

GGDP Index Establishment Based on ARIMA-LSTM Prediction Model

Xiwen Shi^{1, *, #}, Jiaming Wu^{2, #}, Wenxi Zhang^{3, #}

¹School of Business, East China University of Science and Technology, Shanghai, China, 200237

²School of Mathematics, East China University of Science and Technology, Shanghai, China, 200237

³School of Information, East China University of Science and Technology, Shanghai, China, 200237

*Corresponding author: 21011821@mail.ecust.edu.cn

#These authors contributed equally.

Abstract. More and more studies have shown that GGDP may be more representative of the national economic health level and this paper establishes a GGDP calculation system based on the Environmental Economic Accounting System Central Framework. Then, with GGDP as the clustering index, the countries are divided into five categories. The Entropy TOPSIS model is established to quantify the impact of global climate. Then, based on the Ridge regression model, the functional relationship between GGDP and climate assessment values under different clusters is studied. Finally, two steps are taken to analyze the potential advantages and disadvantages of GGDP. In the case of using GDP, a prediction model based on ARIMA-LSTM is established to analyze global climate and economic development trends. In the case of using GGDP, to quantify the change of GGDP, international factor α and domestic factor β are introduced to characterize the impact of global efforts and domestic efforts on GGDP.

Keywords: SEEA Central Framework, Systematic clustering, Entropy TOPSIS model, Ridge Regression, ARIMA-LSTM model.

1. Introduction

To measure the economic condition and development of a country or region, the most famous and widely used indicator is gross domestic product (GDP). However, GDP as a measurement index has some shortcomings, it does not include environmental quality. Suppose governments abolished all environmental regulations so that businesses could produce goods and services regardless of the pollution they caused. In this case, GDP would increase, but welfare would most likely decline, with the deterioration of air and water quality outweighing the welfare benefits of more production [1]. To be specific, GGDP is the gross product after deducting the cost of economic losses caused by environmental pollution, degradation of natural resources, poor education, uncontrolled population, mismanagement, and other factors. This index essentially represents the net positive effect of national economic growth. Wang Xue and Shi Xiaoqing applied the index system method, and they established the indexes of water resource efficiency and environmental efficiency of wastewater and waste.[2] Hu Biao and Fu Yeteng combined the non-expected output SBM model with spatial autocorrelation analysis to measure the ecological efficiency of China.[3] Shen Lulu and Liang Jiale used ARIMA model to predict solar radiation data and extract linear components in the data. Then, the filtered residual is substituted into the LSTM neural network model to obtain the prediction of the nonlinear component [4].

2. Methods

2.1. Data collection and processing

The data we used mainly include GDP, living plane, forest cover, renewables (% growth), greenhouse gas emission, PM10, PM2.5, and fossil fuels (% growth). The data come from Our World in Data Official Website, the United Nations Department of Economic and Social Affairs Official Website, USA Gov Official Website, etc. The exact data resources are shown in table 1. To mine the relevant knowledge more effectively, it must be provided with clean, accurate, and concise data. But data are generally incomplete, inconsistent, and dirty data, which cannot be directly mined, or the results are unsatisfactory[5]. Moreover, the dimension and unit of the data might be inconsistent. So, we must do the data cleaning, and remove noise from the data, and correct inconsistencies. Data reduction is also necessary, including compressing data by aggregating, removing redundant attributes, or clustering[6]. In addition, conversion is needed to unify the dimension of different data. When it regards the missing value, we apply the data interpolation method such as the Lagrange and Newton Interpolation. The data is shown in table 1:

Table 1. Data Source

	Our World	UNEAS	Gov Data	US Gov
Data	Environmental data	Worldwide data	GDP	Natural resources
Web site	https://www.cambridge.org/core/journals/environmental-data-science	https://ourworldindata.org/	https://data.stats.gov.cn/	https://www.britannica.com/science/natural-resource

2.2. GGDP calculation framework

The 2012 Center Framework for Accounting Systems in the Environmental Economy (SEEA Central Framework) uses a wide range of information and is structured to compare source data[7]. The SEEA Central framework brings together information on water, minerals, energy, wood, fish, soil, land and ecosystems, pollution and waste, production, consumption, and accumulation in a single measurement system, assigning specific and detailed measurements to each area so that, The SEEA Central Framework allows environmental information and economic information to be integrated into a single framework to provide a comprehensive view.

