Research on GGDP Evaluation System Based MRA-GPR Analysis

Yueyao Li[1], Ziyu Liu[2], Ziyao Zhu[1,*], Zixuan Zhou[3], Yan Guo[4]

[1]School of Insurance and Economics, University of International Business and Economics, Beijing, 100029, China
[2]Business School, University of International Business and Economics, Beijing, 100029, China
[3]School of Cultural Industries Management, Communication University of China, Beijing, 100024, China
[4]School of Banking and Finance, University of International Business and Economics, Beijing, 100029, China
[*]Corresponding author: zzyafr@163.com

Abstract: *This paper focuses on demonstrating that GGDP is a better indicator of a country's economic health than GDP. In this study, seven variables were selected as secondary indicators, and then CO_2 emissions were used as dependent variables, and the identified seven variables were used as independent variables for multiple regression. In this regression, significant and robust results are obtained, which can prove that using GGDP as a macroeconomic variable is more environmentally friendly. This paper also uses the cross validation method to train the Gaussian process regression model, and obtains better regression results (R^2= 99.9 %, RMSE = 1.364e + 5).*

Keywords: *Green GDP, CO_2, MRA-GPR*

1. Introduction

GDP has long been used to measure the level of economic development of a country and the quality of life of its people. However, since the 1960s, many scholars have pointed out that GDP is not a good indicator of a country's level of development because it ignores concerns about the environment and sustainable development [1]. So some scholars have been tried to create a new indicator, green GDP (GGDP) which takes environmental cost into consideration and tried to prove that GGDP is a better indicator of a country's economic health [2-5].

This problem firstly described the disadvantage of the widely used indicator GDP to introduce an updated concept: GGDP. And it mainly asks competitors to find a best measurement of GGDP and try to give proof that it's actually an incredible indicator of a country's economic health and the sustainable level [6-8].

This study intends to construct a defensible model to estimate the impact of GGDP on global climate mitigation. In this question, we need to build a model to estimate the GGDP's influence of global climate mitigation. So we used a multiple regression model in which CO2 emission is the dependent variable and the GGDP is the independent variable. With this regression result, we can explain the GGDP's influence on environment.

2. Data Processing

As a preparation for the model building, we downloaded the relevant data from the World Bank, took the intersection of the years covered by each indicator, determined the survey year interval from 2000 to 2019, and imported the data into an excel sheet for data processing. The data are first cleaned to remove invalid values and outliers and replace missing values with sample means; then the data are converted into a format that can be understood by the machine learning model and the individual countries are labeled and coded. Further, we perform visual and statistical analysis of the data to gain insight into the characteristics and distribution of the data. For data with skewed distribution, we use taking logarithm to make it easier to be understood by machine learning. Finally, data normalization is performed in order to improve the accuracy and performance of the model [9,10].

3. MRA-GPR Analysis

In contrast to the self-regulation of the ecosphere, the mitigation of climate problems undoubtedly requires the world's countries to take the initiative to make changes. Using CO2 emissions to measure changes in global climate, we hope to show what changes countries should make if they want to mitigate the climate problem by establishing robust regressions of GGDP and the remaining seven environmental indicators on CO_2 emissions.

First we need to show that GGDP has a significant effect on CO2 emissions, and the multivariate regression analysis in econometrics has a more robust theoretical framework and stronger persuasive power in addressing this. We then build a machine learning model for time series regression prediction, since the prediction results of multiple regression are often unsatisfactory. Thus, we repeated the regression of the above variables using Gaussian Process Regression.

3.1 Variable Screening

To set our own GGDP's measurement, we selected seven variables as secondary indicators. The details are shown in the table 1.

