Exploiting Regressive Model for Population Prediction in China

Jiayi Li *

Department of computer science and technology, Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding author email: lijiayi0702@bupt.edu.cn

Abstract. The demographics of China reveal a huge population, which amounted to around 1.4 billion people in 2022. However, evidence and authorities demonstrate that the Chinese population is about to shrink for the first time and the country's total fertility rate keeps decreasing. It introduces great uncertainty to the future development of China. An accurate population prediction is important to estimate the economy, make reasonable policies, and stabilize society. In this paper, three machine learning models, including the polynomial regression, logistic growth model and Autoregressive Integrated Moving Average (ARIMA), are used to forecast the population in China. The effectiveness of different models on population forecasting is compared and analyzed. The results show that ARIMA performs the best, which is about a 0.34% error rate validated on previous population data. The prediction results demonstrate that the population in China will experience a brief rise and then enter negative growth. To embrace the population decline, the government should get prepared for the aging society and propose reasonable policies to increase fertility.

Keywords: Population Prediction; Polynomial Regression; Logistic Growth Model; ARIMA.

1. Introduction

As the third largest country in the world, China's population is second to none. Although China has a huge number of resources, per capita resources are very few. Therefore, in order to control the excessive expansion of the population in China, the family planning policy was proposed in the 1970s, which has achieved remarkable results, mainly reflected in the following aspects: the excessive growth of the population has been effectively controlled; The national economy has developed rapidly; People's living standards have improved substantially; Children's rights and interests have been protected; The living standard of the elderly continues to improve. But at the same time, some problems appear, such as the aging of the population, unreasonable age structure, and the disappearance of demographic dividend. These problems have brought population control strategies back to the forefront of people's minds. Therefore, to promote population growth, China issued the two-child policy and the three-child policy in 2016 and 2021, respectively. However, the effect of the policy is not satisfactory, and the growth rate of China's population has continued to decrease in recent years. According to the UN World Population Prospects 2022, China's population may enter negative growth in 2022 [1]. Therefore, reasonable population prediction is of great significance.

As the basis of economic and social research, population prediction has been widely used in various aspects of national development. Over the past few years, researchers have studied issues such as population aging [2], sex ratios [3], and demographic dividend [4]. Some researchers have focused too much on the effectiveness of policies. However, the effectiveness of policies is not only related to population base but also related to age structure, sex ratio, national development, natural disasters, people's will, and other factors. In this paper, the future population of China will be forecasted based only on previous population data. The polynomial regression model, logistic population growth model, and Autoregressive Integrated Moving Average (ARIMA) model will be used to analyze and fit the population data from 1950 to 2021. By dividing the training set and test set, the error of model prediction will be analyzed. Three models are used to forecast the population for the next 20 years, plot the trend, and select the best model.

2. Method

This section first introduces the data sources, data collection, and processing, and then explains three forecasting methods: polynomial regression model, logistic population growth model, and ARIMA. Finally, the population prediction for the next two decades based on the three methods will be realized, and the prediction results will be compared.

2.1 Dataset

The dataset used in this paper is from the National Bureau of Statistics [5]. It includes data on China's total population from 1950 to 2021. Redundant content in the data set, including annotations and extraneous data, is removed to get the data in the correct format. To evaluate the predictive ability of different models, data from 1950 to 2011 are used as the training set, and data from 2012 to 2021 are used as the test set.

2.2 Model Introduction and Implementation

2.2.1 Polynomial Regression

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In simple linear regression, the model

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

is used, where ε is the error that is normally distributed with mean 0. β_0, β_1 are the intercept and slope of the line, respectively. However, a simple linear regression algorithm only works when the relationship of data is linear. If there is non-linear data, the polynomial regression is a better choice. Polynomial function refers to the function obtained by finite multiplication and the addition of constant and independent variable x. In general, the expected value of y could be modeled as an nth-degree polynomial, yielding the general polynomial regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon \tag{2}$$

The polynomial regression model is a kind of linear regression model, which is widely used in economics and prediction of prediction. Polynomial regression is often used to fit the trend of various indexes and plays an important role in the analysis.

