model will be established according to the total population of Zhejiang Province from 1978 to 2017, and the model will be used to predict the population number and structure in the next few years.

2. ARIMA model

There are two methods to test the stationarity of sequences, one is to judge the stationarity of sequences according to the characteristics of sequence diagram and autocorrelation diagram, and the other is to construct test statistics to test hypotheses. The graph test method is a simple and widely used method to judge the stationarity. Its disadvantage is that the discriminant conclusion has a strong subjective color. Therefore, it is best to use the statistical test method to assist the judgment. At present, the most commonly used statistical test method for stationarity is the unit root test. The Flow chart of ARIMA model is shown in Figure 1.

![Flow chart of ARIMA model](image-url)

**Figure 1.** Flow chart of ARIMA model.
3. Prediction and analysis of population data

3.1. Collection and processing of total population data in Zhejiang Province

In order to analyze and predict the total population of Zhejiang Province over the next five years (assuming that the growth of population data is affected only by its own factors, free from outside interference), we collected the total population of Zhejiang Province from 1978 to 2017 from the statistical yearbook on the website of the Zhejiang Bureau of Statistics (see Table 1), and selected the total population data from 1978 to 2016 for processing and analysis. The total population data of Zhejiang Province in 2017 is used for verification. By using the SAS statistical software, a broken line diagram of the total population of Zhejiang Province from 1978 to 2016, as shown in Figure 2.

Table 1. The total population of Zhejiang Province from 1978 to 2017 (unit: million).

| Year | Total population | Year | Total population | Year | Total population | Year | Total population |
|------|------------------|------|------------------|------|------------------|------|------------------|
| 1978 | 3750.96          | 1988 | 4169.85          | 1998 | 4446.86          | 2008 | 4687.85          |
| 1979 | 3792.33          | 1989 | 4208.88          | 1999 | 4467.46          | 2009 | 4716.18          |
| 1980 | 3826.58          | 1990 | 4234.91          | 2000 | 4501.22          | 2010 | 4747.95          |
| 1981 | 3871.51          | 1991 | 4261.37          | 2001 | 4519.84          | 2011 | 4781.31          |
| 1982 | 3924.32          | 1992 | 4285.91          | 2002 | 4535.98          | 2012 | 4799.34          |
| 1983 | 3963.10          | 1993 | 4313.30          | 2003 | 4551.58          | 2013 | 4826.89          |
| 1984 | 3993.09          | 1994 | 4341.20          | 2004 | 4577.22          | 2014 | 4859.18          |
| 1985 | 4029.56          | 1995 | 4369.63          | 2005 | 4602.11          | 2015 | 4873.34          |
| 1986 | 4070.07          | 1996 | 4400.09          | 2006 | 4629.43          | 2016 | 4910.85          |
| 1987 | 4121.19          | 1997 | 4422.28          | 2007 | 4659.34          | 2017 | 4957.63          |

Figure 2. Annual variation of total population in Zhejiang Province (unit: million).

From Figure 2, it can be seen that the sequence contains a long-term trend of curve increment, which is preliminarily judged to be a non-stationary sequence. The study of ARIMA model for non-stationary time series requires the difference processing of population data. Because the original
sequence diagram series shows an approximate linear trend, the first-order difference is selected. The sequence diagram after the first-order difference is shown in Figure 3.

![Figure 3. Time series diagram after the first-order difference of total population in Zhejiang Province in 1978-2016.](image)

Figure 3. Time series diagram after the first-order difference of total population in Zhejiang Province in 1978-2016.

In order to further determine the stationarity, the autocorrelation diagram of the differential sequence is investigated, as shown in Figure 4.

![Figure 4. Autocorrelation.](image)

**Autocorrelations**

| Lag | Covariance | Correlation | Std Error |
|-----|------------|-------------|-----------|
| 0   | 87.177465  | 1.00000     | 0         |
| 1   | 48.047265  | 0.55114     | 0.162221  |
| 2   | 35.074173  | 0.40233     | 0.205677  |
| 3   | 28.139429  | 0.32278     | 0.243389  |
| 4   | 20.575452  | 0.23602     | 0.251459  |
| 5   | 24.014403  | 0.27547     | 0.262912  |
| 6   | 29.166239  | 0.33456     | 0.269244  |
| 7   | 22.057834  | 0.25302     | 0.270505  |
| 8   | 9.915084   | 0.11373     | 0.270505  |
| 9   | 5.528312   | 0.06341     | 0.270896  |
| 10  | -0.840865  | -0.0965     | 0.270905  |
| 11  | -3.383406  | -0.33881    | 0.271052  |
| 12  | -5.231030  | -0.6000     | 0.271401  |
| 13  | -6.035320  | -0.6923     | 0.271865  |
| 14  | -14.336694 | -1.6445     | 0.274471  |
| 15  | -25.140723 | -2.8839     | 0.282332  |
| 16  | -20.808959 | -2.3870     |           |

