Forecasting The Population of China From 2020 To 2025 Based on Random Forest and Linear Regression

Weibo Wang *

School of Dalian University of Technology, Dalian, China

* Corresponding Author Email: rgrant81743@student.napavalley.edu

Abstract. The population dynamics of a country do play a vital role in its economic and political development. The COVID-19 epidemic has significantly affected the world's population. However, most people only care about the negative influence caused by COVID-19 but ignore the positive influence. This article uses two machine learning models, the random forest model, and the linear regression model, to predict the population change in China if there were no COVID-19 pandemic. With the predicted results, this article can compare the potential positive impact of the pandemic. This paper tries to fit two popular models, namely, Random Roest and Linear Regression to forecast the population of China from 2020 to 2025. The historical birth rate, death rate, and GDP growth rate of China are collected as features adding to the models to decline the error. For the Random Forest model, this paper set up an ensemble of decision trees to predict the future population of China. For the Linear Regression model, our features and population fit a linear connection. The findings of two models are compared in this article, and it is suggested that the random forest model is more suited for population forecasting. In addition, according to this study, the COVID-19 pandemic has some favorable effects on economic growth and birth rates. The article emphasizes the need to not only focus on the negative effects of the pandemic. Furthermore, the article points out that the linear regression model has poor fitting results for non-linear relationships. It suggests exploring more non-linear models for prediction and considering more influential parameters.

Keywords: machine learning; population forecasting; random forest; linear regression.

1. Introduction

The population dynamics of a country play a vital role in its economic and political development [1]. Accurate predictions of population changes provide policymakers and planners with valuable insights. However, The COVID-19 epidemic has significantly affected the world's population. As one of the areas where the pandemic initially took place, China has implemented lots of controlling methods and policies to efficiently stop the rapid spread of the pandemic. Although the pandemic has been successfully controlled, the negative influence on the global population and economy produced by COVID-19 still cannot be ignored.

Hoshino Takuya highlights that the COVID-19 pandemic has been accelerating Japan's population declining trend in the natural population growth rate and carrying out strict border control measures that restrict entry for foreign people [2]. These factors have contributed to the acceleration of population decline and aging in Japan. It is the same as China. The issue of China's aging population, as noted by Warren C. Sanderson and Sergei Scherbov, is particularly problematic. While the population of working age is shrinking, the proportion of those 65 and over in the total population climbed from 7.7% in 2005 to 39.3% in 2060. This leads to long-term sustainability problems, as labor shortages and an increasing elderly population will bring a big burden to society and the economy [3]. What's more, it has been mentioned that such preference for male children has also led to a gender imbalance. This gender imbalance will further affect population structure and social stability [4].

Forecasting by machine learning algorithms has a lot of advantages. Firstly, machine learning algorithms can deal with huge amounts of data automatically. Compared to traditional manual data analysis methodology, machine learning algorithms have successfully improved the efficiency of data analysis. Furthermore, machine learning algorithms can also find out the trends of how the future data is going to change. It means that accurate predictions can be made, and corresponding strategies can be formulated with the help of machine learning. There are numerous benefits to utilizing machine learning algorithms to forecast population dynamics. These algorithms are capable of utilizing vast amounts of data to construct models that can then make predictions based on past population data. Additionally, machine learning algorithms can develop time series models that enhance the accuracy of population predictions [5].

Therefore, it’s necessary to forecast the population of China and understand the population size and characteristics of China in the absence of the COVID-19 pandemic. Then this study can provide possible references for future policy making in China.

This paper utilizes machine learning algorithms to forecast the trends of population change in China. This paper aims to use two popular regression models, namely, Random Forest and Linear Regression to forecast the population of China from 2020 to 2025. In addition, this study tries to explore the influence of population changes by COVID-19, with the hope of providing possible references for future policy-making in China.

