Journal of Sociology and Ethnology (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jsoce.2023.050304
ISSN 2616-2318 Vol. 5 Num. 3

Population Prediction of China Based on ARIMA-LSTM Combined Model

Gu Minghui

College of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

Keywords: ARIMA; LSTM; Population gross

Abstract: In recent years, the birth rate in China has been declining continuously. Accurate population prediction is of great significance for the government to formulate population macro-adjustment policies. In this paper, the residual optimization method is used to establish the combined prediction model of ARIMA-LSTM. It is found that the precision of combinatorial prediction is better than that of each single prediction, and the combinatorial prediction with residual optimization has better generalization ability. In addition, the forecast results show that China's population will continue to decline in 2023.

1. Introduction

1.1 Research Background and Significance

In recent years, China's birth rate continues to decline, the population problem may become an important factor restricting social development. In order to create a favorable population environment, China's fertility policy is also being adjusted. On August 20, 2021, the Standing Committee of the National People's Congress of China voted to implement the "three-child policy".

The total population is an important indicator reflecting the fertility rate. Accurate prediction of the total population is of great significance for the government to formulate the macro-adjustment policy of population.

1.2 Literature Review

The problem of sustainable population development has been widely concerned by scholars, Zhang Tianliang (2000) used GM (1,1) grey series prediction model to predict the birth rate of China from 1994 to 2000 [1].Yin Chunhua and Chen Lei (2005) used BP neural network technology to build a population prediction model and conducted an empirical analysis [2].Rayer, S (2009) et al. use of the ten-year census data of the United States from 1900 to 2000, and provided a fairly accurate prediction of the accuracy of population prediction with the prediction interval and trend extrapolation technology based on experience [3]. Mao Jiaohui (2018) predicted the birth rate of China in the next 5 years by using three prediction methods, namely multiple linear regression, exponential smoothing and ARMA model [4].Chen, LX (2022) et al. used the Malthus model, unary linear regression model, Logistic model and grey prediction model to forecast the population

of 210 prefecture-level cities in China [5].

By combing relevant literature, it can be found that scholars have conducted abundant researches on the total population prediction, but there is still no recognized optimal prediction method so far.Common methods for population prediction include grey prediction, ARIMA model, LSTM neural network and IOWA operator, etc. In this paper, two common single prediction methods, ARIMA and LSTM, are selected, and residual optimization method is used to establish a combined prediction model and empirically test the effectiveness of combined prediction.

2. Introduction to the Model

2.1 Arima Model

ARIMA model, also known as differential integrated moving average autoregressive model, is a common time series prediction method in econometrics and belongs to a linear model. Its expression is as follows:

$$\Delta^d x_t = u + \sum_{i=1}^{p} \phi_i \Delta^d x_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{1}$$

Where, x_t represents the original time series, $\Delta^d x_t$ represents a smooth and reversible ARMA process after d subdifference, μ is the drift term of $\Delta^d x_t$, ε is the white noise random error sequence, p and q is the order of the model.

The model includes AR, MA and ARMA processes, and modeling generally includes five steps: 1) smoothing the original time series; 2) Model identification; 3) parameter estimation to determine the equation; 4) Model test.

2.2 Lstm Model

LSTM, known as long and short term memory network, is a special kind of RNN, which is characterized by its time loop structure, and can describe the sequence data with spatiotemporal correlation well.The memory module of LSTM is composed of a storage unit, a forgetting gate, an input gate and an output gate, so that information can be selectively retained or deleted to better learn long-term dependencies, as shown in Figure 1.

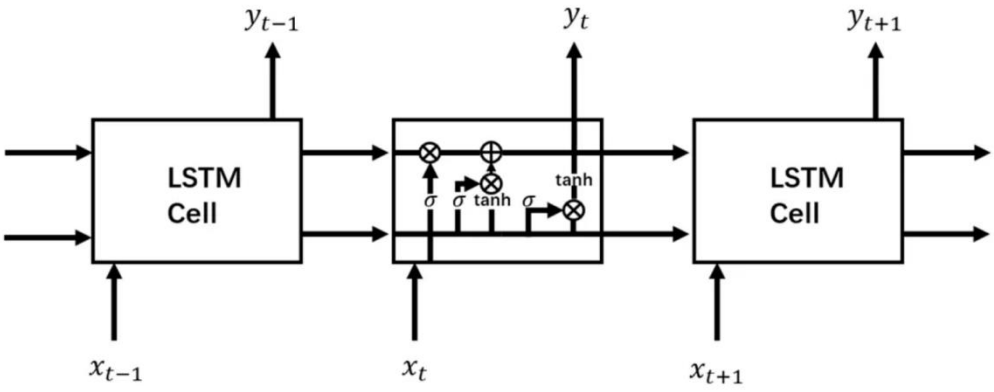


Figure 1: LSTM structure diagram

Where, x_t represents the input of the training sample at the t time, y_t represents the output of the current unit, tanh、δ represents various activation functions of the memory module, $\oplus$、$\otimes$ represents the basic operation of the vector.

