Forecast and Analysis of Reasons for Changes in the Number of Students Studying Abroad

Zhaohui Li

School of Mathematical Sciences, University of Southampton, Southampton, UK

454875250@qq.com

Abstract. Since 2000, the number of students studying abroad in China has been growing continuously. Based on the macro data of National Bureau of Statistics from 2005 to 2021, this paper analyzes the degree of influence of variables such as economy, epidemic, and policies on the number of students studying abroad, and constructs a GM (1, 1) model using SPSSAU to calculate the predicted values of variables for the next two years, and Multiple linear regression models, machine learning models, and time series models were constructed using Python, Eviews, and SPSS software to predict the future trend of the number of study abroad students; and based on this, a combination model based on weight assignment was constructed. The research results prove that the combined model has higher accuracy than the single model.

Keywords: Study Abroad; Multiple Regression; Machine Learning; Time Series; Gray Model; Combined Model.

1. Introduction

1.1 Background and Significance of the Study

Today's world is the world of education globalization, and China's modern "study abroad" education began after the Opium War and flourished between the Sino-Japanese War of 1894-1895 and the War of Resistance against Japan [1]. In 1984, the State Council promulgated the *Provisional Regulations on Self-financed Study Abroad*, which opened the door to self-financed study abroad [2]. In the 21st century, the restriction on self-financed study abroad was completely abolished, and self-financed study abroad became the main force among Chinese students. The booming economy of China has greatly promoted the prosperity of self-financed study abroad, and the *Report on the Development of Study Abroad* in China 2017 shows that the proportion of self-financed study abroad has exceeded 90% in recent years. Studying abroad has become a way for many students to pursue higher education. *The China Study Abroad Development Report 2020-202*1[3] states that 80.7% of those who return to China for employment have a master's degree, and the White Paper on Study Abroad in China states that most of the study abroad population has a target degree of master's degree.

However, the sudden epidemic has added much uncertainty to the path of many students studying abroad. Today, the Ministry of Education has been consistent in cultivating international talent with an open and inclusive attitude, guided by the new era's policy of "supporting study abroad, encouraging return to China, freedom to come and go, and play a role" in response to the controversy surrounding the topic of study abroad. Therefore, it is important to investigate the factors influencing study abroad in order to judge the trend of study abroad, adjust study abroad plans and formulate relevant policies. At the same time, this paper collects data and information for processing, and uses statistical knowledge to predict the number of students studying abroad in the next few years, so that people can have a certain knowledge of the heat and situation of studying abroad, and can look at studying abroad objectively and rationally in the future.

1.2 Review of Domestic and International Studies

Domestic scholars have conducted relatively few studies in this area of the number of students studying abroad, and most of them are based on qualitative analysis. Zhu Zhengzhong (2016) explored the causes, problems, and countermeasures of young students studying abroad in China in

the context of globalization [4]. Li Xuemei (2019) focused on the motivation and influencing factors of younger students studying abroad [5]. Meanwhile, some scholars have achieved some results in studying the influencing factors and the prediction direction of the number of students studying abroad by building models. Liu Xuezhi (2016) used regression analysis to explore in depth the quantitative relationship between study abroad, return from study abroad and economic and social development [6]. Chen Yanglin and Yue Changjun (2020) used logistic regression analysis to investigate the influence of demographic characteristics, family background, human capital, and teaching satisfaction on college students' study abroad [7]. Li Lanxiang (2013) developed a multivariate linear model based on econometric linear regression theory [8] to predict the number of study abroad before the epidemic.

As a contrast, for the study of study abroad, foreign scholars started earlier and study abroad is more normalized in western countries. dwedarDahlia devoted to the study of the impact of study abroad on the learning of second foreign language, and his summarized findings covered multiple aspects such as language and interaction [9], comparing the advantages and disadvantages of international education and local education. Foreign scholars are more interested in the effects of immersion in multilingual spaces and multilingual values than in China, and thus their research focuses mostly on language proficiency and study abroad experiences, often combined with literature, philosophy, pedagogy, sociology, and other fields.

2. Theoretical Foundations

2.1 Gray Prediction GM(1,1)

A gray prediction model is a forecasting method that builds a mathematical model to make predictions from a small amount of incomplete information. The GM(1,1) model is one of the most widely used gray dynamic prediction models in gray system theory, which consists of a univariate first-order differential equation [10]. The gray model does not use the original data series, but generates a new data series from the original series, and thus can be used without considering the distribution pattern and change trend of the original data. The model does not need a lot of data, at least four data are sufficient. It can solve the problems of few original data and low completeness and reliability of the sequences. The advantages of the gray model are high accuracy, the ability to generate regular series with irregular raw data, applicability to exponential growth [11], and ease of testing. The disadvantage is that it can only describe monotonic changes and can only be used for short- and medium-term forecasting.