"." marks two standard errors

In order to further determine the stationarity, the autocorrelation diagram of the differential sequence is investigated, as shown in Figure 4.

The autocorrelation diagram shows that the sequence has a short-term correlation, and then the white noise test is carried out on the sequence. The results shown that the significance level of the test is 0.05, the P value of the delayed statistics is less than 0.0001 and 0.05, so the sequence can’t be regarded as a white noise sequence which is used to fit the ARMA model. The MINIC option specifies that the SAS system outputs the BIC information of all ARMA (p, q) models whose autocorrelation delay order is less than or equal to 5 and the moving average delay order is less than or equal to 5. It is found that the order of the model whose BIC information reaches the minimum is obtained.
It is found that the AR (1) model is fitted to the first order difference sequence and the parameters are significant, which means that the fitting model of the total population data of Zhejiang Province is the ARIMA (1,1,0) model, and the fitting results are as follows:

\[ x_t - x_{t-1} = 31.51505 + 0.56721(x_{t-1} - x_{t-2}) + \varepsilon_t, \]  

(1)

then

\[ x_t = 31.51505 + 1.56721x_{t-1} - 0.56721x_{t-2} + \varepsilon_t. \]  

(2)

Partial Autocorrelations

| Lag | Correlation |
|-----|-------------|
| 1   | 0.55114     |
| 2   | 0.14158     |
| 3   | 0.07974     |
| 4   | 0.0180      |
| 5   | 0.14795     |
| 6   | 0.16655     |
| 7   | -0.04341    |
| 8   | -0.16173    |
| 9   | -0.03725    |

Figure 5. Partial Autocorrelations.

In order to further verify the effectiveness of the model, the SBC and AIC principle are also used in this paper. It can be seen from Figure 5 that the partial autocorrelation coefficient of the stationary sequence shows a trailing phenomenon after \( p=2 \), so \( p \) can be 1 or 2. Meanwhile, it can be found that the AIC value and SBC value of the model reach the minimum value when \( p=1 \) and \( q=1 \) according to Table 2, which means that the ARIMA (1, 1, 0) is the most suitable one.

Table 2. Comparison between AIC and SBC criterion of ARIMA model.

| ARIMA (p, d, q) | ARIMA (1, 1, 0) | ARIMA (2, 1, 0) |
|----------------|----------------|-----------------|
| AIC            | 267.4879       | 268.9201        |
| SBC            | 270.7631       | 273.8329        |

Therefore, the ARIMA (1, 1, 0) model is used to predict the total population of Zhejiang Province in the next six years, and the results are shown in Table 3.

Table 3. Results of predict the total population of Zhejiang Province in 2017-2022.

| Forecasts for variable x                      |
|----------------------------------------------|
| obs  | Forecast | Std Error | 95% Confidence Limits |
|------|----------|-----------|-----------------------|
| 40   | 4945.7654| 7.9649    | 4930.1546 4961.3763   |
| 41   | 4979.2092| 14.8073   | 4950.1875 5008.2309   |
| 42   | 5011.8182| 21.1095   | 4970.4444 5053.1921   |
| 43   | 5043.9538| 26.7921   | 4991.4424 5096.4653   |
| 44   | 5075.8208| 31.9046   | 5013.289  5138.3526   |
| 45   | 5107.5355| 36.5295   | 5035.9389 5179.1321   |

According to Table 1, the actual total population data of Zhejiang Province in 2017 is about 49.5763 million, while the model ARIMA (1,1,0) predicts that the total population data of Zhejiang Province in 2017 is about 49.457654 million people. Obviously, it’s 95% confidence interval is [4930.1546, 4961.3763], and the actual data falls into the confidence interval. As we all know, since the full implementation of the two-child policy in 2016, the total population will certainly be more
than the forecast. Therefore, it can be considered that the prediction accuracy of the model fitted in this paper is high and the prediction results are reliable.