2.3. Arima-Lstm Combined Prediction Model

Time series is composed of linear and nonlinear parts. ARIMA is the most widely used prediction model of linear model, and neural network is the most widely used prediction model of nonlinear model. Based on the prediction results of ARIMA model, this paper introduces LSTM model with long and short term memory function to correct the nonlinear part of ARIMA model. Finally, the combined prediction results of linear optimization are obtained. The basic idea is shown in Figure 2.

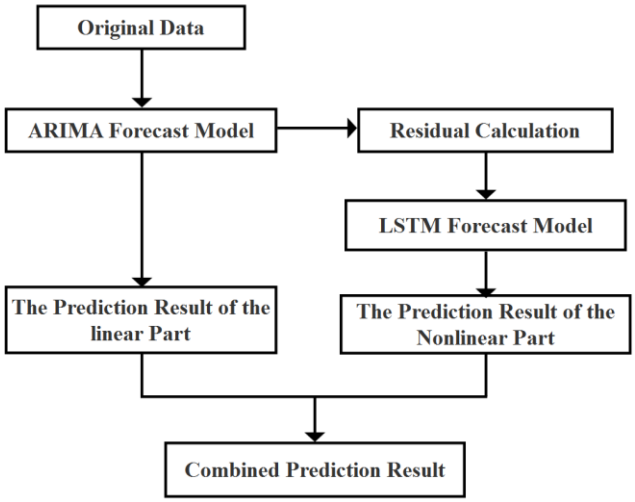


Figure 2: Basic thought of linear optimization combination prediction

3. Empirical Analysis

3.1 Data Source and Explanation

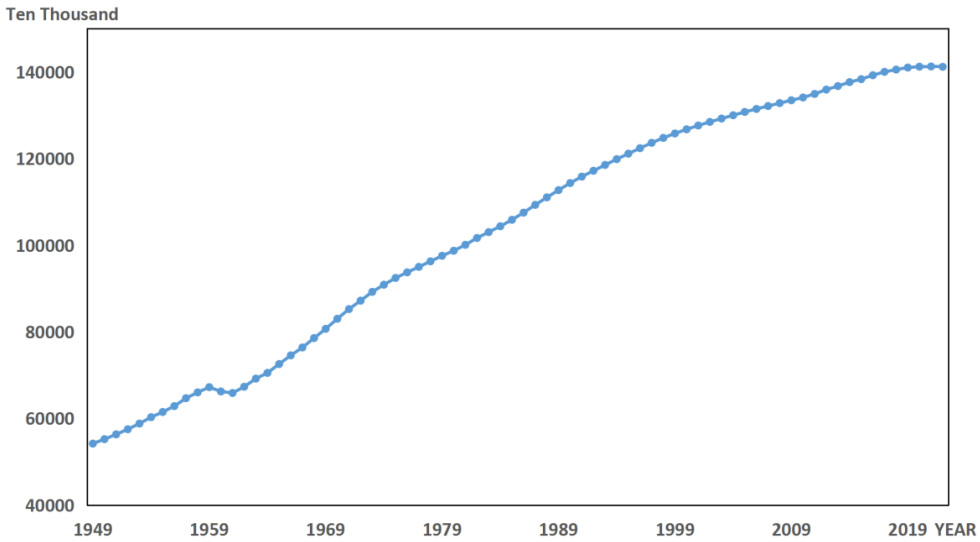


Figure 3: Trend chart of China's total population

In this paper, statistical data of China's total population from 1949 to 2022 are selected from China Statistical Yearbook, and the population trend chart is shown in Figure 3. It can be seen that before 1965, China's population growth rate fluctuated greatly due to environmental influences. After 1965, China's population growth rate showed a downward trend.

3.2 Forecast by Arima Model

3.2.1. Sequence Smoothing Treatment

According to the trend chart of total population, it can be judged that the sequence is not stationary. If the non-stationary sequence is used for regression, there will be a phenomenon of false regression, which requires the stationarity processing of the original sequence.According to the sequential diagram after second-order difference, it is found that the sequential diagram after second-order difference no longer has sequential characteristics, but has non-zero mean value. Therefore, the model with constant term but no trend term is selected for unit root test, and the results are shown in Table 1:

Table 1: Unit root test of second-order difference sequence

		t-statistic	**prob.***
augmenteddickey-fuller test statistic		-8.312385	0.0000
test critical values:	1% level	-3.525618	
	5% level	-2.902953	
	10% level	-2.588902	

According to the test results, when the significance level is 5%, the null hypothesis of the existence of unit root is rejected, indicating that the sequence is stationary after second -order difference. In addition, the autocorrelation and partial correlation graphs of second-order difference series can also be seen that the sequence trend is basically stable, so d=2.