2.2 Multiple Linear Regression

In practical problems, the change in a dependent variable is often influenced by multiple independent variables, and then it is necessary to explain the change in the dependent variable by multiple independent variables [12], which is multiple regression, and when the relationship between multiple independent variables and the dependent variable is linear, the regression analysis performed is multiple linear regression. The equation of the multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where β_0 is the constant term, $\beta_1, \beta_2 \ldots \beta_p$ is the regression coefficient, and $\varepsilon \sim N\left(0, \sigma^2 I_n\right)$. The least squares method is used to find the parameters of the regression coefficients, and takes the partial derivatives of the $\beta_1, \beta_2 \ldots \beta_p$ to obtained the solution of parameters are $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$.

2.3 Support Vector Regression

The linear function y=wx+b is the model function of support-vector-regression. Unlike the strict linear regression, support-vector-regression is more tolerant, and the principle of calculating residuals

is different from that of linear function. It allows the sample point y value and the sample function output y value at most ϵ deviation, deviation is less than or equal to the value of the residuals are not counted. Only when the deviation is greater than ϵ, the residuals are counted. Introducing the slack variables ξ_i and ξ_i^* to denote the residuals, the values of the residuals are obtained as:

$$\xi_i = \begin{cases} y_i - \left(f\left(x_i\right)+\epsilon\right), & y_i > f\left(x_i\right)+\epsilon \\ 0, & \text{otherwise} \end{cases}, \quad \xi_i^* = \begin{cases} \left(f\left(x_i\right)-\epsilon\right)-y_i, & y_i < f\left(x_i\right)-\epsilon \\ 0, & \text{otherwise} \end{cases}$$

The optimized model is obtained by minimizing the total loss as a way to determine the parameters w and b. The optimization objective of the support vector regression is:

$$\min_{w,b,\xi,\xi,*} \frac{1}{2}\| \mathrm{w}\|^2 + \mathrm{c}\sum_{i=1}^{\mathrm{n}}\left(\xi_i + \xi_i^*\right)$$

$$\text{s.t.} f\left(x_i\right) - y_i \leqslant \epsilon + \xi_i; y_i - f\left(x_i\right) \leqslant \epsilon + \xi_i^*; \xi_i \geqslant 0; \xi_i^* \geqslant 0, \quad i = 1,2,\ldots$$

Note that the solved f(x) model may not be linear and is likely to be nonlinear. We use the approach of increasing the dimensionality with a low-dimensional to high-dimensional mapping , so that the nonlinear model becomes linear. The mapping of it is the high-latitude space and then calculate the inner product is very computationally intensive, and if there exists a function , such that

$$K\left(x_i, x_j\right) = \phi\left(x_i\right)^T \phi\left(x_j\right)$$

Then it is called the kernel function. The inner product with in the high-dimensional space is equal to their result calculated in the original sample space by the kernel function [13], which greatly reduces the computational effort.

2.4 Smooth Time Series

The autoregressive model uses the values of the series itself to regress the variables. The p-order autoregressive model AR(p) indicates that there is a linear relationship between X_t and $X_{t-1}, X_{t-2}\ldots X_{t-p}$, independent of X_{t-j} (j=p+1,p+2.....), under the known conditions $X_{t-1}, X_{t-2}\ldots X_{t-p}$ .The mathematical model is given by

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + \varepsilon_t \ \circ$$

where the constant term ϕ_0 , $\phi_0, \phi_1, \phi_2 \ldots \phi_p$ is the parameter coefficient , ε_t is a white noise series independent of $X_{t-1}, X_{t-2}\ldots X_{t-p}$ with mean 0 and variance σ2. MA(q) differs from AR(p) in that in MA(q) is independent of the past moment observations $X_{t-1}, X_{t-2}\ldots X_{t-p}$ and linearly related to the past moment perturbation terms $\varepsilon_t, \varepsilon_{t-1}, \ldots \varepsilon_{t-q}$ .MA(q) model function is:

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \ldots\ldots - \theta_q \varepsilon_{t-q}.$$

The autoregressive moving average model is based on the autoregressive model and the moving average model, and its order consists of p, q. p represents the order of the autoregressive component and q represents the order of the moving average component, and is denoted as ARMA(p, q). the expression of ARMA(p, q) is

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \ldots\ldots - \theta_q \varepsilon_{t-q} \ \circ$$

The ARMA model only needs to depend on endogenous variables without the help of other exogenous variables, but it requires that the time series data are smooth. If the time series is not smooth, the smooth time series can be obtained by differencing [14], and the model based on the smooth time series obtained by differencing is denoted as ARIMA (p, d, q), where d is the order of difference and the other is consistent with the ARMA(p, q) model.