SEEA expands the scope of natural assets accounting in SNA and evaluates the value of natural assets and their changes. The SEEA accounting framework can be seen from the basic structure of SEEA. It is an accounting system based on the total GDP, which subtracts the depletion of natural resources, environmental pollution losses, and environmental governance cost and then adds environment benefits:

$$EDP = GDP - \sum X_1 - \sum X_2 - \sum X_3 + \sum X_4 \quad (1)$$

$$GDP = C + I + G + NX \quad (2)$$

C is consumption; I is investment; G is government consumption ; NX is net export.

The specific index chosen to calculate GGDP is listed in the following table 2:

Table 2. Index Chosen in Calculation GGDP

Subaccount		First-order Index	Second-order Index
Gross GDP		Regional Gross GDP	Regional Gross GDP
Depletion of Nature Resources		Energy Consumption	Energy Consumption
		Cultivated Land	Cultivated Land
		Water Consumption	Water Consumption
Environmental	Pollution Losses	Environmental Degradation	Ecology Benefit Loss Biodiversity Loss
Environmental	Governance Cost	Pollution Governance Cost	Actual Governance Cost Virtual Governance Cost
Environmental	Improvement Gains	National Happiness	National Happiness Index
			Carbon Fixation Effect
		Forest Ecological Benefit	Urban Dust Retention Effect
			Climate Regulation Effect

3. Model establishment

3.1. Cluster analysis model

The clustering model is the process of dividing a sample into multiple classes composed of similar objects. In the last model, we established an evaluation system of GGDP based on the SEEA Central Framework and calculated the GGDP values of all countries in 2020 with the data collected from each country[8]. Due to the large sample size of data, direct regression analysis is not obvious, so system-level clustering is first adopted. Data after dimensionality reduction will belong to multiple categories, the results are as follows in table 3:

Table 3. Clustering categories

Classification	Specific Description
Category 1	Low population density in developed countries
Category 2	Large national size developing countries
Category 3	Super countries
Category 4	Large national size developed countries
Category 5	Countries depending on natural resources

➤ **Category 1: Low population density developed countries**

The first category has low population density, so the absolute value of total GDP is not large. However, these countries are highly developed, so the GDP per capita is very high. Typical representatives include Denmark, Greece, Norway, Singapore, and Spain. They can strike a better balance between economic development and environmental protection.

➤ **Category 2: Large national size developing countries**

The second category includes Armenia, Belarus, Benin, Djibouti, Kenya, Nepal, Solomon, and Vanuatu. The absolute GDP of these countries is small, although the land area is large. Since these countries in Asia, Africa, and Latin America are relatively underdeveloped, they have relatively good environmental protection because they have not been economically developed.

➤ **Category 3: Super countries**

The third category is super countries, including the United States, China, Russia, and Brazil, which have large land area and relatively high degree of development. Their exploitation and utilization of the environment have been nearly saturated, and the adoption of GGDP will have a certain impact on the improvement of the environment and climate.

➤ **Category 4: Large national size developed countries**

The fourth category are Australia, Canada, France, Germany, India, Japan, and the United Kingdom. They are mainly old, developed countries with high gross GDP and per capita GDP. These countries are usually able to balance economic development and environmental protection.

➤ **Category 5: Countries depending on natural resources.**

Finally, the fifth category includes Bahrain, Brunei, Ecuador, Equatorial Guinea, Oman, Peru, Saudi Arabia, and Trinidad and Tobago. These countries' GDP growth is mainly achieved through oil extraction, cutting down trees, and so on at the cost of environmental damage. Therefore, it performs poorly under the evaluation system of green GDP[9].

3.2. Entropy TOPSIS model

TOPSIS, a commonly used comprehensive assessment method, is applied to quantify the impacts of climate change. It can make full use of raw data such as depletion of natural resources, environmental pollution losses, environmental governance cost, and environmental improvement gains. The quantitative results accurately reflect the expected global impacts of such a transition on climate mitigation.

The entropy weight method is also used to correct the subjectivity of weight, based on the principle that the smaller the degree of variation of the index, the less information reflected, and the lower the corresponding weight value[10].