Table 1: Accounting method of green GDP

Account type	Account classification	Notation
Natural resource loss	Cultivated Land Area	CLA
	Annual Freshwater Extraction Total	AFE
	Depletion Value of Mineral Resources	DMR
	Depletion Value of Energy Resources	DER
Loss of environmental degradation	Nitric Oxide Emissions	NOE
	Total Renewable Inland Freshwater Resources	IFR
Source and environment improvement benefits	Domestic General Health Expenditure	GHE

3.2 Multiple Regression Analysis

For the explanatory variable CO2 emissions, we will use it as the core explanatory variable, introduce DMR, DER, AFE, CLA, NOE, IFR, GME as control variables and keep the intercept term, so the preliminary model framework is designed as follows.

$$CO2E = \beta_0 + \beta_1 GGDP + \beta_2 DMR + \beta_3 DER + \beta_4 AFE + \beta_5 CLA + \beta_6 NOE + \beta_7 IFR + \beta_8 GME$$

To avoid the problem of multicollinearity, we output the variable covariance matrix and perform the VIF factor independence test. The maximum value of VIF is: $VIF_{max} = VIF_{IFR} = 17.48$. There is a multicollinearity problem, and to ensure the integrity of the control variables, we do not adjust it here for the time being (Table 2).

Table 2: Covariance Matrix

e(V)	GGDP	GME	IFR	NOE	CLA	AFE	DER	DMR
GGDP	6.65E-15							
GME	-7.43E-14	8.86E-13						
IFR	1.01E-15	-1.75E-14	2.62E-15					
NOE	-6.25E-14	1.17E-12	-7.85E-14	4.93E-12				
CLA	2.27E-15	-5.50E-14	1.58E-15	-2.07E-13	3.29E-14			
AFE	-7.94E-15	-1.92E-13	3.87E-14	-2.70E-12	5.53E-14	2.04E-12		
DER	-1.50E-14	4.07E-13	-4.31E-14	2.93E-12	-8.65E-14	-2.54E-12	1.12E-11	
DMR	5.23E-14	-1.05E-12	1.14E-13	-5.45E-12	2.97E-13	3.72E-12	-1.75E-11	3.00E-11
_cons	0.001108	-0.028739	0.001172	-0.137671	0.0108598	0.072617	-0.106656	0.179436

Considering the autocorrelation problem of the model itself, we run the Dubin-Waston test, and the original hypothesis of no autocorrelation is rejected by the test results. Autocorrelation may lead to heteroskedasticity problems, so we use heteroskedasticity robust standard errors for subsequent regression adjustment (Figure 1).

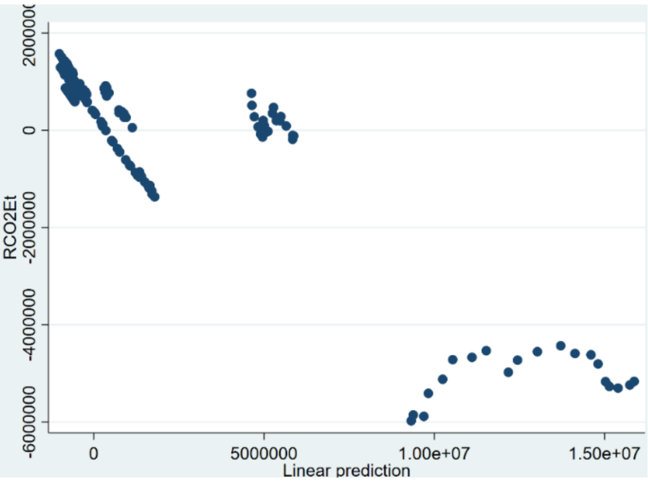

Figure 1: Regression Error Distribution

Considering the autocorrelation problem of the model itself, we run the Dubin-Waston test, and the original hypothesis of no autocorrelation is rejected by the test results. Autocorrelation may lead to heteroskedasticity problems, so we use heteroskedasticity robust standard errors for subsequent regression adjustment.

In order to prevent omitted variable bias from appearing and causing variable significance to be affected, we set up a fixed-effects versus random-effects control for the analysis. Fixed effects take into account the omitted endogenous variables associated with individuals in the panel data ($cov(x_1, c_i) \neq 0$) and absorb them into the intercept term, which is the method of within-group estimation. Fixed effects control for all variables that do not vary over time, getting rid of the spurious relationship problem, only variables that vary over time (x_{it}), and variables that do not vary over time (x_1) will be removed from the model.