2. Method

2.1. Data source

This paper will use historical population data from China from 2010 to 2019 to forecast the population change trends between 2020 and 2025[6]. The historical data from 2010 to 2019 of population and birth rate, death rate, and GDP growth rate are taken from the website of World Bank data (https://data.worldbank.org/). The data are stored in a CSV file and loaded by a Data Frame to be visualized.

2.2. Model

(1) Random Forest

Random Forest can be applied to both classification and regression applications. [7]. It creates a collection of many decision trees to produce a more accurate and robust model to perform predictions [8]. In the construction of each decision tree, The features and data points are chosen at random. This subset of features and data points can increase the tree diversity and reduce overfitting [8] [9]. Additionally, the final prediction result is obtained by voting or the mean or weighted average of all decision trees' prediction results. Thus, Random Forests are robust to noisy data and outliers with the help of the averaging effect of multiple trees [10].

For each decision tree: Choose n samples at random to create a subset from the training dataset. Then randomly select p features and perform feature selection using these p features. Finally, create the decision tree using the subset and the features that were chosen. For each tree, input the data to be predicted, and output a regression result. The final prediction result should then be calculated as the average of all tree regression findings. Each node in the regression tree is divided by specific features and thresholds, which can be represented as:

$$h(x;\theta_j) = \begin{cases} c_j & if\ leaf\ node \\ h_l(x;\theta_j) & if\ x_i \le s_j \\ r(x;\theta_j) & if\ x_i > s_j \end{cases} \tag{1}$$

Where c_j is the constant term of the leaf node, s_j is the feature-splitting threshold, $h_l(x;\theta_j)$ and $h_r(x;\theta_j)$ are the prediction functions for the left and right subtrees, and θ_j represents the parameters of the tree. If there are d features, d decision trees will be built, and the final prediction result will be the average of these decision trees:

$$\hat{y_r} f(x) = \frac{1}{d}\sum_{j=1}^{d} h(x;\theta_j) \tag{2}$$

where d is the number of trees, and $\hat{y_r} f(x)$ is the prediction result for the point x.

(2) Linear Regression:

Linear Regression is a widely used method to forecast future data. It can find a linear function to describe the relationship between two or more variables [11]. The linear regression model is widely applied in the field of data analysis. It can be used to do prediction, relationship analysis, parameter estimation, and descriptive analysis [11].

One independent variable and one dependent variable make up a simple linear regression model:

$$y = \beta 0 + \beta 1x + \varepsilon \tag{3}$$

Where y is the response variable, x is the predictor variable, β0 is the intercept, β1 is the slope, and ε is the error term.

In multiple linear regression models, there are multiple independent variables (x1, x2, ..., xn) and one dependent variable (y):

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon \tag{4}$$

Where β0 is the intercept, β1 to βn are the slope coefficients, and ε is the error term.

3. Results

3.1. Visualization

Data on China's population, birthrate, deathrate, and GDP growth rate from 2010 to 2019 are visualized here.

Fig.1(a) is a line graph depicting the change in China's population from 2010 to 2019. China's population has continued to grow from 2010 to 2019. Moreover, the population growth rate has somewhat declined since 2018.

Fig.1(b) is a line graph depicting the change in China's birth rate from 2010 to 2019. It shows that the birth rate first rose and then fluctuated up and down from 2010 to 2019. This trend is generally downward, and the birth rate hit its lowest point in 2019.

Fig.1(c) is a line graph depicting the change in China's death rate from 2010 to 2019. It indicates that the death rate first increased, then decreased, and then increased again from 2010 to 2019. It hit its lowest point in 2016, and then gradually increased, presenting an overall downward trend.

Fig.1(d) is a line graph representing changes in China's economic growth rate from 2010 to 2019. It illustrates that China's economic growth rate showed a downward trend from 2010 to 2019, with the decline from 2011 to 2018 being slower, followed by a sharp drop from 2018 to 2019.