In order to get the best polynomial fitting curve based on the known data, the least squares method is applied. The least square method is a mathematical optimization technique, which finds the best function that fits the data by minimizing the sum of squares of the errors. The least square method is the most common method to solve the curve fitting problem. The formula is as follows:

$$f(x) = \alpha_1 \varphi_1(x) + \alpha_2 \varphi_2(x) + \cdots + \alpha_m \varphi_m(x) \tag{3}$$

There, $\varphi_k(x)$ is a set of linear functions which do not affect each other. α_k is undetermined coefficient. The fitting criterion is to minimize the sum of squares of the distance between $y_i (i = 1,2, \dots, n)$ and $f(x_i)$.

In this program, the following codes, using the numpy package of Python, are leveraged to fit the data and make predictions:

mymodel = numpy.poly1d(numpy.polyfit(year, population, 3))
myline = numpy.linespace(1950, 2041, 92)

It is implemented by the NumPy package in the Python programming language. The third parameter of function *polyfit* is the degree of the polynomial. There, the basic form of the generated curve is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$, and the degree of the polynomial is 3, which has proven to be better than 2 and 4.

2.2.2 Logistic Growth Model

The logistic growth model is generated when people study the problem of population growth. This model investigates a standard model of population growth in a constrained environment.

In the 18th century, Malthus believed that the growth rate in the process of natural population growth was constant (r), and based on this finding, he established a population model: suppose that the number of people is $N(t)$ at the time t, then the population growth during the period from t to $t + \Delta t$ is $rN(t)\Delta t$. Thus, he got the formula $N(t) = N_0 e^{n}$.

However, a constant rate of population growth is unrealistic. For example, if $r < 0$, the population will dwindle to extinction. Thus, to make the model more accurate, Verhulst revised the model according to the view that "the population cannot exceed the environmental capacity", which resulted in the logistic model: $Y'(t) = rY(t)\big(M - P(t)\big),\ r,\ M > 0$. According to this model, When the population is small, the growth is stable. When the population approaches the maximum capacity M, the growth rate gradually decreases to 0. Through transformation, the basic form of the model is as follows:

$$f(x) = \frac{M}{1+Ce^{-rt}} \tag{4}$$

The Logistic model has the properties of limited and monotonically increasing, which is consistent with most population growth patterns. There are two parameters in the formula: the maximum environmental capacity of population M and the natural growth r. The relative population growth rate can be expressed as $r(1 - \frac{M}{P})$, thus the C can be expressed as $\frac{M}{P_0} - 1$.

In the project, the author defines the model "logistic_increase_function" according to the function mentioned before:

$$f(x) = \frac{M}{1+Ce^{-rt}} \tag{5}$$

In order to get the best value for the parameters M, C, and r, the author uses least square method with the function *leastsq* from *scipy* package in Python:

logistic_params = leastsq(err_f, logistic_p0, args=(t, China_y)),

where *logistic_p0* is the initial value for $M,\ C_3$, and r t is the date (from 1950 to 2021) and *China_y* is the population. *err_f* is an error function defined as:

logistic_increase_function(p, t) - y

There, *logistic_increase_function(p, t)* is predictive values of the population with the date t and p, which contains parameters M, C and r that are obtained through *leastsq*.

Nowadays, logistic growth model has been widely used. Triambak S. used logistic growth model to predict COVID-19 mortality [6], Asraful Haque Md. used logistic growth model to analyze software reliability under uncertain factors [7], and Abdul Latif Nurul Syaza used the logistic growth model to establish agricultural management strategies [8].

2.2.3 ARIMA

ARIMA model is a time series forecasting model combined with autoregressive model (AR), moving average model (MA) and different methods. If the series is a stationary time series, the model is a mixed model with p-order autoregressive and q-order moving average, denoted by ARIMA (p, q). In practice, many economic series are non-stationary time series, and d is the order of differences. Such a model is called ARIMA model, denoted as ARIMA (p, d, q). It can be expressed by the following formula:

$$\left(1-\sum_{i=1}^{p}\emptyset_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{p}\theta_i L^i\right)\varepsilon_t \tag{6}$$

where L is the lag operator, $d \in Z, d > 0$.