3.2. Collection and processing of agricultural population and non-agricultural population data in Zhejiang Province

In order to further verify the effectiveness of the model, the agricultural population and non-agricultural population of Zhejiang Province from 1978 to 2014 were collected from Zhejiang statistical database (see Table 4 and Table 5). In this subsection, the data of the two different groups of data are analyzed and processed respectively.

Since Zhejiang household registration has no longer been divided into agricultural and non-agricultural hukou from 2015, relevant data from 1978 to 2014 can only be collected in the database. Therefore, the agricultural population data and non-agricultural population data of Zhejiang Province from 1978 to 2013 were selected for analysis, and the data for 2014 were retained for verification.

Table 4. Agricultural population data of Zhejiang Province from 1978 to 2014 (unit: million).

| Year | Agricultural population | Year | Agricultural population | Year | Agricultural population | Year | Agricultural population |
|------|-------------------------|------|-------------------------|------|-------------------------|------|-------------------------|
| 1978 | 3321.96                 | 1988 | 3487.61                 | 1998 | 3539.78                 | 2008 | 3292.37                 |
| 1979 | 3332.57                 | 1989 | 3515.46                 | 1999 | 3519.79                 | 2009 | 3282.23                 |
| 1980 | 3346.40                 | 1990 | 3538.13                 | 2000 | 3506.20                 | 2010 | 3279.06                 |
| 1981 | 3362.04                 | 1991 | 3555.37                 | 2001 | 3473.63                 | 2011 | 3279.43                 |
| 1982 | 3387.79                 | 1992 | 3560.13                 | 2002 | 3438.76                 | 2012 | 3277.74                 |
| 1983 | 3413.05                 | 1993 | 3563.24                 | 2003 | 3394.08                 | 2013 | 3281.48                 |
| 1984 | 3425.47                 | 1994 | 3565.19                 | 2004 | 3353.16                 | 2014 | 3279.11                 |
| 1985 | 3395.35                 | 1995 | 3567.14                 | 2005 | 3335.30                 |      |                         |
| 1986 | 3417.19                 | 1996 | 3570.17                 | 2006 | 3317.26                 |      |                         |
| 1987 | 3455.14                 | 1997 | 3557.19                 | 2007 | 3308.21                 |      |                         |

Table 5. Non-agricultural population data of in Zhejiang Province from 1978 to 2014 (unit: million).

| Year | Non-Agricultural population | Year | Non-Agricultural population | Year | Non-Agricultural population | Year | Non-Agricultural population |
|------|-----------------------------|------|-----------------------------|------|-----------------------------|------|-----------------------------|
| 1978 | 429.00                      | 1988 | 682.24                      | 1998 | 907.08                      | 2008 | 1395.48                    |
| 1979 | 459.76                      | 1989 | 693.42                      | 1999 | 947.67                      | 2009 | 1433.95                    |
| 1980 | 480.18                      | 1990 | 696.78                      | 2000 | 995.02                      | 2010 | 1468.90                    |
| 1981 | 509.47                      | 1991 | 706.00                      | 2001 | 1046.21                     | 2011 | 1501.88                    |
| 1982 | 536.53                      | 1992 | 725.78                      | 2002 | 1097.22                     | 2012 | 1521.61                    |
| 1983 | 550.05                      | 1993 | 750.06                      | 2003 | 1157.50                     | 2013 | 1545.41                    |
| 1984 | 567.62                      | 1994 | 776.01                      | 2004 | 1224.06                     | 2014 | 1580.06                    |
| 1985 | 634.21                      | 1995 | 802.49                      | 2005 | 1266.81                     |      |                         |
| 1986 | 652.88                      | 1996 | 829.92                      | 2006 | 1312.17                     |      |                         |
| 1987 | 666.05                      | 1997 | 865.09                      | 2007 | 1351.13                     |      |                         |
3.2.1. Processing of agricultural population data in Zhejiang Province. Using the calculation steps like that given in subsection 3.1, it can be concluded that the fitting model of the agricultural population data of Zhejiang Province is the ARIMA (1,1,0) model, and the fitting results are as follows:

\[ x_t - x_{t-1} = 0.77577(x_{t-1} - x_{t-2}) + \epsilon_t, \]  \hspace{1cm} (3)

then

\[ x_t = 1.77577x_{t-1} - 0.77577x_{t-2} + \epsilon_t. \]  \hspace{1cm} (4)