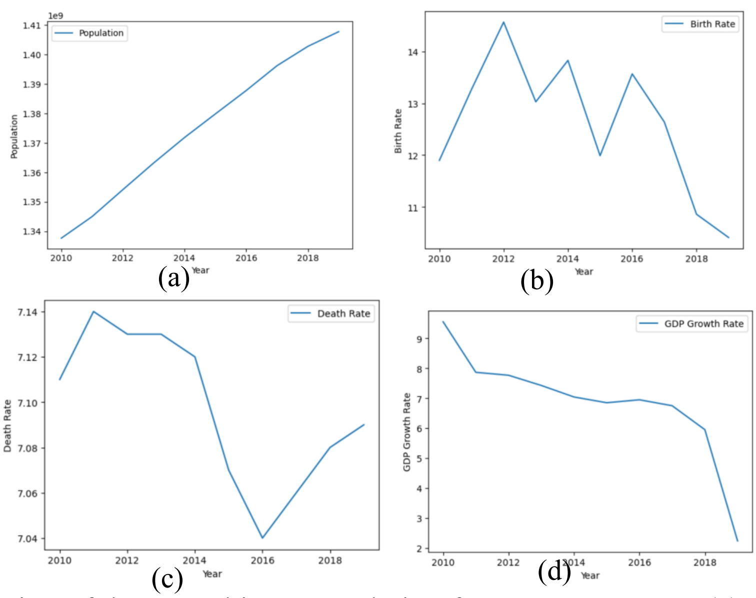


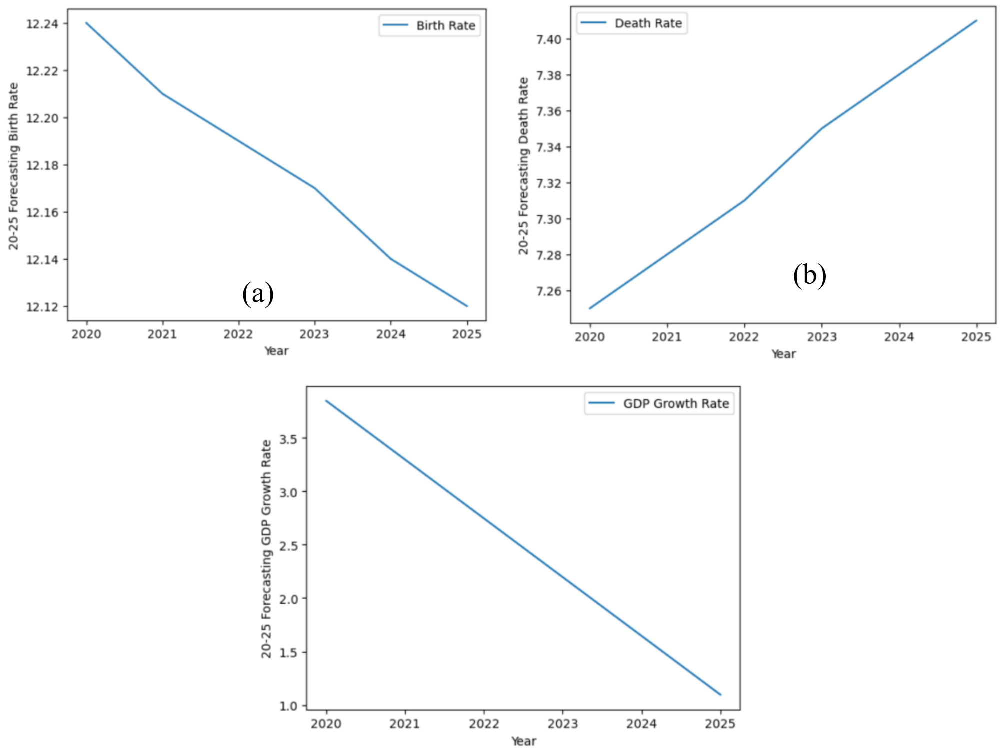
Fig. 1 The visualization of data on China's population from 2010 to 2019: (a) population, (b) birth rate, (c) death rate, and (d)GDP growth rate

Then this study uses linear regression models to predict the values of various features from 2020 to 2025 based on data from 2004 to 2019.