The stationarity test is performed by the Augmented Dickey-Fuller (ADF) ADF test and observing the p-value. If the p-value is less than 0.05, the data can be considered stationary. Otherwise, the data needs to be differenced. The function uses to make the ADF test is as follows:

from statsmodels.tsa.stattools import adfuller
result = adfuller(dta)

By executing the codes, the author gets p-value≈1.749e-08, which indicates that the data already has stationarity. Therefore, the parameter is set to d=0.

The identification of parameters p,d,q in the model mainly depends on the analysis of autocorrelation and partial autocorrelation, Figure1 is the plot of correlation and autocorrelation. It can be observed that the partial autocorrelation of the sequence decreases rapidly after the second order, showing a second-order censoring state, so p=2. However, the correlation coefficient of autocorrelation map is still significantly non-zero in the lag of ten orders, showing trailing decay characteristics, so q=0.

In order to get a more accurate selection of p and q values, *auto_arima* function in *pmdarima* package was used for obtaining the corresponding parameters with the smallest AIC and BIC values, which were p=3, q=2 and p=2, q=0, respectively. Therefore, ARIMA (2,0,0) model and ARIMA (3,0,2) model were selected to fit data and make predictions.

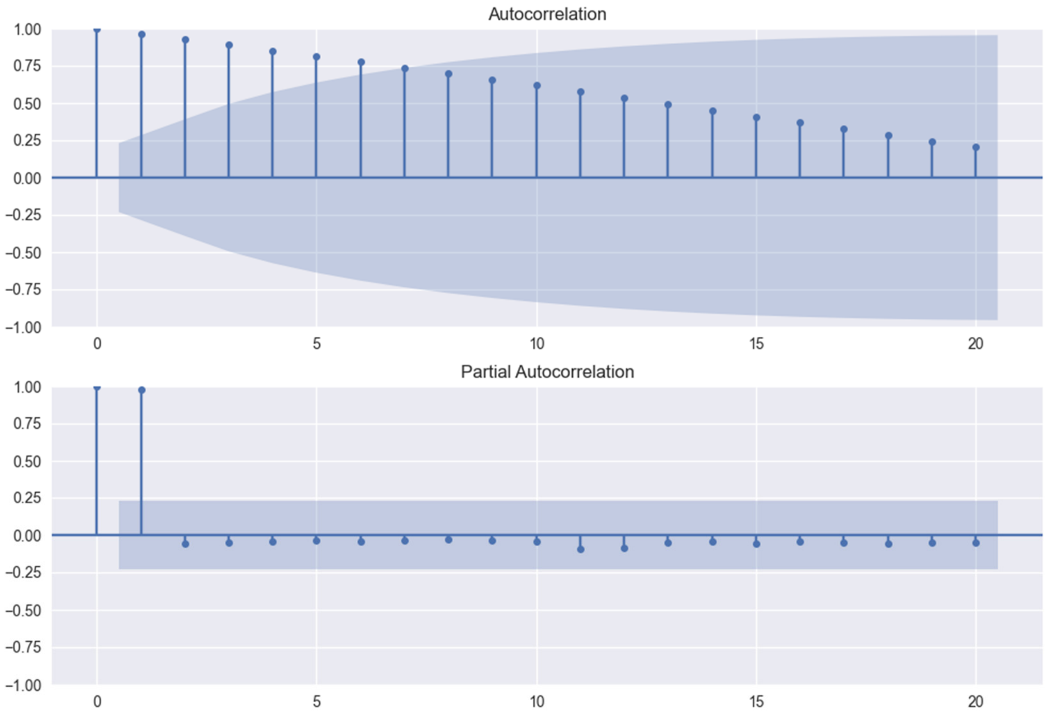

Fig 1. Autocorrelation and Partial Autocorrelation

The author predicts the population with the two models mentioned above using the following codes:

model = ARIMA(temp, order=(2, 0, 0)).fit()
predict = model.predict(start=1, end=91, dynamic=False)
model = ARIMA(temp, order=(3, 0, 2)).fit()
predict = model.predict(start=1, end=91, dynamic=False)

The advantage of the ARIMA model is its simplicity, requiring only endogenous variables without resorting to other variables. However, it requires the data to be stable and can only capture the linear relationship inherently. At the same time, the ARIMA model has been studied in detail by predecessors: Romanuke Vadim 's paper *Arima Model Optimal Selection for Time Series Forecasting* describes how ARIMA model selects the optimal fit [9]; Goyal Megha uses ARIMA model to forecast India's agricultural exports [10], Spyrou Evangelos D. compares ARIMA and LSTM forecast model and expounds the advantages and disadvantages of the two methods [11].