For objects including living plane, forest cover, renewables (% growth), greenhouse gas emission hundred million tons), PM10, PM2.5, and Fossil fuels (% growth), to be evaluated and 4 evaluation indexes, a forward matrix is constructed:

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad (3)$$

Define standardized matrix as Z, with the formula $Z = X_{ij} / \sqrt{\sum_{i=1}^n x_{ij}^2}$, we can transform each element in Z with each element in X, finally we have the standardized matrix as follows:

$$Z = \begin{bmatrix} Z_{11} & \cdots & Z_{1m} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \cdots & Z_{nm} \end{bmatrix} \quad (4)$$

To carry out weight assignment objectively, the entropy weight method is used to carry out weight assignment objectively. The specific principle is that the smaller the degree of variation of indicators, the less information reflected, and the lower the corresponding weight should be.

So, the information entropy is

$$H(x) = \sum_{i=1}^n [p(x_i)I(x_i)] = - \sum_{i=1}^n [p(x_i) \ln(p(x_i))] \quad (5)$$

It can be used to calculate the j th indicators' information entropy, e_j

Next, the information utility value of each indicator is calculated using the formula:

$$d_j = 1 - e_j \quad (6)$$

The larger the information utility value is, the more information will correspond to. Then, it is necessary to normalize the information utility value to obtain the entropy weight of each indicator. The specific formula is as follows:

$$W_j = \frac{d_j}{\sum_{j=1}^m d_j (j = 1, 2, \dots, m)} \quad (7)$$

The results are as follows in table 4:

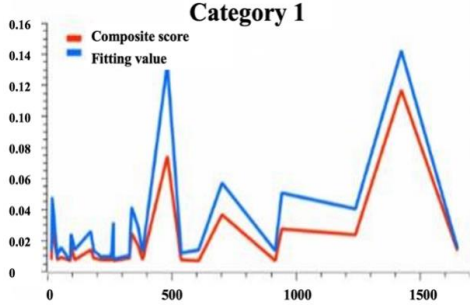
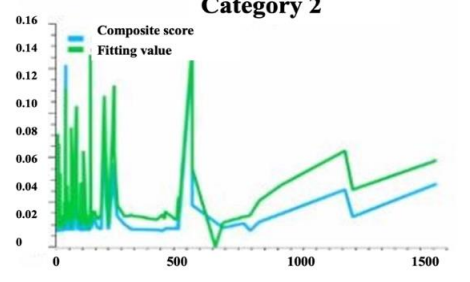
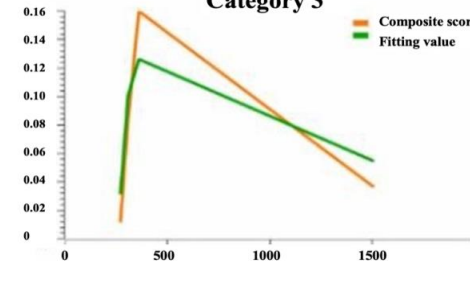
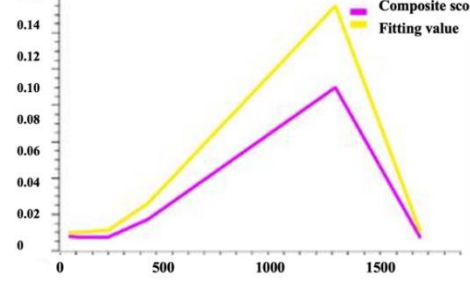
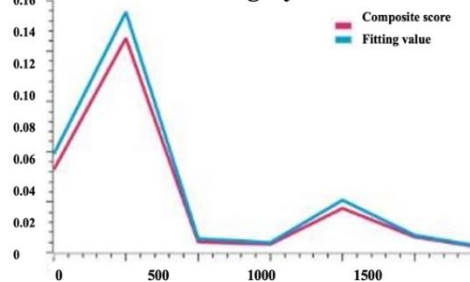
Table 4. Regression Score

Country	Category	Composite Score	GGDP
Austria	1	0.008733	330.0286
Belgium		0.007336	382.1691
Denmark		0.007368	251.1586
Armenia	2	0.007362	10.1960
Belarus		0.012872	58.2560
U.S.	3	0.380523	60336.1765
China		0.258042	10832.7790
France	4	0.163447	2397.8259
Germany		0.015579	2656.87462
Bahrain	5	0.007322	26.0401
Brunei		0.007366	10.3821
Equatorial		0.007963	12.3519