Fig.2 (a) shows that the birth rate in China is expected to decline from 2020 to 2025, but it will remain relatively stable at over 12%. Compared to the birth rate in 2019, there is a noticeable increase.

Fig.2 (b) displays a gradual increase in the mortality rate in China from 2020 to 2025, although the rate of increase is relatively slow.

Fig.2 (c) indicates that China's economic growth rate is projected to further decline from 2020 to 2025. However, the overall trend shows that the country's economy is still growing positively and is expected to stabilize in the future.



c

Fig. 2 The visualization of data prediction on China's population from 2020 to 2025 :(a) birth rate, (b) death rate, and (c)GDP growth rate

3.2. Evaluation

Before performing specific parameter evaluations on the two models, this article used historical data from 2004 to 2015 to predict the data from 2015 to 2019 and used this to evaluate the models.

The visualization in Fig.3 (a) displays the predicted population data from 2015 to 2019 using the random forest model. It shows a slight decrease in population from 2014 to 2015, followed by an upward trend in 2016 before experiencing a decline, and finally a slight recovery.

Fig.3 (b) shows the results of using a linear regression model to predict the population from 2015 to 2019, which indicates a consistent upward trend in population during this period. However, It's crucial to remember that these estimated values might not match the population figures perfectly.

Fig.3 (c) depicts a visualization chart that combines the actual population numbers from 2015 to 2019, the predicted population numbers from the random forest model, and the predicted population numbers from the linear regression model. It is apparent from the chart that the random forest model fits the actual situation better.

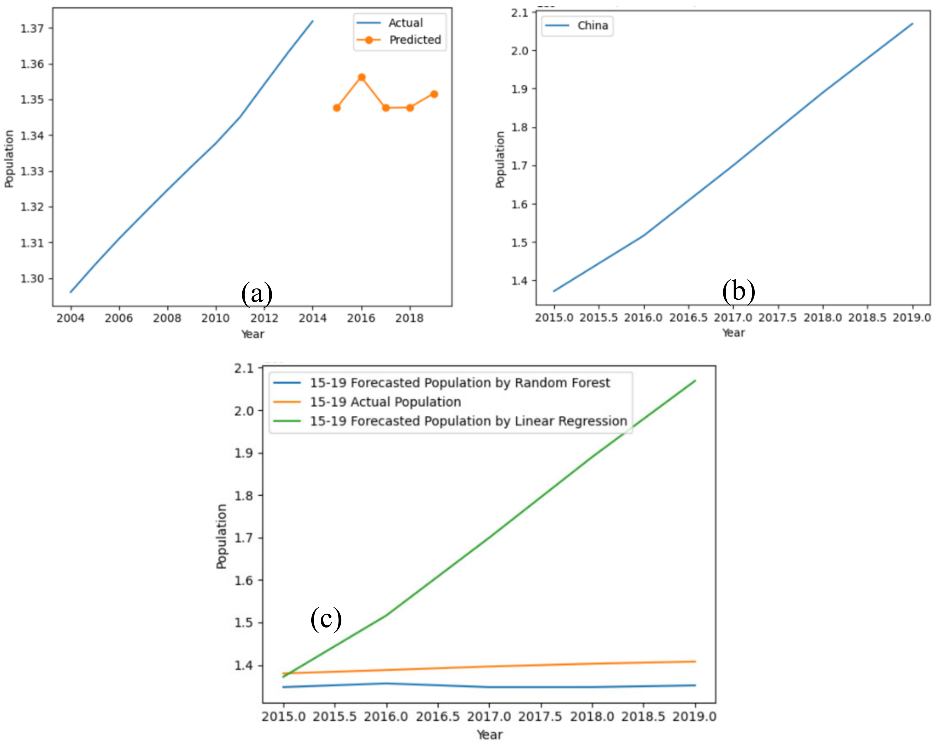


Fig. 3 The predicted population

3.3. Accuracy Evaluation

3.3.1 Equation

Before making final predictions, this study attempts to evaluate the prediction performance of random forest and linear regression models using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and other analysis parameters to quantitatively assess the effectiveness of the two models. First, the concepts of MSE, RMSE, and MAE will be introduced, as well as the judging criteria.