2.3 Evaluation

To evaluate the fitting curve with each model, the author draws the error scatter charts of the fitting values and real data of each model. The best fitting model is determined by the performance of the test set. In the charts which are mentioned in the next section, the author uses p for percentage error, which is defined as $p = \frac{error}{true\ value}$.

3. Results

In section 4.1, the author first validates and compares the effectiveness of the three aforementioned models, the polynomial regression model, the logistic growth model, and the ARIMA model with variance parameters. In section 4.2, the author predicts the population with the three models.

3.1 Model Comparison

To evaluate the effectiveness of the three models, the author first set the data from 1950 to 2011 as the training set and use the data from the training set and the three models mentioned above to forecast the population from 2012 to 2021. The author then recorded the results and plotted the trends of the prediction to compare with the original data. Table 1 shows the predictions of three models. The real value is true population. The polynomial, logistic, ARIMA (3,0,2) and ARIMA (2,0,0) show the difference between the predicted value and the true value under the three models respectively, and p equals to prediction deviation/true value of the corresponding model.

Table 1. Performance of different models

date	real value	polynomial	p	logistic	p	ARIMA (3,0,2)	p	ARIMA (2,0,0)	p
2012	135922	-1630.07	-1.20%	1431.419	1.05%	922.6437	0.68%	893.5471	0.66%
2013	136726	-2278.53	-1.67%	1445.871	1.06%	961.0349	0.70%	845.4838	0.62%
2014	137646	-3126.3	-2.27%	1322.052	0.96%	806.2104	0.59%	602.4587	0.44%
2015	138326	-3819.62	-2.76%	1416.27	1.02%	817.14	0.59%	544.3664	0.39%
2016	139232	-4826.74	-3.47%	1262.849	0.91%	530.7268	0.38%	214.759	0.15%
2017	140011	-5796.91	-4.14%	1215.127	0.87%	302.8071	0.22%	-29.7166	-0.02%
2018	140541	-6610.37	-4.70%	1395.455	0.99%	258.151	0.18%	-65.8292	-0.05%
2019	141008	-7455.36	-5.29%	1618.197	1.15%	213.4624	0.15%	-79.3215	-0.06%
2020	141212	-8134.14	-5.76%	2083.725	1.48%	371.3795	0.26%	129.6871	0.09%
2021	141260	-8755.94	-6.20%	2685.422	1.90%	627.4755	0.44%	453.9465	0.32%
average	139188	-5253.4	-3.75%	1587.639	1.14%	581.1031	0.42%	350.9381	0.25%

Figure 2 shows the trend of the real population and the predicted results under the three models. The X-axis represents the date and the Y-axis represents the population.