3.2.2 Determine the Model

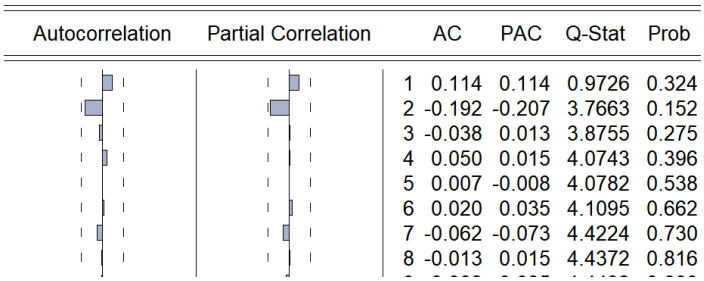
By observing autocorrelation graphs and partial autocorrelation graphs, it is found that both autocorrelation graphs and partial autocorrelation graphs of second-order difference are trailing, and the autocorrelation coefficient is trailing of order 1, while the partial autocorrelation graph is trailing of order 1.Therefore, ARIMA (1, 2, 1) model can be selected for fitting. According to the model, EVIEWSsoftware is used for regression, and the regression results are shown in Table 2. It can be found that each coefficient has passed the significance test of 5%.

Table 2: Regression results of ARIMA (1, 2, 1)

Variable	**Coefficient**	**Std. Error**	**T-statistic**	**Prob.**
Ar(1)	0.684022	0.200097	3.418455	0.0011
Ma(1)	-0.904845	0.202357	-4.471534	0

3.2.3 Residual Sequence White Noise Check

The white noise test results of the residual sequence are shown in Figure 4. In the test results, Q-Stat and P values show that there is no autocorrelation in the residual sequence and it is white noise. The ARIMA model is significantly effective.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.114	0.114	0.9726	0.324
		2	-0.192	-0.207	3.7663	0.152
		3	-0.038	0.013	3.8755	0.275
		4	0.050	0.015	4.0743	0.396
		5	0.007	-0.008	4.0782	0.538
		6	0.020	0.035	4.1095	0.662
		7	-0.062	-0.073	4.4224	0.730
		8	-0.013	0.015	4.4372	0.816

Figure 4: Residual sequence white noise test

3.2.4 Model Prediction

Static forecasts are used in the sample (1952-2022) and dynamic forecasts are used out of the sample (2023-2025).The comparison results between the predicted value and the real value of the ARIMA (1, 2, 1) model are shown in Figure 5. It can be seen that the predicted value and the actual value basically fit, and the ARIMA model predicts that the total population of China at the end of 2023 is 1412, 616, 783.

3.3 Forecast by Lstm Model

Before establishing the LSTM model, the original data is first normalized to avoid partial maximum and minimum values affecting the prediction results.Then, Python software is used forprediction. In this paper, four hidden layers are selected, the output node is 1, and 500 times of training are conducted. Finally, the prediction results are obtained as shown in Figure 5, and the total population is predicted to be 1420,969,780 in 2023.

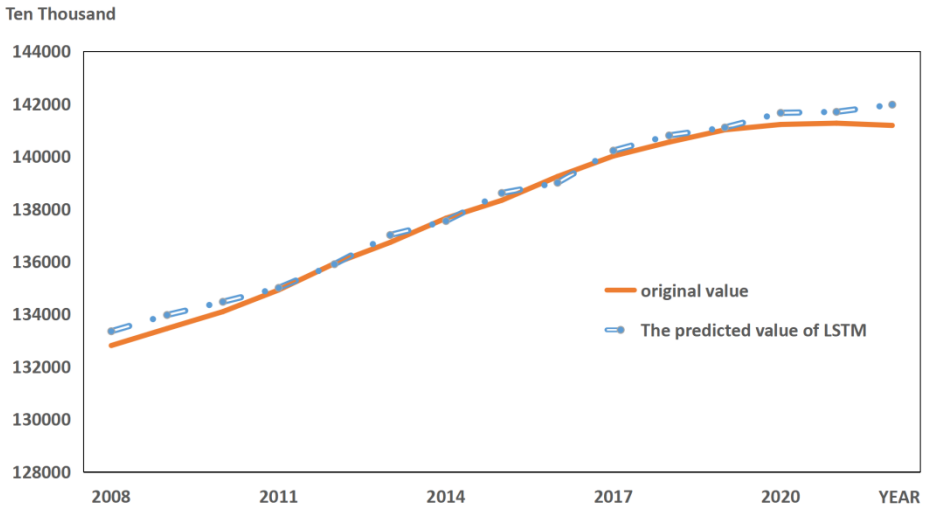


Figure 5: LSTM model prediction results

3.4 Prediction Results of Armi-Lstm Combination

According to the ARIMA prediction results, the residual sequence was obtained after subtracting the ARIMA prediction results from the original data. 2/3 data were selected as the training set and 1/3 data as the test set, which were put into the LSTM model for prediction. The predicted value of the nonlinear part was added to the predicted value of the linear part of ARIMA to obtain the predicted value of the linear optimization combination, as shown in Table 3.