Table 3: Multiple Regression Results

	Ordinary Multiple Regression	Heteroscedasticity Robustness	Random Effect	Fixed Effect
GGDP	6.04e-07*** (21.64)	6.04e-07*** (14.95)	6.04e-07*** (21.64)	4.03e-07*** (10.74)
DMR	9.26e-06** (3.20)	9.26e-06** (2.90)	9.26e-06** (3.20)	6.41e-06* (2.57)
DER	3.60e-06* (2.36)	3.60e-06 (1.94)	3.60e-06* (2.36)	3.31e-06* (2.41)
AFE	7.23e-06*** (13.01)	7.23e-06*** (9.18)	7.23e-06*** (13.01)	1.03e-05** (3.01)
CLA	-8.77e-08 (-0.56)	-8.77e-08 (-0.57)	-8.77e-08 (-0.56)	-2.57e-06** (-2.99)
NOE	1.20e-06 (1.66)	1.20e-06 (1.60)	1.20e-06 (1.66)	1.33e-05*** (7.00)
IFR	-1.43e-07*** (-7.10)	-1.43e-07*** (-6.68)	-1.43e-07*** (-7.10)	
GME	-6.01e-06*** (-17.58)	-6.01e-06*** (-15.25)	-6.01e-06*** (-17.58)	-4.25e-06*** (-10.64)
_cons	-20158.8 (-0.61)	-20158.8 (-0.65)	-20158.8 (-0.61)	-591816.7 (-1.27)
N	200	200	200	200
p	2.98e-175	4.42e-175	0	1.19e-97
R2	0.987	0.987		0.922
F	1794.1	1786.6		307.9
RMSE	304927.1	304927.1	304927.1	232613.8
Note: t statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001				

From the regression results, the results of fixed effects are significantly better than random effects, and the significant positive effect of GGDP on CO2 emissions is revealed after absorbing the effect of individual characteristics. And the multicollinearity problem was solved after removing the IFR variables($\overline{VIF}$ <10).

The multiple regression analysis of individual fixed effects demonstrates a significant effect of GGDP on CO2 in terms of time variation and that this effect is country dependent (Table 3).

3.3 Gaussian process regression

The main machine learning models used in regression are linear regression, distributed linear regression, decision tree models, support vector machines, integration algorithms, Gaussian process regression, etc. Here we regress CO_2 emissions one by one using the mainstream algorithm substituting GGDP with the seven control variables mentioned above. The regression results are shown as table 4, where the Gaussian process regression is the best fit (here, the lowest RMSE level).

Table 4: Machine Learning Parameter Comparison

Model	RMSE	R2	MSE	MAE	Note
Gaussoan Process Regression	1.3604e+05	0.99	1.8506e+10	66807	Exponenyial
Linear Regression	3.0978e+05	0.99	9.5965e+10	143760	Interactions
Stepwise Linear Regression	3.4138e+05	0.98	1.1654e+11	245490	
Tree	3.2790e+05	0.98	1.0752e+11	121770	Fine
SVM	2.4583e+05	0.99	6.0433e+10	109630	Quadratic
Ensemble	2.6851e+05	0.99	7.2095e+10	123460	Boosted

GPR is a machine learning regression method developed in recent years, which has a strict statistical theoretical basis and good adaptability to deal with complex problems of high dimensionality, small samples, and nonlinearity, and has strong generalization ability, and has the advantages of easy implementation, adaptive acquisition of hyperparameters, flexible nonparametric inference, and probabilistic significance of output compared with neural networks and support vector machines(Park C W et al. ,2011)The Gaussian process regression principle consists of two main parts: prediction and training.

The function value prediction by Gaussian process regression starts from the function space perspective by defining a Gaussian process describing the function distribution and performing Bayesian inference directly in the function space. GP is the set of any finite number of random variables having a joint Gaussian distribution, whose properties are completely determined by the mean and covariance functions, i.e.

$$\begin{cases} m(\boldsymbol{x}) = \boldsymbol{E}[f(\boldsymbol{x})] \\ k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{E}\big[(f(\boldsymbol{x}) - m(\boldsymbol{x}))\big(f(\boldsymbol{x}') - m(\boldsymbol{x}')\big)\big] \end{cases} \tag{1}$$

where $\boldsymbol{x}, \boldsymbol{x}' \in R^d$ is an arbitrary random variable. Thus GP can be defined as $f(\boldsymbol{x}) \sim \mathrm{GP}\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big)$. We standardize the data when we introduce them so that the data mean function is zero, simplifying the operation.