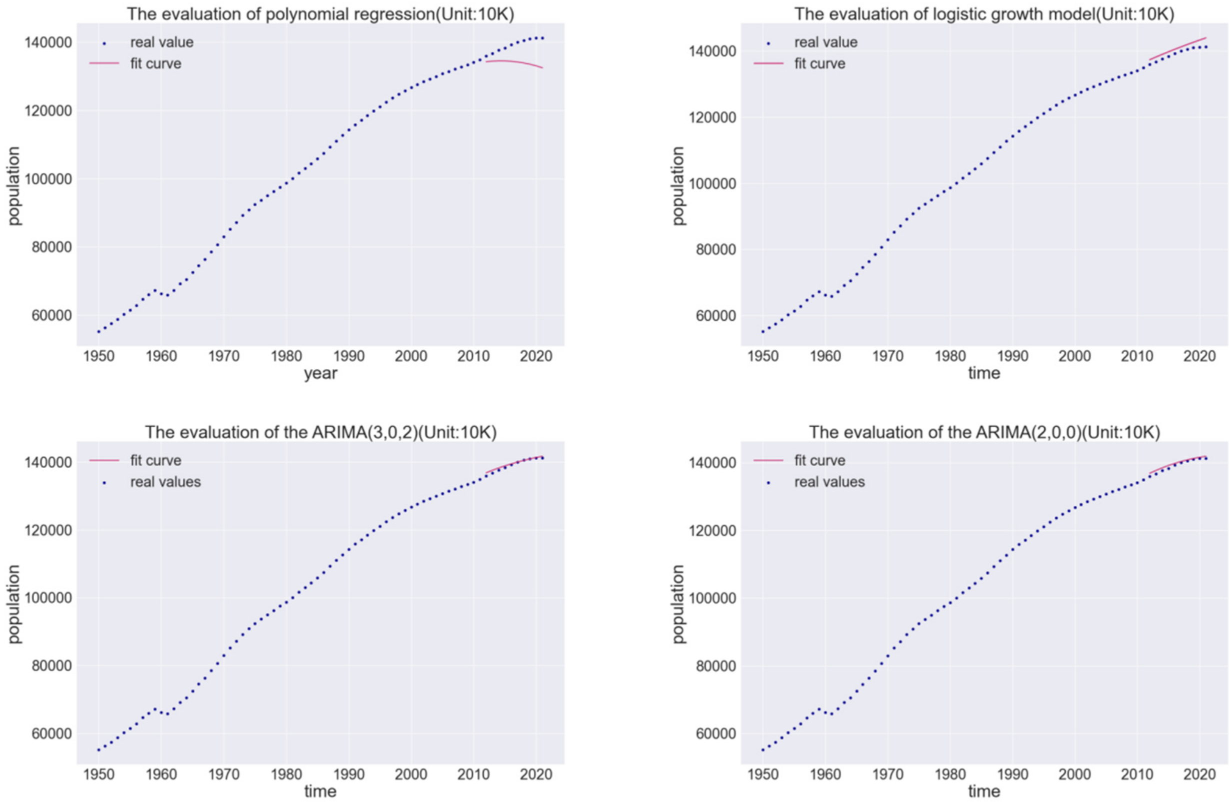

Fig 2. Evaluation results of different regression model

Previous data show that if the prediction error is within 2%, the model can be considered to have a good prediction effect. As can be seen from the above results, the qualified prediction models include the logistic growth model, ARIMA (3,0,2), and ARIMA (2,0,0). Therefore, the prediction results of these three models can be considered valid. Among them, ARIMA (3,0,2) and ARIMA (2,0,0) models have better prediction results, whose error range is within 1%, so the reference value is more significant.

Table 2 shows the maximum prediction error, the average of the error of the preceding three years, the average of the preceding five years, and the average of the preceding 10 years. Table 2 shows that ARIMA is the best-performing model in terms of both short-term and long-term results. The unit of value is 10 thousand.

Table 2. Evaluation of error of 4 model

		max error	3 years on average	5 years on average	10 years on average
polynomial regression	value	8755.942	-2344.97	-3136.25	-5243.4
	percentage	-6.20%	-1.71%	-2.27%	-3.75%
logistic growth model	value	2685.422	1399.781	1375.692	1587.639
	percentage	1.90%	1.02%	1.00%	1.14%
ARIMA(3,0,2)	value	961.0349	896.6296	807.5511	581.1031
	percentage	0.70%	0.66%	0.59%	0.42%
ARIMA(2,0,0)	value	893.5471	780.4965	620.123	350.9381
	percentage	0.66%	0.57%	0.45%	0.26%

3.2 Population Prediction

Table 3 shows the population of China from 2022 to 2041 predicted by the three models. According to the evaluation results, the two better models, ARIMA (3,0,2) and ARIMA (2,0,0), show slightly different trends. According to the ARIMA (3,0,2), the total population of China increases first and then decreases. The population will reach 14.3991 billion in 2022 and will reach peak at

14.24358 billion in 2030, an increase of about 103.67 million. The growth will be reduced to 13.97909 billion in 2041. According to the ARIMA (2,0,0), the total population will decrease to 14.12659 billion in 2022, and then continue to decrease until it reaches 13.57766 billion in 2041, whose decrease is about 548.93 million.