It can be concluded from Table 6 that that the predicted value of agricultural population data in Zhejiang Province in 2014 predicted by ARIMA (1, 1, 0) model is 32.843814 million which means that the result is at 95% confidence level, and the interval is [3257.8470, 3310.9158]. The actual agricultural population in Zhejiang Province in 2014 was 32.7911 million according to Table 4. It can be found that the difference is 52714, and the actual data fall into the confidence interval, so it can be considered that the prediction accuracy of the fitted model is high and the prediction results are reliable.

Table 6. Results of predict the agricultural population of Zhejiang Province in 2014-2017.

|obs| Forecast  | Std Error | 95% Confidence Limits |
|---|-----------|-----------|----------------------|
|37 | 3284.3814 | 13.5382   | 3257.8470            |
|38 | 3286.6321 | 27.5905   | 3232.5557            |
|39 | 3288.3782 | 42.3948   | 3205.2860            |
|40 | 3289.7328 | 57.2733   | 3177.4791            |

Finally, the forecast results of agricultural population in Zhejiang Province from 2015 to 2017 are shown in Table 7.

Table 7. Results of predict the agricultural population of Zhejiang Province in 2015-2017 (unit: million).

| Year | Agricultural population | Year | Agricultural population | Year | Agricultural population |
|------|-------------------------|------|-------------------------|------|-------------------------|
| 2015 | 3286.6321               | 2016 | 3288.3782               | 2017 | 3289.7428               |

3.2.2. Processing of Non-agricultural population data in Zhejiang Province. The same calculation steps as for the total population data in Zhejiang Province apply here, it can be founded that the fitting model of the non-agricultural population data of Zhejiang Province is the ARIMA (1,1,0) model, and the fitting results are as follows:

\[ x_t - x_{t-1} = 31.45073 + 0.62253(x_{t-1} - x_{t-2}) + \epsilon_t, \]  \hspace{1cm} (5)

then

\[ x_t = 31.45073 + 1.62253x_{t-1} - 0.62253x_{t-2} + \epsilon_t. \]  \hspace{1cm} (6)

By the same way, the non-agricultural population data for the next four years from 2013 are shown in Table 8.
Table 8. Results of predict the non-agricultural population of Zhejiang Province in 2014-2017.

| obs | Forecast | Std Error | 95% Confidence Limits |
|-----|----------|-----------|-----------------------|
| 37  | 1572.0979| 12.7189   | 1547.1694 1597.0264   |
| 38  | 1600.5837| 24.2414   | 1553.0715 1648.0959   |
| 39  | 1630.1886| 35.2314   | 1561.1362 1699.2410   |
| 40  | 1660.4903| 45.4002   | 1571.5075 1749.4731   |

Finally, the forecast results of non-agricultural population in Zhejiang Province from 2015 to 2017 are shown in Table 9.

Table 9. Results of predict the non-agricultural population of Zhejiang Province in 2015-2017 (unit: million).

| Year | Non-Agricultural population | Year | Non-Agricultural population | Year | Non-Agricultural population |
|------|------------------------------|------|------------------------------|------|------------------------------|
| 2015 | 1600.5837                    | 2016 | 1630.1886                    | 2017 | 1660.4903                    |

3.3. Verification of model

According to the subsections 3.1 and 3.2, it can be found that the optimal model of the agricultural population data, non-agricultural population data and the total population data of Zhejiang are all ARIMA (1, 1, 0). As can be seen from Table 10, the prediction of the total population data of Zhejiang Province from 2015 to 2017 is more accurate, which further verifies the validity of the ARIMA (1,1,0) model. Therefore, the ARIMA (1, 1, 0) model of the total population data of Zhejiang Province has good effectiveness, and the fitting results are as follows:

\[ x_t = 31.51505 + 1.56721x_{t-1} - 0.56721x_{t-2} + \epsilon_t. \]  

Table 10. Results of predict the population of Zhejiang Province in 2015-2017 (unit: million).