3. Analysis of Study Abroad Data and Prediction of the Number of Students

3.1 Selection of Variables

In this article, the factors influencing study abroad are grouped into four categories: economic factors, schooling factors, policy influence, and epidemic influence. The time period intercepted in the article is 2005-2021, and the relevant data is used as the benchmark for analysis. We use the per capita disposable income of residents and the RMB exchange rate (with the US dollar as the reference) to measure the economic indicators. This paper argues that economic factors are the key factors that dominate Chinese students' intention to study abroad. For the willingness to study, this paper uses the number of domestic graduate students as its indicator. Study abroad policy is also one of the factors examined in this paper. The main influence of foreign policy is to change the destination of students, which does not have much effect on the total number of students and does not affect the scale of international students. In 2013, President Xi emphasized the policy of "supporting study abroad, encouraging return to China, freedom to come and go, and playing a role" at the 100th anniversary of the founding of the European and American Association of Students [15], and this paper will focus only on this policy as a representative. This paper will focus only on this policy.

3.2 Data Collection

After completing the analysis of causes and selecting variables, we consider building a model to make appropriate and reasonable predictions on the changes of the number of students studying abroad. In order to ensure the accuracy of the prediction, we will involve as many variables as possible, both qualitative and quantitative, while maintaining the accuracy of the prediction. As mentioned above, this paper will focus on the data from 2005 to 2021, and for the numerical variables, the following data were collected (Table 1).

Table 1. Data table of variables related to studying abroad

Year	Number of students studying abroad	Per capita disposable income	Exchange rate
2005	11.85	6385.00	819.2
2006	13.40	7229.00	797.2
2007	14.40	8584.00	760.4
2008	17.98	9957.00	694.5
2009	22.93	10977.00	683.1
2010	28.47	12520.00	677.0
2011	33.97	14551.00	645.9
2012	39.96	16510.00	631.3
2013	41.39	18311.00	619.3
2014	45.98	20167.00	614.3
2015	52.37	21966.00	622.8
2016	54.45	23821.00	664.2
2017	60.84	25974.00	675.2
2018	66.21	28228.00	661.7
2019	70.35	30733.00	689.9
2020	45.09	32189.00	689.8
2021	52.37	35128.00	645.1

3.3 Time Series Analysis

Data on multiple variables have been collected in the above section. Now consider: is it feasible to directly predict changes in the number of international students without identifying multiple independent variables? We use ARIMA model to analyze the data. After plotting the change curve of the number of international students over the years (Figure 1), it is found that the series is a non-stationary time series and thus needs to be differenced. Second-order differencing was performed on the data to confirm whether it was a smooth time series. The unit root test is used to prove that the

second-order difference series is a smooth time series with intercept of 0. After that, model sizing is performed and the preliminary time series model may be ARIMA(1,2,1), ARIMA(1,2,0) or ARIMA(0,2,1).