Table 3. Population prediction for the next 20 years (Unit:10K)

year	polynomial regression	logistic growth model	ARIMA(3,0,2)	ARIMA(2,0,0)
2022	139998.9	143305.6	141399.1	141265.9
2023	139948.2	143871.2	141609.9	141231.7
2024	139828.1	144418.5	141826.4	141159.5
2025	139637.2	144948	142021.2	141051.1
2026	139374.2	145460.2	142182.8	140908.5
2027	139037.6	145955.5	142306.5	140733.5
2028	138625.9	146434.3	142390.4	140527.8
2029	138137.7	146897	142433.6	140293.2
2030	137571.6	147344.2	142435.8	140031.4
2031	136926.1	147776.1	142397	139743.9
2032	136199.9	148193.3	142317.2	139432.3
2033	135391.4	148596.2	142196.4	139098.1
2034	134499.3	148985.1	142035	138742.8
2035	133522.1	149360.5	141833.2	138367.7
2036	132458.4	149722.8	141591.1	137974.4
2037	131306.7	150072.4	141309.2	137563.9
2038	130065.6	150409.6	140987.8	137137.8
2039	128733.7	150734.9	140627.3	136697
2040	127309.6	151048.6	140228.2	136242.9
2041	125791.7	151351.1	139790.9	135776.6

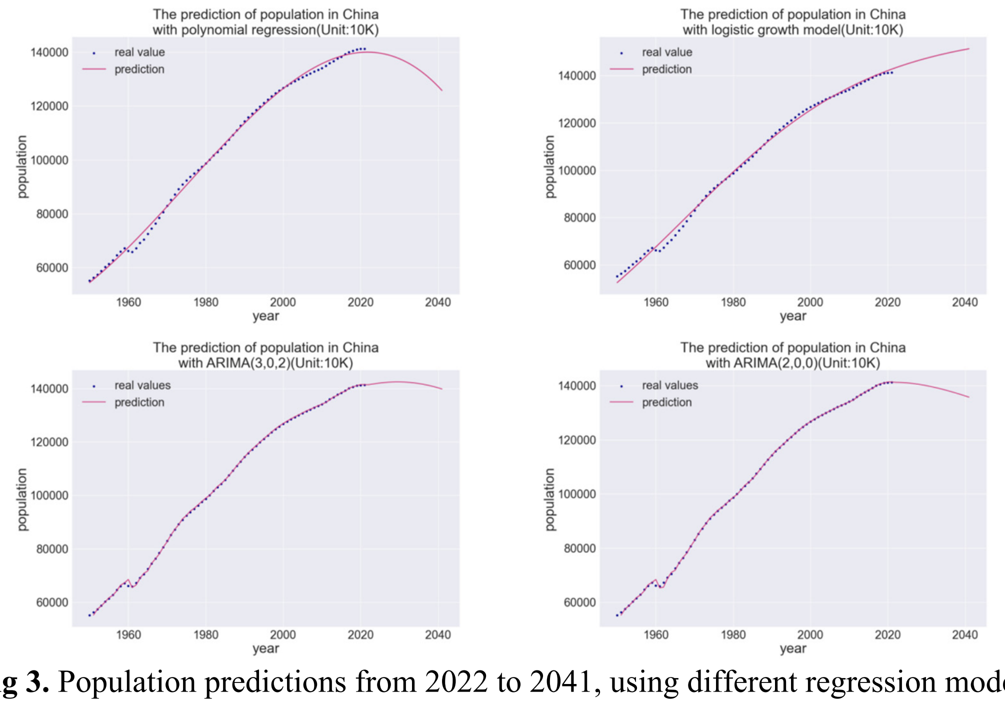

Fig 3. Population predictions from 2022 to 2041, using different regression models.

According to the Logistic Grow Model, the population will continue to grow in the next 20 years, reaching 15.13511 billion in 2022, an increase of about 1009.11 million compared with 2021; According to Polynomial Regression, the total population of China tends to decrease, and it will decrease to 12.57917 billion in 2041, which is about 1547.4 million lower than that in 2021. Table 3 shows the population predicted by different models for the next 20 years. Figure 3 shows the trends in the predictions of the four models.