| Year | Agricultural population | Non-Agricultural population | Total | Actual value | Phase difference (absolute value) | 95% Confidence Limits |
|------|-------------------------|-----------------------------|-------|--------------|-----------------------------------|-----------------------|
| 2015 | 3286.6321               | 1600.5837                   | 4887.2158 | 4873.34 | 13.8758 | 4785.6272 4988.8045 |
| 2016 | 3288.3782               | 1630.1886                   | 4918.5668 | 4910.85 | 7.7168 | 4766.4222 5070.7114 |
| 2017 | 3289.7328               | 1660.4903                   | 4950.2231 | 4957.63 | 7.4069 | 4748.9866 5151.4595 |

4. Summary

The population problem has always been an important issue in the development of human society. The prediction of the population data of a country and a state in the next few years will not only play a guiding role in the economic and social development of the region but also provide scientific basis for the evaluation of relevant policies. As an economically developed coastal city, Zhejiang Province urgently needs to grasp the development trend of its own population. The study of its population structure can play a positive role in promoting the development of local economic and cultural undertakings.

In this paper, with the help of SAS statistical software, the ARIMA (1,1,0) model is constructed according to the total population data of Zhejiang Province from 1978 to 2016. In order to test whether...
the selected parameters of the model are reasonable, the population data of Zhejiang Province in 2017 are validated. The experimental results show that the model has a good fitting effect. Then, in order to further verify the practical effectiveness of the model, the data of agricultural population and non-agricultural population in Zhejiang Province are processed, analyzed and verified according to the same modeling method. Through verification and combining with practice, it is concluded that the actual fitting effect of ARIMA (1, 1, 0) model based on total population data is good. Finally, we predict the population data of Zhejiang Province in the next few years. The prediction results show that the population of Zhejiang Province is generally increasing year by year. It is hoped that the study of this topic can provide a theoretical reference for the planning and adjustment of relevant policies in Zhejiang Province.

References
[1] He X H, Han T, Dong Q and Liang W K 2017 Prediction of the number of Chinese population in the Future under the Comprehensive two-child Policy based on the time Prediction Model[J]. Course education research 36 247-248
[2] Li Z K, Zhang L N, Zheng J S and He X H 2017 The influence of China's population on Economic Development in the Future[J]. Shenzhen 31 219
[3] He X H, Zhang S H and Wang Ying X J 2017 Prediction of China's population under the Comprehensive two-Child Policy based on the two-tier "small World" Model[J]. Scientist 17 11,44
[4] Tu X L and Tu H Y 2009 A Comparative study of ARIMA and exponential smoothing method in population Prediction in China[J]. Statistics and Decision 16 21-23
[5] Tu X L 2011 A Comparative study on the Prediction models of income Gap between Urban and Rural residents in China[J]. Journal of Anhui Agricultural Sciences 39(24) 15057-15062
[6] Tu X L and Huang Y L 2011 Application of ARIMA Multiplicative Seasonal Model to the Forecast for Travel Demand-A Case Study of Guilin[J]. Journal of Guangxi University of Finance and Economics 24(1) 112-117
[7] Han S T and Zhou Y X 2014 A Comparative study of multiple Linear regression and ARIMA in China's population Prediction[J]. China Management Informationization 17(22) 100-103
[8] Mumbare S S, Gosavi S and Almale B 2014 Trends in Average Living Children at the Time of Terminal Contraception: A Time Series Analysis Over 27 Years Using ARIMA (p, d, q) Nonseasonal Model[J]. Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine 39(4) 223-228
[9] Chen P and Lu W Q 2016 Analysis on the Application of ARIMA Model in Regional population Prediction[J]. Contemporary Economics 6 111-113
[10] Jadhav V, Chinnappa Reddy B V, Gaddi G M 2017 Application of ARIMA Model for Forecasting Agricultural Prices [J]. Journal of Agricultural Science and Technology 5 981-992
[11] Karthika M, Krishnaveni and Thirunavukkarasu V 2017 Forecasting of meteorological drought using ARIMA model [J]. Indian Journal of Agricultural Research 2 103-111
[12] Xia F 2011 Prediction of Total population in Zhejiang Province based on ADL[J]. Statistical Science and Practice (11) 43-45
[13] Zhang J J and Ma D D 2011 Application of Equal dimensional Grey number compensation Model in population Prediction: a case study of Zhejiang Province[J]. Journal of Wenzhou University(Natural Sciences) 32(4) 37-42