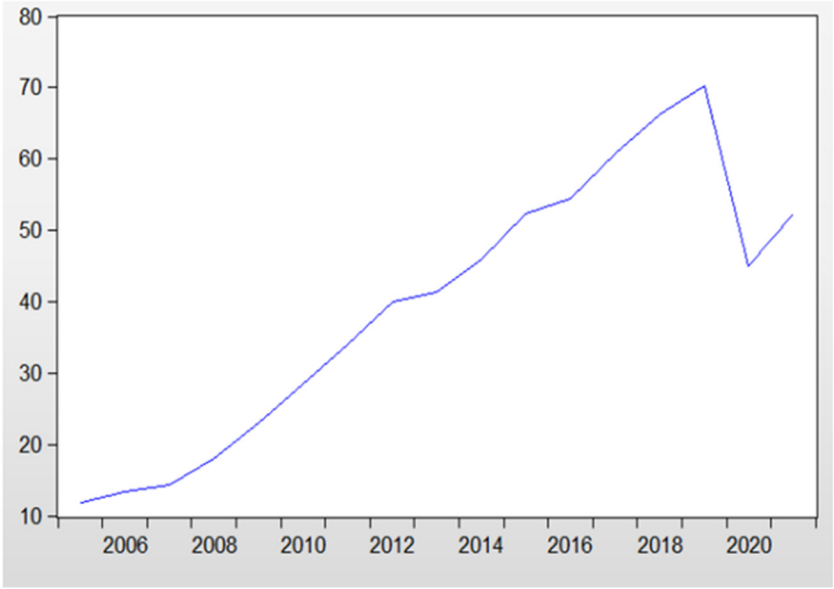

Figure 1. number of students studying abroad

The significance of the coefficients of the ARIMA(1,2,0) model was tested and passed, and let the number of international students be Yt and the second-order difference series be Xt (Xt=(1-B)2Yt-1). The coefficient test for Xt is shown in Table 2. The function of Xt obtained after the coefficient test is $Y_t=Y_{t-1}+0.642Y_{t-2}-0.821Y_{t-3}+\varepsilon_t$, then the expectation of the function about Yt is $E(Y_t)=1.179Y_{t-1}+0.642Y_{t-2}-0.821Y_{t-3}$.

Table 2. Parameter Inspection

Variable	Coefficient	Std. error	t-Statistic	Prob.
AR(1)	-0.821227	0.232066	-3.538758	0.0036
SIGMASQ	71.43961	11.54747	6.186604	0.0000
R-squared	0.465319	Mean dependent var		0.382000
Adjusted R-squared	0.424189	S.D. dependent var		11.96475
S.E. of regression	9.079115	Akaike info criterion		7.448205
Sum squared resid	1071.594	Schwarz criterion		7.542611
Log likelihood	-53.86154	Hannan-Quinn criter		7.447199
Durbin-Watson stat	2.304924			

We compared the model predicted values with the actual true values, and in Table 3 we see that the error in the data for 2020 is severe, with a relative error of 67.31%. This indicates that the sudden appearance of the epidemic affects the normal growth pattern of the dependent variable and thus the data for the three years 2019, 2020 and 2021 cannot be used to predict the values for 2022. The model indicators in Table 4 show the superiority and accuracy of the time series model. This paper has reservations about the accuracy of the model. However, this does not mean that the model loses its significance. The model function is

$$E(Y_t)=1.179Y_{t-1}+0.642Y_{t-2}-0.821Y_{t-3}\ \circ$$

According to the model parameters, if the number of people studying abroad increases in 2022, it will continue to increase in 2023 and 2024. It is believed that with the gradual stabilization of the world epidemic, the change of the number of Chinese going abroad to study will also return to the original rule.

Table 3. Relative error

Year	True value	Predicted value	Relative error
2005	11.85	8.77	25.95%
2006	13.40	9.23	31.16%
2007	14.40	11.99	16.73%
2008	17.98	13.85	22.96%
2009	22.93	17.44	23.93%
2010	28.47	24.76	13.05%
2011	33.97	31.53	7.20%
2012	39.96	39.5	1.15%
2013	41.39	45.54	10.03%
2014	45.98	46.56	1.26%
2015	52.37	47.98	8.38%
2016	54.45	57.28	5.20%
2017	60.84	60.07	1.27%
2018	66.21	63.7	3.79%
2019	70.35	72.41	2.93%
2020	45.09	75.44	67.31%
2021	52.37	43.96	16.06%

Table 4. Statistical indicators

Statistical indicators	
MAPE	15.20%
MAE	4.82
MSE	66.73
RMSE	8.17

3.4 Prediction of Variables based on Gray Prediction

After establishing the time series model, we find that a single time variable cannot accurately explain the change in the number of international students. So we need to use a model containing multiple independent variables. For the three numerical variables of per capita disposable income of residents, RMB exchange rate, and the number of graduate students, this paper uses the gray prediction model GM (1, 1) to predict their values in 2022 and 2023. We model the per capita disposable income of the residents, and Table 5 shows the test indicators of the model with a posteriori difference ratio C value of 0.001 <= 0.35, implying a very good model accuracy grade. The small error probability p-value is 1.0 > 0.95, implying that the model accuracy is very good, and the model can be used to predict the per capita disposable income of residents in 2022 and 2023. The GM (1, 1) models were established for the number of examiners and the RMB exchange rate, respectively, and the above steps were repeated to finally obtain the predicted values of the variables in 2022 and 2023 (Table 6 Predicted values of variables).