(1) MSE

The MSE is the average of the squared differences between the predicted value and the true value.

The formula for calculating MSE is as follows:

$$MSE = (1/m) * \sum(yi - \bar{y})^2 \tag{5}$$

Where m is the number of samples, yi is the actual observed value, and ȳ is the predicted value.

The smaller the value of MSE, the smaller the difference between the predicted value and the true value, indicating that the model has higher prediction accuracy.

(2) RMSE

(3) The RMSE is the square root of the MSE. The formula for calculating the RMSE is

$$RMSE = \sqrt{(MSE)} \tag{6}$$

The calculation method of the RMSE is similar to that of the MSE, but due to the square root operation, the value of the RMSE is in the same unit as the original data. The discrepancy between the anticipated value and the true value is smaller the lower the RMSE value, indicating that the model has higher prediction accuracy.

(4) MAE

The MAE is the average of the absolute differences between the predicted value and the true value. The formula for calculating the MAE is:

$$MAE = (1/m) * \sum|yi - \bar{y}| \tag{7}$$

Where yi is the actual observed value, and ȳ is the predicted value.

The difference between the projected value and the true value decreases with decreasing MAE values, suggesting improved prediction accuracy for the model.

• MSE and RMSE: Both indicators are non-negative. Generally speaking, the model's prediction accuracy increases with decreasing MSE and RMSE values. However, since the values of MSE and RMSE are related to the units of the data, it is not possible to judge the quality of the model based solely on the size of these values. It is necessary to compare them with other models.

• MAE: The model's prediction accuracy increases as the MAE value decreases. Unlike MSE and RMSE, the value of MAE is not affected by the units of the data, so the quality of the model can be directly judged based on the numerical evaluation.

3.3.2 Evaluation

Next, this study will compare the predicted population changes from 2015 to 2019 using the random forest and linear regression models with the actual population changes based on the data of the population, birth rate, death rate, economic growth rate from 2004 to 2014 over a period of 10 years, and the birth rate, death rate, and economic growth rate from 2015 to 2019 in Fig.4.It is evident from the chart that the Random Forest model has a better fit in predicting the population quantity and its trends from 2015 to 2019, while the Linear Regression model has a poor fit. This can largely be attributed to the complexity of the models. The Random Forest model is a collective learning approach that integrates various predictive decision tree methods. This leads to higher flexibility in capturing different data patterns and non-linear modeling capabilities to capture non-linear relationships. In contrast, because the Linear Regression model assumes a linear relationship between variables, its ability to fit data when faced with a non-linear relationship is considerably reduced. As a result, it may not generate a good population change pattern.

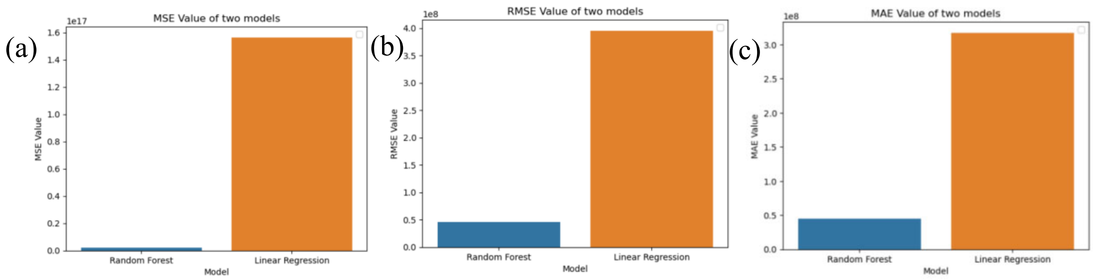


Fig. 4 The value of MSE, RMSE and MAE

Comparing the values of MSE, RMSE, and MAE, it is evident that the random forest model outperforms the linear regression model in terms of accuracy in population prediction. The precision of the random forest model is higher, making it more advantageous in predicting population trends.