In the regression process we consider the model:

$$y = f(\boldsymbol{x}) + \varepsilon \tag{2}$$

where x is the input vector, f is the function value, and y is the observed value affected by noise, further assuming that the noise $\varepsilon \sim N(0, \sigma_n^2)$.The prior distribution of the observation $\boldsymbol{y}$ can be obtained as:

$$\boldsymbol{y} \sim N(0, K(X, X) + \sigma_n^2 I_n) \tag{3}$$

and the joint prior distribution of the observed $\boldsymbol{y}$ and predicted values f_* is

$$k(\boldsymbol{x}, \boldsymbol{y}) = exp\left(-\frac{\|x-y\|}{2\sigma_f^2}\right) \tag{4}$$

The hyperparameter θ is obtained by the great likelihood method. We firstly established the negative log-likelihood function of the conditional probability of the training sample. $L(\boldsymbol{\theta}) = -\log p(\boldsymbol{y} \mid X, \boldsymbol{\theta})$, and let its partial derivative with respect to the hyperparameter θ. The conjugate gradient method optimization method was used. After obtaining the optimal hyperparameters, the predicted values f_* corresponding to the test points and their variances $\hat{\sigma}_{f_*}^2$ are obtained using $\bar{f}_*$ and $\mathrm{cov}(f_*)$.

3.4 Analysis of model regression results

Combining the above model derivation process, we used MATLAB and selected the Gaussian process regression model with a linear basis function and an exponential kernel function as described above, and used the conjugate gradient method to find the hyperparameters. Since the sample size is relatively small, we use the cross-validation method by dividing the training set into 10 parts, of which 9 parts are used for training and 1 part for validation, and the cycle is taken 10 times to end. The regression results are shown in the figure below, and the Gaussian process regression can effectively fit the CO_2 emissions data. The relationship between record number and sample information in the left figure is shown in the table below. As the CO_2 emissions, i.e., the target value, increases, the deviation of the model prediction results widens, but the widening of the error relative to the overall value does not have an excessive impact on the prediction accuracy (Table 5 and Figure 2).

Table 5: Record number versus sample information

Record Number	Sample Information	Record Number	Sample Information
0-20	Australia	100-120	Britain
21-40	Brazil	121-140	Japan
41-60	Canada	141-160	New Zealand
61-80	China	161-180	U.S.A
81-100	France	181-200	South Africa
Note: The increase of serial number in a single sample represents the increase of the number of years (2000-2019)			

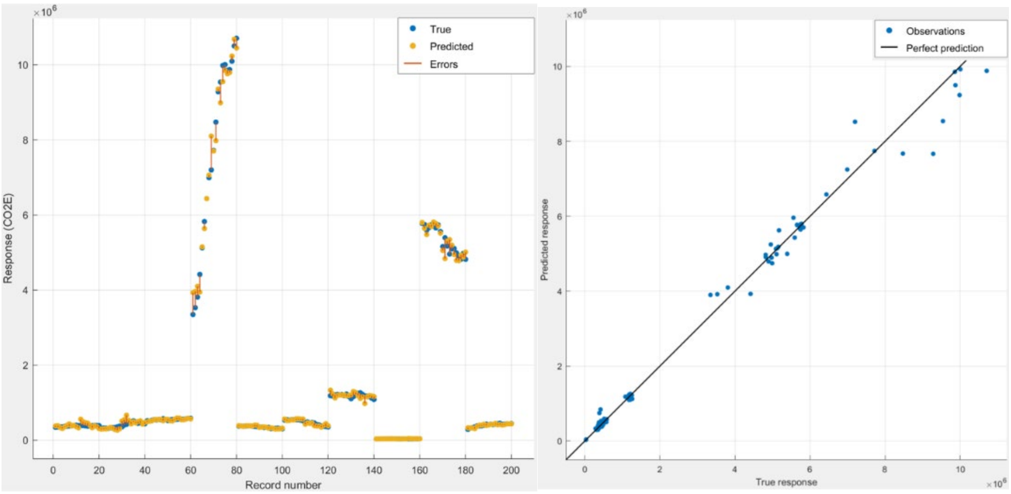

Figure 2: Gaussian process regression (left), Best predicted value control (right)

The conclusions obtained from the analysis of the errors of the regression are similar to the results of the best prediction value control: the prediction error values show a trend of increasing and then decreasing with the increase of GGDP and are basically stable, and the prediction errors have a significant tendency to increase with the increase of the true value.