Table 5. Results of GM (1,1) model of per capita disposable income of 5 residents

The development coefficient a	gray action b	Posteriori difference ratio C	Small error probability p-value
-0.0341	39859.5275	0.001	1.000

Table 6. Predicted values of variables

Year	Per capita disposable income	The exchange rate	Number of postgraduate students
2022	37388.78	629.803	511.071
2023	39907.93	624.933	593.985

3.5 Support Vector Model

Among the variables we want to explore, only the three independent variables of per capita disposable income, RMB exchange rate, and the number of examinations have quantiles, while epidemic and policy have no numerical characteristics. We first try to build the model with known quantitative variables, but this will inevitably ignore the influence of epidemic and policy, so we choose a more lenient support vector model to try. Let the dependent variable y be the number of international students, the independent variable x1 be the per capita disposable income of residents, x2 be the RMB exchange rate,x3 be the number of examinations. It is stated in econometrics that most indicators of economic variables grow exponentially [16], and taking this factor into account, we take the logarithm of the per capita disposable income of the population to convert its growth trend to linear growth. For the support vector model, choosing different kernel functions will get different models and different fitting results, we choose linear kernel function, polynomial kernel function, and Gaussian kernel function for fitting, and the fitting effect is shown in Figure 2 SVR fitting curve shows that the predicted value curve is closer to the true value curve when choosing Gaussian kernel function than choosing linear kernel function and polynomial kernel function. Table 7 measures the model fitting effect by MAE (mean absolute error), RMSE (root mean square error), MAPE (mean absolute percentage error), MSE (mean squared error), and R-squared (coefficient of determination). Table 7 shows that the R2 of the support vector model is closer to 1 when the Gaussian kernel function is selected, and the mean absolute error MAE, root mean squared error RMSE, mean absolute percentage error MAPE, and mean squared error MSE are much smaller than the values when the linear kernel function and polynomial kernel function are used at this time. Therefore, it indicates that the support vector model fits best when the Gaussian kernel function is selected.

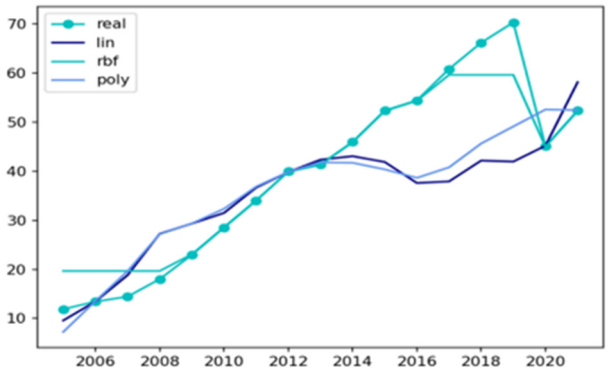

Figure 2. SVR fitting curve

Table 7. Statistical indicators

Index of evaluation	Linear	Poly	Rbf
MAE	8.27	7.90	2.38
RMSE	147.92	114.19	17.02
MAPE	19.60%	20.62%	11.03%
MSE	196.30	258.96	17.01
R squared	0.56	0.66	0.95

The support vector model with Gaussian kernel function is analyzed and the relative error between the predicted value and the real value is compared. It is pointed out in Table 8 that the relative error between the real value and the predicted value in recent years is controlled within 20%, which proves that the prediction of the data in recent years by the support vector model is in line with the requirements. But there are some problems with this model. On the one hand, as shown in Table 8, the maximum relative error exceeds 50%, which is obviously not required. On the other hand, we find that the model has the problem of multiple groups with the same predictive value. So we can't rule out the possibility of large error when using this model to predict.

Table 8. Relative error of the model

Year	True value	rbf	Relative error
2005	11.85	19.64	65.77%
2006	13.40	19.64	46.60%
2007	14.40	19.64	36.42%
2008	17.98	19.64	9.26%
2009	22.93	23.03	0.43%
2010	28.47	28.57	0.35%
2011	33.97	34.07	0.29%
2012	39.96	39.86	0.25%
2013	41.39	41.29	0.24%
2014	45.98	45.88	0.22%
2015	52.37	52.27	0.19%
2016	54.45	54.35	0.18%
2017	60.84	59.64	1.97%
2018	66.21	59.64	9.92%
2019	70.35	59.64	15.22%
2020	45.09	44.99	0.22%
2021	52.37	52.27	0.19%