3.3. Ridge regression model

The task of regression analysis is to try to explain the formation mechanism of Y by studying the correlation between independent variable X and dependent variable Y, to achieve the purpose of predicting Y through X. Ridge regression analysis is adopted in this paper, which is more suitable for small sample data and can effectively reduce data overfitting compared with linear regression. The prediction of unknown data is more robust, and the regression coefficient obtained by regression is more realistic. In addition, ridge regression can reduce the fluctuation range of estimated parameters and make them more stable, which is of great practical value in research with more ill-conditioned data. The results are as follows in table 5:

Table 5. Regression Results

The figure of each category	Description
Category 1  Positively related	Due to their geographical advantages and national size, these countries do not need to destroy the environment for development and have the spare power to pay attention to climate and environmental protection. Therefore, the increase of GGDP is accompanied by the improvement of climate and environment.
Category 2  Positively related	They score relatively low because of low levels of development, poor natural conditions and other factors. Part of the anomaly is that some tourism-driven countries, despite their low levels of development, are extremely climate-conscious.
Category 3  Positively related	Due to the first-mover advantage, countries with high GGDP have the ability to carry out environmental protection and pollution control at the same time of economic development, so they have higher climate evaluation value.
Category 4  Positively related	In developed countries, economic growth is powered by high-end manufacturing and service industries, and there is no need to over-exploit the environment.
Category 5  Negatively related	The industrial structure of oil-producing countries is deformed, and the domestic economy mainly comes from the export of resources.

3.4. Prediction model

3.4.1 ARIMA model

ARIMA model is a combination of autoregressive moving average (ARMA) model and difference model. The ARIMA (p, d, q) model is mainly to solve the prediction problem of non-stationary time series. The d-order difference operation is carried out on the non-stationary time series to get stationary series, and then the ARMA (p, q) model is used to predict the stationary series.

The formula for ARMA is:

$$X_t = c + U_t + \sum_{i=1}^p \alpha_i X_{t-i} - \sum_{i=1}^q \beta_i \varepsilon_{t-i} \quad (8)$$

α_i is autoregressive coefficient, β_i is sliding average coefficient, residual parameter is c

We need to perform ACF and PACF tests on the sequence. Suppose there are two random variables X and Y, and the time series is stationary

$$\begin{cases} E(x_t) = E(x_{t-s}) = u \\ Var(x_t) = Var(x_{t-s}) = \sigma^2 \\ Cov(x_t, x_{t-s}) = \gamma \end{cases} \quad (9)$$

Define the autocorrelation coefficient:

$$\rho_s = \frac{cov(x_t, x_{t-s})}{\sqrt{Var(x_t)}\sqrt{Var(x_{t-s})}} \quad (10)$$

Suppose $\{x_t\}$ is still stable, the ACF reflects the linear relationship between x_t and x_{t-s} , the PAFC reflects the relationship between x_t and x_{t-s} without the linear property of the middle values.

$$x_{t-s} \rightarrow x_{t-s+1} \rightarrow x_{t-s+2} \rightarrow \cdots \rightarrow x_{t-1} \rightarrow x_t \quad (11)$$

3.4.2 LSTM model

The difference between LSTM deep learning algorithm and a recursive neural network lies in that the former adds two structures of cell state and gate to predict based on the latter, which is used to solve the problem that the weight of RNN disappears in the backward propagation over time. LSTM introduces three gated units, including Forget Gate, Input Gate, Output Gate, and a memory cell for storing information. The function of the three gated units is realized by the sigmoid function and the dot product operation. The gated unit provides no additional information and is only responsible for filtering the amount of information passed.

x_t is the network output at t, h_{t-1} is the network output at t-1, c_t is the condition of cell at t, f_t is output of forget gate at t, O_t is t output of output of output gate at t.