4. Discussion

Through the experiment, it can be observed that ARIMA model is the most accurate one. The prediction error of the ARIMA model is within 1% in both short and long term, while the polynomial regression model has an average error of -3.75% and the maximum error is over 6%. The average error of logistic growth model is 1.14%, and the maximum error is 1.9%. The projections suggest a brief rise in the total population, followed by a gradual decline. According to the UN World Population Prospects and the China Population Forecast Report 2021, China's population will achieve negative growth in 2022 and 2029 respectively, which is in line with the forecast results of ARIMA (3, 0, 2) and ARIMA (2, 0, 0), respectively. The forecast results of ARIMA (3, 0, 2) and ARIMA (2, 0, 0) show that China's population will experience transient growth, followed by negative growth in 2022 and 2031, respectively.

Although the results of the test sets show that the prediction of logistic growth model and ARIMA model are reasonable, they still need to be improved. For example, when using logistic growth model, variable slope can be calculated by gradient algorithm to obtain more accurate parameters. When using ARIMA for prediction, randomness test and white noise test should be carried out to ensure the accuracy and reasonableness of prediction results.

As the advent of population declines, China will also face some problems, the most significant of which is the disappearance of the demographic dividend. This will affect the development and decision-making of social economy, politics and other aspects. At the same time, the aging of the population leads to the shortage of labor force, which makes pension pressure become greater. In order to avoid this problem, it is necessary for the government to encourage fertility, improve the quality of the population, fully develop human resources for all ages and actively participate in economic globalization.

5. Conclusion

The population problem plays a key role in the development of various social undertakings. Predicting the size of the population is conducive to make appropriate population policies and has a far-reaching impact on social development. In this paper, three machine learning models are used to predict the population, and the effectiveness of different models on forecasting is compared. The results show that the ARIMA is the best prediction model. The test shows that the maximum prediction error of this model is less than 1%. However, the parameters of the model still need to be optimized to get more accurate prediction results.

The prediction demonstrates that the population in China will experience a brief rise, followed by a gradual decline. The decline of the population will lead to a series of social problems, such as aging population, insufficient labor force, and so on. Therefore, the government should prepare for the aging society as well as encourage fertility.

References

[1] The United Nations. World Population Prospects 2022: Summary of Results, https://www.un. org/ development/ desa/pd/content/World-Population-Prospects-2022, 2022.

[2] Gao Hong. Population aging prediction in Nanjing based on grey forecasting system. Jiangsu business theory, 2021(08): 137-140.

[3] Li Chunling. "Boy crisis", "leftover women phenomenon" and "Difficult employment for female college students" -- Social challenges brought by the reversal of gender ratio in education. Collection of Women's Studies, 2016(02):33-39.

[4] Huang fan, Duan cheng rong. From Demographic Dividend to Demographic Quality Dividend -- Based on the analysis of the Seventh National Census data. population and development, 2022, 28(01): 117-126.

[5] National Bureau of Statistics. Total population of China from 1950 to 2021. https://data. stats. gov. cn/ easyquery. htm?cn=C01, 2021.

[6] Triambak S., Mahapatra D.P., Mallick N., Sahoo R. A new logistic growth model applied to COVID-19 fatality data. Epidemics, 2021.

[7] Asraful Haque Md., Ahmad Nesar. A Logistic Growth Model for Software Reliability Estimation Considering Uncertain Factors. International Journal of Reliability, Quality and Safety Engineering, 2021.

[8] Abdul Latif Nurul Syaza, Anuar Mushoddad Noor Asma' Mohd, Mior Azmai Norin Syerina. Agriculture Management Strategies Using Simple Logistic Growth Model. IOP Conference Series: Earth and Environmental Science, 2020.

[9] Romanuke Vadim. Arima Model Optimal Selection for Time Series Forecasting. Maritime Technical Journal, 2022.

[10] Goyal Megha, Goyal S.K., Agarwal Subodh, Kumar Nitin. Forecasting of Indian Agricultural Export Using ARIMA Model. Journal of Community Mobilization and Sustainable Development, 2022.

[11] Spyrou Evangelos D., Tsoulos Ioannis, Stylios Chrysostomos. Applying and Comparing LSTM and ARIMA to Predict CO Levels for a Time-Series Measurements in a Port Area. Signals, 2022.