3.6 Multiple Linear Regression

We created a support vector model with only quantitative variables, which has a certain accuracy in predicting the data in recent years. However, this model also has disadvantages that cannot be ignored: (1) it ignores the role of qualitative variables. The epidemic has the most direct impact on the number of students studying abroad. After the outbreak of the epidemic, the number of students studying abroad plummeted in 2020, with about 450,900 choosing to continue studying abroad, down 252,600 from 703,500 in 2019. It is clear that the impact of the epidemic cannot be questioned, so the impact of the epidemic cannot be ignored when studying the number of people studying abroad. (2) Because of the existence of kernel function, it is impossible to observe the interpretation degree of independent variable to dependent variable. (3) Multiple forecast data are equal. It can be confirmed that qualitative variables have an impact on the problem studied in this paper, so it is lack of rigor to interpret dependent variables only with quantitative data. At this time, dummy variables are introduced into multiple linear regression model to solve this problem. For epidemic situation, the policy two dummy variables are D1 and D2 respectively. The policy of "Support overseas study, encourage return to China, come and go freely, play a role" was promulgated in 2013, and the epidemic appeared in early 2020. Therefore, it is assumed that: the dependent variable y is the number of international students, the independent variable x1 is the per capita disposable income of residents, x2 is the exchange rate of RMB, and x3 is the number of postgraduate students. The model function is:

$$y = f\left(\ln\left(x_1\right), x_2, x_3, D_1, D_2\right)$$

Table 9. Stepwise regression

Model	R^2	Adjusted R^2	Error in standard estimates	Variable of prediction
1	0.873	0.864	6.97084	Ln(disposable income)
2	0.968	0.963	3.63624	Ln(disposable income),pandemic
3	0.989	0.986	2.24005	Ln(disposable income), pandemic, exchange rate
4	0.992	0.990	1.90474	Ln(disposable income), pandemic, exchange rate,number of postgraduate students

Multiple linear regression was conducted with SPSS, and the results of coefficient test were not significant. Step-based regression was used to eliminate insignificant independent variables, and the results were shown in Table 9. The policy variable is excluded, indicating that the linear relationship between the variable and the dependent variable is not significant. The gradual regression is carried out in four steps. In the first step, the variable ln (per capita disposable income of residents) is introduced, with R2 greater than 0.85, indicating that per capita disposable income of residents is the key, and this variable largely explains the change of the number of Chinese students studying abroad.

After the policy variables were removed by stepwise regression, the multiple linear regression model was established with SPSS, and the model coefficient was significant. The multiple linear regression function is:

$$y = -390.914 + 39.994\ln(x_1) + 0.054x_2 + 0.043x_3 - 30.358D_1 \text{。}$$

The coefficient of this model shows that the change of the number of overseas students is positively correlated with the per capita disposable income of residents, the number of postgraduate students and the exchange rate. Compared with before the epidemic, the number of international students decreased by about 300,000 in the short term, but it did not affect the growth rate. According to the survey report "White Paper on Overseas Students under the Influence of COVID-19" released by QS, a global higher education analysis institution, 66% of Chinese students' study abroad plans have been affected [17]. However, only a small number of students gave up their plans completely, with 50 percent saying they would delay their plans or change their country of study, while less than 10 percent of the total said they would cancel their plans. Only 4 percent of Chinese students have decided not to study abroad. It can be seen that the epidemic has not shaken the majority of students' idea of studying abroad. We believe that the impact of the epidemic is only temporary, and the study abroad industry and study abroad will soon recover the original situation.

In Figure 3, the predicted value curve of the model and the real value curve of the number of Chinese students from 2005 to 2021 are drawn, and the two curves are very close to each other. The statistical indicators MAPE, MAE, MSE and RMSE of the model are all smaller than those of the support vector model (Table 10), which indicates that the fitting effect of the multivariate linear model is better than that of the support vector model. The predicted value of variables in the next two years in the GM(1,1) model above is put into the multiple linear regression model, and the number of Chinese studying abroad in 2022 and 2023 is 558,200 and 616,500, respectively. With the rapid development of Chinese economy, with the gradual increase of residents' per capita disposable income [18], the appreciation of RMB and the increasing number of postgraduate students, we have reason to believe that the number of overseas students will continue to increase in the future.