The calculation formula is as follows:

$$f_t = \sigma \left(W_f^T \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + b_f \right) \quad (12)$$

f_t is the weight of forget gate, T is transpose, b_f is the offset of forget gate, σ is the activating function of sigmoid, the expression is:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (13)$$

The expression for input gate is:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_c) \quad (14)$$

$$c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (15)$$

W_i is the weight of input gate, T is transpose, b_i is the offset of input gate, W_c is the weight of tanh function, b_c is the offset of tanh function

Expression of $\tanh(x)$ is:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (16)$$

the expression of unit condition of c_t at t is:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (17)$$

LSTM use the output gate to output h_t and $\tanh$ function:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (18)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (19)$$

W_o is the weight of the output gate, b_o is the offset of output gate

LSTM time series prediction model first divides the time series data into two verification sets: training set and test set, and then preprocesses the data. After setting the initial parameters and loss function of the model, the prediction model is trained, and the training model is predicted.

3.4.3 ARIMA-LSTM model

To achieve the best prediction effect, a new nonlinear ARIMA-LSTM model is constructed by combining the advantages of the neural network and the traditional time series prediction model.

First, the raw data is decomposed into linear and nonlinear parts and treated separately for given time series Z_t can be divided into linear part L_t and nonlinear part N_t .

$$Z_t = L_t - \tilde{L}_t \quad (20)$$

The linear part L_t is fitted by ARIMA, and the nonlinear part N_t is fitted by LSTM, and we can obtain the prediction $\tilde{L}_t$ and residual e_t

$$e_t = Z_t - \tilde{L}_t \quad (21)$$

Then the residual sequence is fitted by LSTM model to get the prediction $\tilde{N}_t$

$$\tilde{N}_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t \quad (22)$$

Take the linear sum and get the prediction $\tilde{Z}_t$

$$\tilde{Z}_t = \tilde{L}_t + \tilde{N}_t \quad (23)$$

Therefore, this paper constructs a nonlinear composite ARIMA-LSTM model with the help of neural network tools. After the prediction results of linear and nonlinear parts are obtained, the LSTM model is used again, and the predicted values obtained in the above steps are used as the input to find the complex relationship between them through the deep neural network.

Final output prediction result:

$$\tilde{Z}_t = g(\tilde{L}_t, \tilde{N}_t) \quad (24)$$

To objectively evaluate and judge the model, the average absolute percentage error (MAPE) index was selected in this paper to evaluate the prediction results. The formula is as follows:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{f_t - d_t}{d_t} \right| \times 100\% \quad (25)$$

We put GDP, greenhouse gas emissions and other index data into ARIMA and ARIMA-LSTM respectively for prediction and calculated the MAPE value. The ARIMA-LSTM is 0.0013%, and the ARIMA is 0.0145%. The MAPE value of the former one is much smaller than the fitting result of the single ARIMA model, so the prediction is very accurate and reliable. Due to space limitation, only the forecast results of the ARIMA-LSTM model of US GDP are shown here. After GGDP replaces GDP, the old and new GGDP of each cluster is obtained.

We select Norway for display, as shown in Figure 1 below:

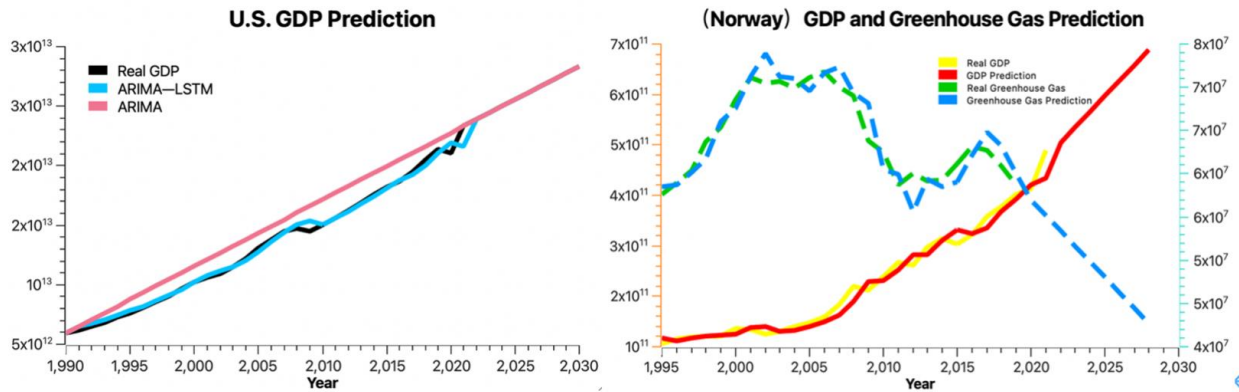


Figure 1. GDP and GGDP prediction results

When global GDP is replaced by GGDP, it is predictable that global economic development and climate change will result. Specifically, for any given country, its development will be influenced by both foreign and domestic influences. We quantified these two influencing factors as α and β . [10]

