China’s CO_2 Emission Prediction by Population and GDP Based on MLR and BP Neural Network

Yaning Zhang

Mathematics and Applied Mathematics, Wenzhou Kean University, Wenzhou, Zhejiang, China
**Corresponding author. Email: 1129863@wku.edu.cn*

ABSTRACT

Carbon dioxide (CO_2) emission is the amount of greenhouse gas emitted in processes, such as trading, production, or transportation. With the industry development, carbon dioxide emission grows significantly. However, too much CO_2 emission will cause a series of environmental issues, such as climate warming, glacial melting, and sea-level rising. Hence, it is urgent to realize influencing factors and take corresponding measures to protect the environment. Previous research has found that population and GDP are two major factors that cause CO_2 emission to increase, so building models to predict CO_2 emission is feasible and necessary. This paper will test the correlation between China’s population, GDP, and CO_2 emission, then use multiple linear regression (MLR) and backpropagation (BP) algorithm to establish CO_2 emission prediction models by inputting population and GDP data, and finally compare the advantages and disadvantages of the two models. The research shows that the BP algorithm is more suitable for prediction and the result is more accurate.

Keywords: *Population, GDP, CO_2 emission, multiple linear regression (MLR), backpropagation (BP) neural network.*

1. INTRODUCTION

Global warming caused by an excess of CO_2 emission has become a serious environmental problem, making many countries think about solutions. Since the second half of the 20th century, China's carbon dioxide emissions have increased very rapidly. In 2020, China's share of global carbon dioxide emissions is 30.65%, which is a very high value compared with other countries [1]. Much previous research has found a close relationship between gross domestic product (GDP), population, and CO_2 emission. Liu et al. [2] concluded that economic activity is a dominant factor for each region in China. Ma et al. illustrated that the economy’s rapid development and urbanization process are not favourable to reducing CO_2 emission [3]. Fan et al. [4] associated the impact of population on CO_2 emission with the income level. It shows that the population has the highest impact on emissions in upper-middle-income level countries, and lower impact on low-income level countries, and the least at those with lower-middle-income level. [5]. In addition, several models and their extensions have been built to study the current CO_2 emission situation. For example, STIRPAT has been used to find main factors [6] and predict the peak of CO_2 emission in China [7]. With the extended IPAT model, Song et al. [8] concluded that promoting technical development in the economic process and accelerating industrial restructuring is inevitable. LMDI is another helpful method widely used to analyse how the transport sector [9] or energy consumption influences CO_2 emission [10]. Previous studies thoroughly understand the influencing factors, but few established models based on these factors to predict the future values of CO_2 emission. This essay chooses population and GDP as two main factors, then builds the Multiple Linear Regression (MLR) model and Backpropagation (BP) neural network model to predict CO_2 emission, and finally makes a comparison of the two models about the prediction precision. The model building is necessary because it can provide researchers with a reference for future data, and helps them make decisions about how to optimize the industrial structure or control the amount of emission.

2. METHOD

2.1 The principle of the Multiple Linear Regression (MLR) method

Multiple linear regression (MLR) is used to estimate the value of observed variables through the linear relationship between the dependent variable and several independent variables. The formula can be expressed as:

$$Y = X\beta + \varepsilon \quad (1)$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad (2)$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (3)$$

In the equation, y is the dependent variable matrix, and X is the matrix of the independent variables. β is a constant term, which is also called intercept term; $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients. ε is the error value. The model needs to find out regression coefficients when the loss function minimizes.

Loss function:

$$Q(\hat{\beta}_0, \hat{\beta}_0, \ldots, \hat{\beta}_0) = \sum_{i=0}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2 \quad (4)$$

Regression coefficients expression calculated by the least square method is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5)$$

2.2 The principle of the Backpropagation (BP) algorithm

The backpropagation (BP) network algorithm is the feedforward network with multiple layers, generally containing the input, hidden, and output layers. Its self-learning process has two basic procedures, forward and backpropagation. Forward propagation transmits data from the input layer to the output layer. The input data (x_i) will be transferred to the hidden layer as I_j by multiplying initial weights (w_{ij}) and plus the bias (b_j) through the following formula:

$$I_j = \sum_{i=1} w_{ij} x_i + b_j \ (i, j = 1,2,\ldots,n) \quad (6)$$

After the transmission from the input layer to the hidden layer, the network will calculate the output values (o_j) through the sigmoid function:

$$o_j = \frac{1}{1+e^{I_j}} \quad (7)$$

After the network gets the output data, it will calculate loss between predicted values (o_j) and actual values (o_{true}), then it would transmit back to the hidden layer and input layer with certain functions and weights. Here is the loss function:

$$L = \frac{1}{n} \sum_{i=1}^{n} (o_{true} - o_j)^2 \quad (8)$$

Then, it will get updated weights ($w_{ij}^{'}$) by the gradient descent optimization algorithm with the learning rate (η) ranging from 0 to 1. Here is the function:

$$w_{ij}' = w_{ij} - \eta \frac{\partial L}{\partial w_{ij}} \quad (9)$$

3. MODEL BUILDING

3.1 Study design

This paper chooses the total Chinese population, GDP (current US $) from 1960 to 2020 as influencing factors. Datasets were downloaded from WorldBank.org. The target predicted variable is the annual amount of CO_2 emission in producing process, which does not include emissions embedded in traded goods [1]. It is got from Our World in Data website.

3.2 MLR model establishment

3.2.1 Zero-centered and standardization

Data needs zero-centered and standardization in the regression analysis before building the model. Zero-centered is the process that moves the center of the dataset into the original point, and the central value is obtained by using each value in the dataset minus its mean. If the model finishes the zero-centered process, it will not have a constant term, but it does not change coefficient values. After the zero-centered process, standardization is used to change the model's standard deviation to 1. Standardization can help the regression model eliminate the negative effect of orders of magnitudes and units, making the coefficients comparable. Here is the formula to standardization:

$$x' = \frac{x-\mu}{\sigma} \quad (10)$$

μ is the sample's mean, σ represents standard error, and x' is the value after standardization.

3.2.2 Correlation analysis

Before building the MLR models, variables need to take correlation analysis to test how they influence each other. Table1 shows the result:

Table 1. Correlation Among the Population, GDP, and CO_2 emission

Correlation Index	Population	GDP ($)	CO2 Emission
Population	1.00	0.68	0.86
GDP ($)	0.68	1.00	0.93
CO2 Emission	0.86	0.93	1.00

If the absolute value of the index is ranged 0.8 to 1, the variables are highly correlated. If the number ranges from 0.5 to 0.8, they are moderately related. Therefore, the table shows that population and GDP strongly correlate with CO_2 emission at 0.86 and 0.93, respectively. There is no multicollinearity between GDP and CO_2, because their correlation index is 0.68, lower than 0.8.

3.2.3 MLR model building

With the least square method, the model is built with the equation:

$$y = 0.4243x_1 + 0.6405x_2 \qquad (11)$$

y is the regression value of CO_2 emission, x_1 is the population variable, x_2 is the GDP variable, and the corresponding regression values are 0.4243 and 0.6405, respectively. The regression coefficients are both positive means that CO_2 emission increases with population and GDP rising. Besides, a larger coefficient with GDP illustrates that the impact of GDP on CO_2 emission is more significant than that with variable population, but the difference is not very large, because there is not much difference between the two values. The graph is drawn (figure 1) to show the fitting result:

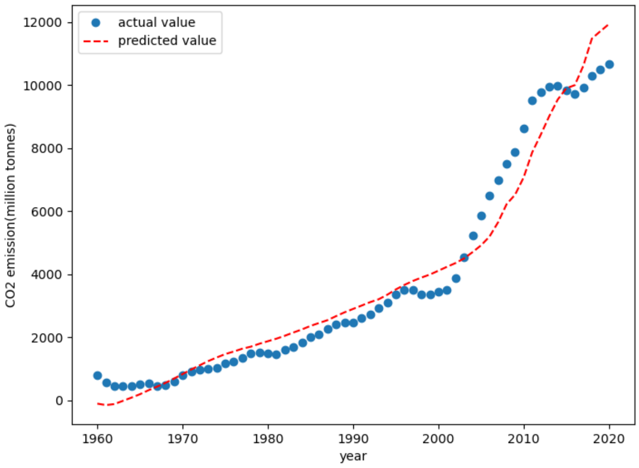

Figure 1 Fitting result with MLR model

The overall fitting result is good, because predicted values follow the trend of actual values. However, there are still many errors between predicted and actual values, so it cannot fit specific values well.