Table 3: Predicted value of linear optimized combination (ten thousand people)

Year	Predicted Value	Year	Predicted Value
2008	132702.2881	2016	139069.8897
2009	133404.51	2017	140000.8357
2010	134014.2189	2018	140735.9997
2011	134649.7144	2019	140954.5284
2012	135687.0602	2020	141364.4697
2013	136869.4086	2021	141402.091
2014	137499.0041	2022	141136.3979
2015	138402.7602	2023	140986.1364

3.5 Prediction Effectiveness Analysis

In order to verify the prediction effect of the model, two error indicators were selected to compare the results of single prediction and combination prediction, namely mean absolute error (MAE) and mean square error (MSE), and the formula was:

$$MAE = \frac{1}{\mathrm{n}} \sum_{t=1}^{\mathrm{n}} \left| x_t - \hat{x}_t \right| \tag{2}$$

$$MSE = \frac{1}{\mathrm{n}} \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(x_t - \hat{x}_t \right)^2} \tag{3}$$

Where, x_t represents the predicted value, x_t represents the actual value, and n represents the number of predicted years. The predicted values of ARIMA, LSTM and linear optimization combination were calculated respectively for MAE and MSE from 2008 to 2022, and the results are shown in Table 4.

Table 4: Comparison of predicted effects

	Mae	Rmse
Arima	181.020318	199.5648347
Lstm	310.3254936	369.9603485
Arima-lstm	127.4313006	144.6774147

The results show that the error value of the combined prediction is smaller than that of the single prediction, which indicates that the effect of the combined prediction is better than that of the single prediction. It also proves the complementary relationship between the ARIMA model and the LSTM model, and it is effective to establish the combined prediction model according to the characteristics of the single model.

3.6 The Total Population Forecast of China

The population of China in 2023 is predicted by using each single prediction method and combination prediction method respectively, as shown in Table 5.

Table 5: Predicted results of China's total population in 2023 (Ten thousand)

	Arima	Lstm	Arima-lstm
2023	141261.6783	142096.978	140986.1364

4. Conclusions and Recommendations

Due to the limitation of individual prediction model, the accuracy of prediction results needs tobeimproved.According to the characteristics of each individual prediction model, combining each individual prediction with certain methods, the combined prediction results obtained are betterthan the results of each individual prediction, so that the final prediction accuracy is improved.

In this paper, ARIMA model and LSTM neural network are combined with linear optimization to predict the total population of China.The empirical results show that the ARIMA-LSTM combination prediction results are better than each single prediction, and the linear optimization combination prediction effect is better, indicating that the time series problem is divided into linear part and nonlinear part respectively forecast and optimization is feasible, and can be further studied.

Population is an important factor affecting economic development. According to the combined forecast results, China's population growth rate will still decline further. Based on this, several policy suggestions are put forward: Relax the restriction on the number of births. The control mentality formed under the long-term family planning policy needs to be opened up in top-level design to gradually reverse. To promote the release of the desire to have children, we should not rely on any coercive or semi-coercive means, but to loosen birth restrictions. So that families who are willing and able to have more children will have no worries at all; improve the social security system. The population problem is not only a problem of birth limits, but also a deeper problem of birth burden. Therefore, behind the low birth rate is actually a continuous decline in the willingness to have children. Improving the social security system, extending maternity leave and guaranteeing the effective supply of labor force can, to a certain extent, improve people's willingness to have children and thus increase the population.

References

[1] Zhang T L. Application of GM (1, 1) model in predicting population birth rate [J]. Chinese Journal of Health Statistics, 2000(02):26-27.

[2] Yin C H, Chen L. Research and application of population prediction model based on BP neural network [J]. Population Journal, 2005(02):44-48.

[3] Rayer S, Smith S K, Tayman J. Empirical Prediction Intervals for County Population Forecasts [J]. Springer Open Choice, 2009, 28(6): 773-793.

[4] Mao J H. Combination prediction of birth rate in China based on IOWA operator [J]. Journal of Changchun Institute of Technology (Natural Science Edition), 2018, 19(01):120-124.

[5] Chen L, Mu T, Li X, et al. Population Prediction of Chinese Prefecture-Level Cities Based on Multiple Models [J]. Sustainability, 2022, 14.