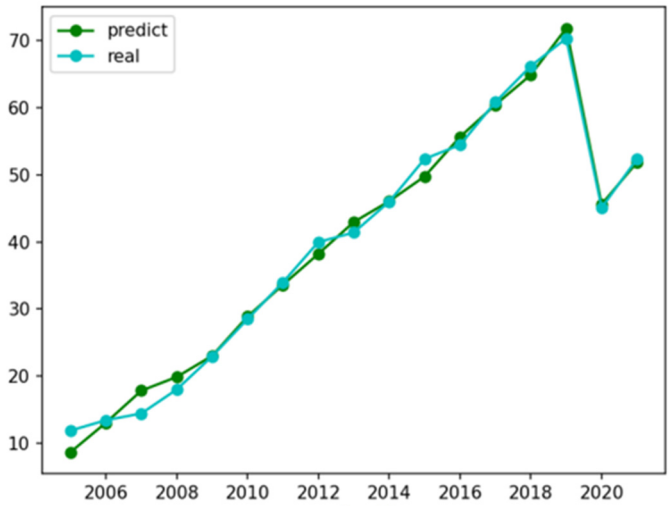

Figure 3. fitting and real value curve

Table 10. Statistical index

statistical index	
MAPE	5.27%
MAE	1.23
MSE	2.14
RMSE	1.46

3.7 Combined Model

In order to combine the advantages of multiple linear regression, ARIMA, and support vector regression models, different weights are assigned to each model, and then the weighted average of the predicted values of each model is used as the prediction result of the combined model [19], which is calculated by the following formula

$$Y_i^c = W_1 Y_i^1 + W_2 Y_i^2 + W_3 Y_i^3$$
$$st. W_1 + W_2 + W_3 = 1$$
$$W_1, W_2, W_3 .. 0, \quad i = 1, 2 \ldots$$

where is the prediction of the combined model, the prediction of the multiple linear regression, ARIMA, and support vector regression, respectively, W is the weight of the original model, and the function is the optimization objective of the combined model. Where Yi is the true value, this objective function takes the minimum error sum of squares as the basic criterion [20], and the weights are solved using Python to obtain the combined model expressions.

$$Y_i^c = 0.8276 Y_i^1 + 0.0098 Y_i^2 + 0.1626 Y_i^3$$

Table 11 and Table 12 show the error and statistical indicators of the combined model, and it can be seen that the multiple error indicators of the combined model are smaller than those of the multiple linear regression model, and it can be considered that the combined model works better. The combined model predicts that the number of Chinese students going abroad to study in 2022 and 2023 will be 529,700 and 578,200 respectively.

Table 11. Combined models predict values

Year	True value	Combined model	Relative error
2005	11.85	10.54	11.02%
2006	13.40	14.03	4.68%
2007	14.40	18.14	25.94%
2008	17.98	19.83	10.28%
2009	22.93	23.30	1.60%
2010	28.47	28.90	1.49%
2011	33.97	33.84	0.39%
2012	39.96	38.90	2.65%
2013	41.39	42.46	2.58%
2014	45.98	46.04	0.13%
2015	52.37	50.05	4.43%
2016	54.45	55.44	1.83%
2017	60.84	60.54	0.49%
2018	66.21	64.05	3.27%
2019	70.35	70.05	0.42%
2020	45.09	45.92	1.83%
2021	52.37	51.82	1.05%

Table 12. Statistical indexes of combined models

statistical indexes	
MAPE	4.36%
MAE	1.06
MSE	1.70
RMSE	1.30

4. Conclusion

4.1 Conclusion

In this paper, three single models and combined models are used for comparison, and the final conclusions obtained are: (1) In the single model, the multiple linear regression model with dummy variables has higher accuracy, and this model predicts the number of people going abroad to study in China in 2022 and 2023 is 558,200 and 616,500 respectively. (2) The combined model has better accuracy than the single model. The combined model gives the predicted values of 529,700 and 578,200 for the number of students going abroad to study in the next two years. (3) The main factors affecting the number of students going abroad to study in China are economic factors, willingness to study, policy guidelines, and the influence of epidemic. Among them, the economy is the decisive factor affecting the change of the number of students going abroad to study, and the economic development promotes Chinese students to go abroad to study. (4) The sudden outbreak of the epidemic has dealt a blow to study abroad in the short term, but in the long term, the impact of the epidemic is only temporary, and the change of the number of students studying abroad will resume the original growth pattern. (5) The change of willingness to study is positively correlated with the change of the number of students studying abroad, and the increase of willingness to study will promote the increase of the number of students studying abroad to a certain extent. (6) Compared with the three variables mentioned above, the policy of "supporting study abroad, encouraging return to China, freedom to come and go, and play a role" has the least influence on the number of study abroad and does not lead to any significant change in the number of study abroad, but this policy reflects the support and confidence of the country in studying abroad. (7) According to the forecast values of 2022 and 2023 calculated in this paper, more students will choose the path of studying abroad in the future.