α indicates that because GGDP replaces GDP, each country adjusts its development route, which will influence the impact of GGDP in other countries. For example, when country B transfers its waste treatment industry abroad, it has an impact on the GGDP of country A, or when global consensus is issued, and conventions are signed. We define these external influencing factors as α .

β represents the change of GGDP due to the replacement of GDP by GGDP. For example, country A vigorously develops renewable energy, promulgating laws to prohibit the indiscriminate exploitation of resources, and so on. We define this kind of internal influencing factor as β .

Based on the transfer trend of high-pollution industries in the past, the trend of energy development, and the impact of the signing of the Kyoto Protocol on various countries and regions, the definition of α and β is summarized in the following table 6:

Table 6. Value of α & β

	Value	Description
Category 1	$\alpha=1\%$ $\beta=5\%$	After the adoption of GGDP, we can improve and implement relevant environmental protection laws, and the health level of the domestic economy can be significantly improved.
Category 2	$\alpha=-5\%$ $\beta=-2\%$	For domestic economic development, most of these countries need to rely on natural resource exploitation, which will reduce the GGDP.
Category 3	$\alpha=3\%$ $\beta=5\%$	They have a strong international influence and can sign international environmental protection conventions; For GGDP, they can adopt a sustainable development strategy, to make the economy healthy.
Category 4	$\alpha=2\%$ $\beta=3\%$	The impact is roughly the same as that of superpowers, but the impact value is slightly smaller due to the smaller size of the former country.
Category 5	$\alpha=-6\%$ $\beta=-10\%$	Excessive energy exploitation and environmental damage are very serious, so GGDP drops sharply.

For the above control factors, sensitivity analysis of (α, β) was conducted as (0%, 0%) (5%, 0%) (0%, 3%) (5%, 3%). Due to space limitations, we only present the sensitivity analysis results of the US as follows in figure 2:

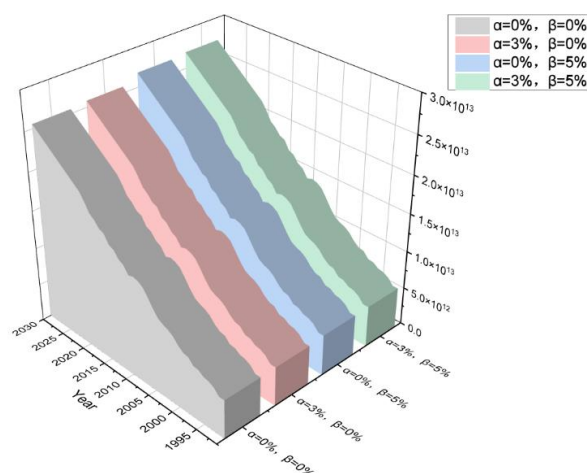


Figure 2. Sensitivity analysis of U.S. data

Adding α has a significant positive effect on GGDP, adding β has a less significant positive effect on GGDP, removing α has a significant negative effect on GGDP, removing β has a less significant negative effect on GGDP, which is similar to the significant reduction of greenhouse gas emissions in the index. The non-renewable energy consumption is not significantly reduced, which is also consistent with the actual climate environment is significantly improved, and the consumption of natural resources is not significantly reduced.

4. Calculation results

After getting the new GGDP of all countries in the world, we can get the changed GDP and climate change assessment value based on the regression model established above and compare the two values with the unchanged situation. The result of U.S. are shown as an example as follows in figure 3:

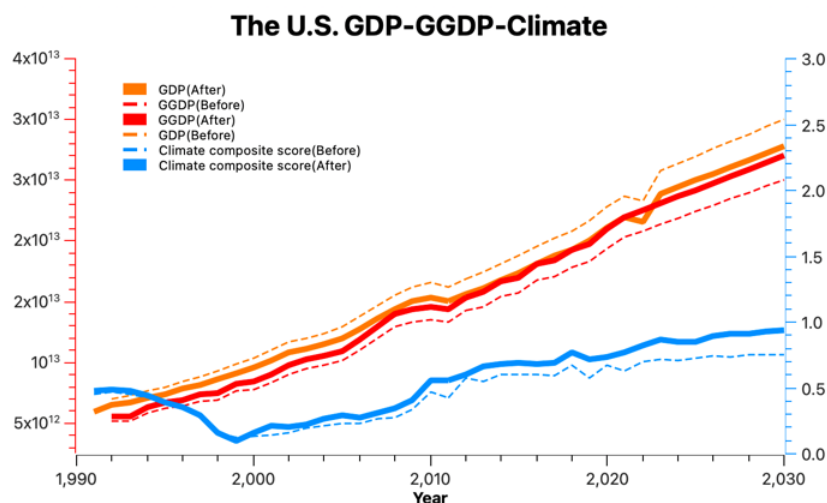


Figure 3. GGDP calculation result of the U.S.

5. Conclusion

In this paper, the United States is selected for further analysis of the possible impact of this transition. It is expected that the United States will take measures to reduce the consumption and development of natural resources (such as fossil fuels and other non-renewable resources). To prove this point, we carried out gradual regression for each indicator in the GGDP evaluation system and finally obtained the index of population density, that is, population density is the most important indicator in the GGDP evaluation system.

In GGDP evaluation, most of the indicators, such as fossil fuel consumption, forest reserve, water resource consumption, and cultivated land area, are highly correlated with human production activities. The increase in population density is often accompanied by the consumption of natural resources, environmental pollution, and the impact of climate change. This is consistent with the reality.

Here, correlation analysis is conducted between population density and GDP and GGDP respectively. Through Spearman correlation analysis, the correlation between GGDP and population density is 0.999, and that between GDP and population density is 0.998. The correlation between GGDP and population density is more significant, which is consistent with our conclusion.

References

- [1] Gavard C, Kirat D. Flexibility in the market for international carbon credits and price dynamics difference with European allowances [J]. *Energy Economics*, 2018(76): 504-518.
- [2] Wang Xue, Shi Xiaoqing et al. EUA and CER phase II price drivers: Unveiling the reasons for the existence of the EUA-CER spread [J]. *Energy Policy*, 2011, 39(3): 1056-1069.
- [3] Hu Biao, Fu Yeteng et al. Causes of the EU ETS price drop, Recession, CDM, renewable policies or a bit of everything? —New evidence [J]. *Energy Policy*, 2014(73): 676-685.
- [4] Shen Lulu, Liang Jiale, Zhou Wen. Energy prediction algorithm based on ARIMA-LSTM [J]. *Radio Communication Technology*, 2023, 49(01): 150-156.
- [5] Xing Qi, Zhu Daolin, Wang Jian. Based on the natural resources of SEEA assets, international comparison of accounting index [J]. *Journal of environmental management in China*, 2022, 14 (01): 68-77.
- [6] Chen Yongzhi, Lin Yingxing, Cheng Hang. Does public evaluation really reflect the level of regional green development? Analysis based on α level set fuzzy TOPSIS method [J]. *Journal of Fuzhou University (Philosophy and Social Sciences Edition)*, 201, 35(03): 46-52.
- [7] Niu Xuena, Shi Zhanhong, Zhou Shengwei et al. Grassland ecological security assessment in eastern Qilian Mountains of Gansu Province based on TOPSIS method of combinatorial weighting: A case study of Tianzhu Tibetan Autonomous County, Gansu Province [J]. *Gansu Science and Technology*, 20, 36(22): 23-27.
- [8] Yang Kail. Datong green GDP calculation based on SEEA-2012 [C]//. China environmental science society proceedings of science and technology annual meeting (a). In 2022, 2022: 114-119.
- [9] LIU Xiang. Research on Prediction and Early Warning Technology of Pipeline Stress Based on ARIMA-LSTM [J]. *Petroleum Engineering Construction*, 202, 48(05): 38-43.
- [10] Song Geifeng, Liu Yanzhong, Zhu Xiaonan, Chen Yong, Zhang Zuo. Dynamic change of ecological efficiency in Wuhan based on emergy analysis [J]. *Ecological Economy*, 2019, 35(07): 103-109.