The GGDP and other seven environmental indicators shown above provide a precise regression of CO_2 emissions, which helps countries to analyze what approach should be taken to solve the climate problem, which in turn provides a better assessment of the expected global impact of climate mitigation (Figure 3).

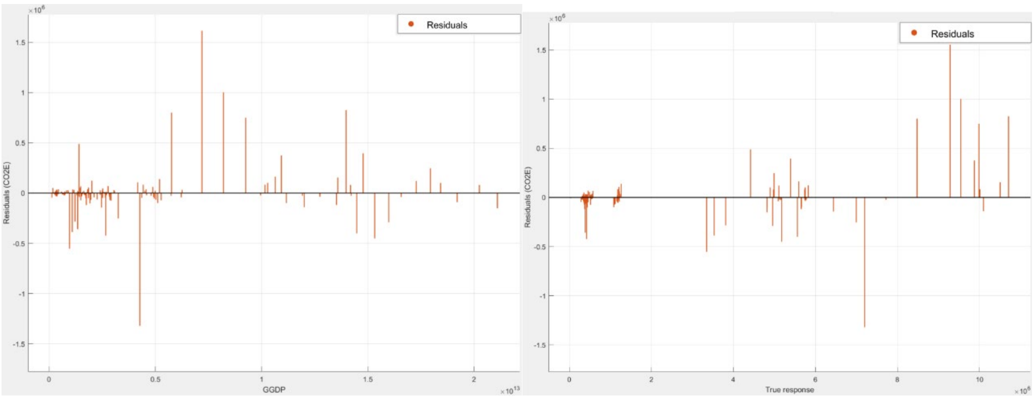

Figure 3: Regression Error Analysis

4. Conclusions

In this study, seven variables were first selected as secondary indicators, then carbon dioxide emissions were used as the dependent variable and a multiple regression was conducted with the seven variables identified as independent variables. In this regression, significant and robust results were obtained and it can be proved that using GGDP as a macroeconomic variable is more favorable to the environment. This paper also used the cross-validation method to train the Gaussian process regression model and obtained better regression results (R2 = 99.9 %, RMSE = 1.364e + 5). In summary, this study constructed a reasonable model to estimate the impact of GGDP on global climate mitigation, which can provide some reference value for measuring the economic health and sustainability level of a country.

References

[1] Marcus, R. D., Kane, R. E., 2007. US national income and product statistics: born of the great depression and World War II. Bureau Econ. Anal. Surv. Curr. Bus. 87, 32-46.

[2] McCulla, S. H., Smith, S., 2007. Measuring the Economy: a Primer on GDP and the National Income and Product Accounts. Bureau of Economic Analysis: US Department of Commerce, Washington, DC.

[3] Stiglitz, J. E., Sen, A., Fitoussi, J. -P., 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress. Commission on the Measurement of Economic Performance and Social Progress, Paris.

[4] Stiglitz, J. E., Sen, A., Fitoussi, J. P., 2010. Mismeasuring our lives: why GDP doesn't add up. The New Press, New York.

[5] Giannetti, B. F., Agostinho, F., Almeida, C. M. V. B., & Huisingh, D. (2015). A review of limitations of GDP and alternative indices to monitor human wellbeing and to manage eco-system functionality. Journal of cleaner production, 87, 11-25.

[6] Daly, H. E., Cobb, J., 1989. For the common good: redirecting the economy toward community, the environment, and a sustainable future. Beacon, Boston, MA.

[7] World Bank, 1997. Expanding the measure of wealth: indicators of environmentally sustainable development. World Bank, Washington, DC.

[8] Costanza, R., Farley, J., Templet, P., 2002. Background: the quality of life and the distribution of wealth and resources. Ecosummit 2000: Understanding and Solving Environmental Problems in the 21st Century. Elsevier.

[9] United Nations, 2003. Integrated environmental and economic accounting 2003. United Nations, New York.

[10] Ayres, R., 2004. On the life cycle metaphor: where ecology and economics diverge. Ecological Economics 48, 425–438.