3.4. The result of prediction

The following table will show in detail the predicted population numbers from two models.

Table 1 The predicted population numbers from two models

Year	The population of Random Forest	The population of Linear Regression
2020	1 billion, 390 million, 369 thousand	1 billion, 407 million, 745 thousand
2021	1 billion, 390 million, 370 thousand	1 billion, 563 million, 817 thousand
2022	1 billion, 390 million, 373 thousand	1 billion, 737 million, 738 thousand
2023	1 billion, 390 million, 372 thousand	1 billion, 931 million, 608 thousand
2024	1 billion, 390 million, 374 thousand	2 billion, 147 million, 781 thousand
2025	1 billion, 390 million,551 thousand	2 billion, 388 million, 895 thousand

From Fig.5 (a), the predictions from the random forest model suggest a decline in China's population from 2020 to 2025 compared to 2019, followed by a relatively stable state until a slight increase in 2025. Fig.5 (b) The linear regression model predicts a significant population growth in

China, surpassing 2 billion in 2024. In the end, this article employs data visualization to draw a chart combining the predicted results from two models.

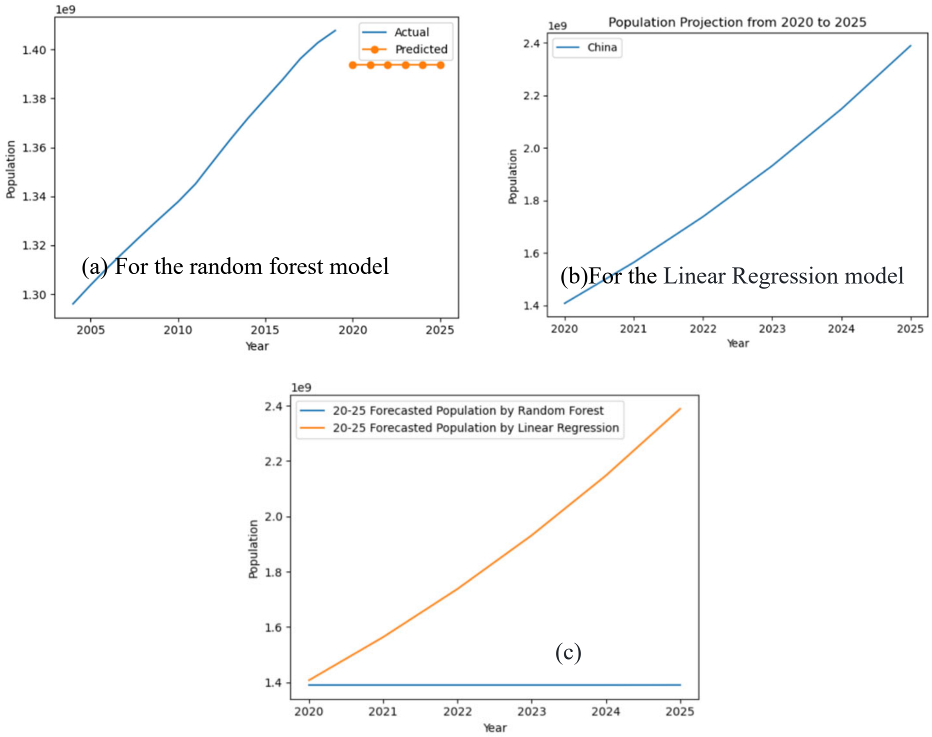


Fig. 5 The visualization of prediction by two models

4. Discussion

From a theoretical standpoint, the random forest model is a collection of decision tree models, which allows for better modeling of nonlinear relationships between data and improved fitting accuracy compared to linear regression models [9]. Therefore, this article mainly focuses on the predictive data generated from the random forest model. Based on this data, if there had been no COVID-19 outbreak, China's population and economic growth rate would have decreased from 2019 to 2020, and the population would have remained relatively stable at around 1.3 billion from 2020 to 2025. However, the COVID-19 pandemic has led to an increase in the mortality rate but has also stimulated an increase in China's birth rate to some extent. From 2019 to 2020, China's population increased instead of decreased, and the economic growth rate increased from 2.23 to 8.44. This highlights the positive impact of the COVID-19 pandemic on certain aspects of China and serves as a reminder that we should not only consider the negative effects of the pandemic but also the positive ones.

In terms of predicting population trends, several suggestions are proposed. Firstly, when conducting population forecasts, a variety of influencing factors should be considered, such as birth rates, mortality rates, economic growth rates, and migration rates to obtain more accurate predictions [1]. Secondly, greater attention should be paid to nonlinear relationships when predicting population trends. As population growth is often characterized by nonlinear relationships, it is crucial to establish models that can better fit the data. Lastly, it is essential to note that different models may produce different results, highlighting the importance of accounting for uncertainty when predicting population trends.

5. Conclusion