3.2.4 MLR model result analysis

F-test and t-test are chosen as two judgments to test the fitting results. F-test is used to test whether coefficients β_p are zeros. The null hypothesis is that $\beta_1, \beta_2, \dots, \beta_p = 0$, and the alternative hypothesis is that not all $\beta_1, \beta_2, \dots, \beta_p$ is equal to zero. If the null hypothesis is true, F-statistics is that:

$$F = \frac{SSR/p}{SSE/(n-p-1)} \qquad (12)$$

SSR is the regression sum of squares, and SSE is the residual sum of squares. Following the formulas, the F-statistics of this model is equal to 690.4, and the probability of F-statistics is 1.18e-41. The value of the probability of F-statistics is too small that the null hypothesis is hard to occur, so it can reject the null hypothesis. Significantly, a strong linear correlation exists between dependent and independent variables.

However, the F-test can only show that the general linear relationship is significant. It cannot illustrate that each independent variable has a significant linear relationship with the dependent variable, so it is necessary to use t-test to judge each variable's significance. Similar as F-test, the null hypothesis of t-test is that $\beta_j = 0$ for the specific variable, and here is the result:

Table 2. T-test result for MLR model

	coefficient	Std error	t	P > |t|
Population	0.4243	0.036	11.829	0.000
GDP	0.6405	0.036	17.854	0.000

Table 2 illustrates that the p-value of population and GDP is less than 0.05, so it can reject the null hypothesis, and the two independent variables have a significant linear relationship with CO_2 emission.

3.3 Establishment of BP neural network model

3.3.1 Normalization

In the BP algorithm, normalization can promote the convergence process faster and increase precision. Therefore, normalization process is necessary for datasets before building the model. As shown in the formula, x' is the value after normalization, x is the original data, and x_{max} and x_{min} are the maximum and minimum values of the original datasets.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (13)$$

3.3.2 Parameter Selection

3.3.2.1 Learning rate

Learning rate is an important parameter for building a suitable BP neural network model, because it is used to adjust the weights. If the learning rate is set too high, the network may fall into an optimal local solution. In contrast, if the learning rate is set too low, the time for finding optimal weights will be too long, but it is more likely to find the optimal global solution. Therefore, this model chooses 0.01 as the learning rate to find optimal weights.

3.3.2.2 Number of nodes for each layer

This model sets one input layer, one hidden layer, and one output layer. There are three nodes in the input layer, corresponding population data and GDP data from 1960 to 2020, respectively. Besides, one node is set for the output layer, which calculates the CO_2 emission values within 60 years. Therefore, if the number of nodes in the input layer and output layer has been determined, the number of nodes in the hidden layer can be calculated by the following formula:

$$m = \sqrt{n + l} + a \quad (14)$$

m represents the number of nodes in hidden layers, n is the number of nodes in input layers, l is the number of nodes in output layers, and a is a constant between 1 and 10. Therefore, the hidden layer's node is ranged from 3 to 12. By the training experiences, the model found that 11 is the optimal number for producing precious values.

3.3.2.3 Training epochs and target error

There are two conditions to terminate the training process. One is that the network reaches the maximum of training epochs. The other condition is that the network reaches the error less than the target error. If the BP network fulfills one of two conditions, it will stop. Therefore, the target error is set as 1e-4 and training epochs are set as 10000 to fulfill the target error.

3.3.3 Training result

Using the parameters, the network stops the training process because it reaches the target error. Here is the figure:

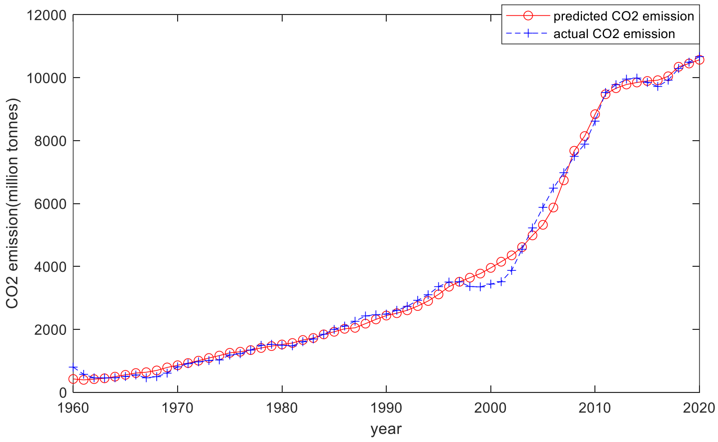

Figure 2 Fitting result with BP algorithm

As the figure shows, the model predicts the rising trend of CO_2 emission well. Almost all predicted values fit actual values, so the sum of errors between predicted and actual values is much smaller than the MLR model.