4.2 Innovations and Shortcomings

Based on the previous research, this paper looks for the factors affecting study abroad and uses statistical knowledge to build a model to measure the number of overseas students in China over the years and predict the future trend on this basis. Secondly, most of the existing quantitative analyses of the number of study abroad in China appeared before the epidemic; statistical analyses of the number of study abroad after the epidemic are very rare, and this paper can supplement this aspect. The current analysis of the number of study abroad mostly uses a single modeling; this paper uses multiple models to try and combine them, thus discovering a combination model with better results.

There are still many shortcomings in this paper. First, this paper has very limited research on policy variables, focusing on only one policy, and does not have the ability to expand the scope of research on policy variables. Second, it has only been three years since the epidemic occurred, and the post-epidemic sample is too small compared to the pre-epidemic sample. The impact of the epidemic on the number of students studying abroad still needs long-term observation. There are still many aspects of the study on the number of students studying abroad in China that deserve optimization, and we hope that we can improve them in future studies.

References

[1] Wang Ping. The historical influence of the first foreign students on the development of China in modern China[J]. Journal of Anhui University of Technology (Social Science Edition), 2007(01):88-89.

[2] Cao Xiuying, Liu Li. How to open up a new road for the exchange of high technology talents [N]. Science and Technology Daily,2021,08(10):05.

[3] Li Huiyu. Blue Book of China Study Abroad Development Report (2020-2021) [J]. Study Abroad, 2021 (06): 15-16.

[4] Zhu Zhengzhong. The causes, problems and countermeasures of underage study in China in the context of globalization[J]. Journal of Adolescence,2016(04):26-29.

[5] Li Xuemei. Study on the Motivation and Influence of Underage Students Studying Abroad in Mainland China [D]. Xiamen University,2018.

[6] Liu Xuezhi. Development trends of Chinese students studying abroad and returning to China[J]. China Administration, 2016(01):52-57.

[7] Chen Yanglin, Yue Changjun. Trends and factors influencing students' study abroad in China's universities: 2005-2017[J]. Journal of Educational Scholarship,2020(05):40-45.

[8] Li Lanxiang. Econometric analysis of the factors influencing study abroad[J]. Journal of Economic Research,2013(24):274-276.

[9] Dwedar Dahlia. book Review: researching second language acquisition in the study abroad learning environment: an introduction for student researchers[J]. Research in Comparative and International Education, 2022,17(1):143-146.

[10] Xu, Z. H. A model for predicting the performance of asphalt pavement leveling by grey system theory[J]. Transportation Standardization,2013(09):112- 114.

[11] Liu, Lin. Spss data regression analysis and gray prediction [J]. Mathematical Learning and Research, 2021 (29): 136-137.

[12] He Xiaoqun, Liu Wenqing. Applied regression analysis (5th ed.) [M].2019:183-187.

[13] Fu, Changfeng. Research on online purchase behavior prediction based on machine learning fusion algorithm [J]. Journal of Beijing Printing Institute,2020,28(S1):262-264.

[14] Liu Jiang, Wang Zhengbo, Liu Cheng, Miao Changqi. Application of combined gray-timeseries model in building settlement prediction[J]. Beijing surveying and mapping,2018,32(02):214-217.

[15] Luo Jun. Exploration on the construction of support system for studying abroad in higher education[J]. Education Modernization,2020,7(55):149-152.

[16] Tao Mizhong, Liu Yifan. An empirical analysis of the factors influencing study abroad based on statistical data [J]. Journal of Higher Education, 2021(11):116- 120.

[17] QS survey: Has the number of Chinese students decreased due to the epidemic? [J]. China Foreign Trade, 2020 (09):72-73.

[18] Li Hui. Research on the integration mode of agricultural cold chain logistics based on the background of big data era [J]. Liaoning Economy,2019(08):18- 20.

[19] Feng Xiao, Huang Yuxin. China-Southeast Asia Total Import and Export Trade Forecast Based on Machine Learning Combined Model[J]. Business Economics,2022(03):103-105.

[20] You Yuxin, Chen Jihong. Research on real estate price forecasting method based on ARIMA-BP combined model [J]. Computer Knowledge and Technology,2020,16(09):264-269.