This paper uses two machine learning models, Random Forest, and Linear Regression, to forecast the population of China in the absence of COVID-19 from 2020 to 2025. This article used the birth rate, death rate, and economic growth rate from China between 2010 and 2019 to predict future population trends. The random forest model predicted a slight population decline followed by stability, while the linear regression model predicted continued population growth. Based on the MSE, RMSE, and MAE considered in this article, the random forest model had higher fitting accuracy than the linear regression model. However, from the actual population trends between 2020 and 2022, the linear regression model appeared to better fit the population changes.

There are several limitations to this article. Firstly, the linear regression model is a machine learning model that fits linear relationships, so it may not perform well in fitting non-linear data, such as population changes. Additionally, the parameters considered may not be comprehensive enough to fully explain the factors affecting population changes. There are many other factors such as changing social attitudes, the reluctance of young people to have kids, and immigration and emigration patterns that might also affect population changes, which this study did not address. Therefore, the interpretation ability of the model used in this study may be limited.

Going forward, future research can adopt more advanced machine learning models and consider more comprehensive parameters to enhance the fitting ability of the model. Predicting population changes without the influence of the COVID-19 pandemic can also help better understand the pandemic's impact on the population and assist countries in handling future changes and setting relevant policies to adjust population structure.

References

[1] Keyfitz N. The social and political context of population forecasting. 1984.

[2] Hoshino T. The COVID-19 Pandemic is Accelerating Japan’s Population Decline: A Statistical Analysis, 2021. https://www.nippon.com/en/in-depth/d00701/

[3] Li Q, Reuser M, Kraus C, et al. Ageing of a giant: a stochastic population forecast for China, 2006–2060. Journal of Population Research, 2009, 26: 21-50.

[4] Luo Y, Su B, Zheng X. Trends and challenges for population and health during population aging—China, 2015–2050. China CDC Weekly, 2021, 3(28): 593.

[5] Grossman I, Bandara K, Wilson T, et al. Can machine learning improve small area population forecasts? A forecast combination approach. Computers, Environment and Urban Systems, 2022, 95: 101806.

[6] Daghistani T, Alshammari R. Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. Journal of Advances in Information Technology ,2020, 11(2): 78-83.

[7] Schonlau M, Zou R Y. The random forest algorithm for statistical learning. The Stata Journal, 2020, 20(1): 3-29.

[8] Biau G, Scornet E. A random forest guided tour. Test, 2016, 25: 197-227.

[9] Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random Forest[C]//Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3. Springer Berlin Heidelberg, 2012: 246-252.

[10] Biau G. Analysis of a random forests model. The Journal of Machine Learning Research, 2012, 13: 1063-1095.

[11] Montgomery D C, Peck E A, Vining G G. Introduction to linear regression analysis. John Wiley & Sons, 2021.