3.4 Comparison of two models

To compare the accuracy of two models, the author chooses Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared as three judgments. Here is the result:

Table 3. MSE, MAE, and R-squared of two models

	MSE	MAE	R-squared
MLR model	0.0403	0.1578	0.9590
BP neural network	0.0042	0.0468	0.9979

All the parameters are calculated after normalizing predicted values and actual values. The table shows that both MSE and MAE of the BP neural network model are smaller than those in the MLR model, which means that the BP neural network can predict CO_2 emission with more precious values and fewer errors. Besides, the BP neural network model's R-squared is higher than that in the MLR model, which can also prove that the fitting result in BP neural network is better than the MLR model. Therefore, BP neural network model is more feasible for predicting CO_2 emission with more accurate predicted values.

4. DISCUSSION

Comparing the CO_2 emission prediction results of two models, it is found that BP neural network has higher accuracy than the MLR model. The reason is that the BP neural network has a self-learning process, and it is not limited to the linear relationship between independent and dependent variables. It can update the weights and bias through the learning process, decreasing the errors to produce a better fitting result. The MLR model can only fit the linear relationship among variables, and noises or interaction effects may cause the predicted results lack accurate. One common advantage of the two models is that they both do not need a large sample size to train the model, and they can find the best result quickly. However, the BP neural network has higher requirements for parameters choices. If the parameters are not selected optimally, the network may cause many problems, such as over-fitting and optimal value loss. Therefore, each model has its strengths and weaknesses, but BP neural network is more likely to fit the actual values with higher accuracy in this situation.

5. CONCLUSION

In conclusion, this paper finds that the correlation among the population, GDP, and CO_2 emission is high, so it builds the MLR and the BP neural network model to use population and GDP data from 1960 to 2020 as independent variables for predicting the CO_2 emission. The F-test and t-test check the significant linear relationship between independent variables and the target variable. The predicted values fit the actual values moderately well, whose R-squared equals 0.9590. However, BP neural network model can produce predicted values closer to actual values with much lower errors, so the fitting result is better than the MLR model with R-squared equal to 0.9979. Therefore, the BP neural network is the more suitable model for predicting CO_2 emission. However, this paper still has some limitations. The MLR model in this paper has relatively lower accuracy, so it still has potential for improvement. In the future, it should optimize the MLR model and BP neural network model to predict the CO_2 emission with higher precision. In addition, more research will focus on how to decrease the CO_2 emission with the influencing factors, and then find out the feasible and optimal value as the emission reduction goal.

REFERENCES

[1] Hannah Ritchie and Max Roser 2020, CO_2 and Greenhouse Gas Emissions. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

[2] Liu, B., Shi, J., Wang, H., Su, X., & Zhou, P. Driving factors of carbon emissions in China: a joint decomposition approach based on meta-frontier. Applied Energy, 256, (2019) 113986.

[3] Ma, X., Wang, C., Dong, B., Gu, G., Chen, R., Li, Y., … & Li, Q. Carbon emissions from energy consumption in China: its measurement and driving factors. Science of the total environment, 648, (2019) 1411-1420.

[4] Fan, Y., Liu, L. C., Wu, G., & Wei, Y. M. Analyzing impact factors of CO_2 emissions using the STIRPAT model. Environmental Impact Assessment Review, 26(4), (2006) 377-395.

[5] Puliafito, S. E., Puliafito, J. L., & Grand, M. C. Modeling population dynamics and economic growth as competing species: An application to CO2 global emissions. Ecological Economics, 65(3), (2008). 602-615.

[6] Li, H., Mu, H., Zhang, M., & Li, N. Analysis on influence factors of China’s CO2 emissions based on Path–STIRPAT model. Energy Policy, 39(11), (2011) 6906-6911

[7] Li, L., Lei, Y., He, C., Wu, S., & Chen, J. Prediction on the Peak of the CO2 Emissions in China Using the STIRPAT Model. Advances in Meteorology, 2016(8):1-9.

[8] Song, M., Wang, S., Yu, H., Yang, L., & Wu, J. To reduce energy consumption and to maintain rapid economic growth: Analysis of the condition in China based on expended IPAT model. Renewable and Sustainable Energy Reviews, 15(9), (2011) 5129-5134.

[9] Wang, W. W., Zhang, M., & Zhou, M. Using LMDI method to analyze transport sector CO2 emissions in China. Energy, 36(10), (2011) 5909-5915.

[10] Xu, J. H., Fleiter, T., Eichhammer, W., & Fan, Y. Energy consumption and CO_2 emissions in China's cement industry: A perspective from LMDI decomposition analysis. Energy policy, 50, (2012) 821-832.