Environ. Eng. Res. 2017; 22(4): 407-416
https://doi.org/10.4491/eer.2017.023

pISSN 1226-1025
eISSN 2005-968X

An investigation and forecast on CO_2 emission of China: Case studies of Beijing and Tianjin

Lei Wen[1,2†], Zeyang Ma[2], Yue Li[2], Qiao Li[2]

[1]*The Academy of Baoding Low-Carbon Development, Baoding 071003, China*
[2]*Department of Economics and Management, North China Electric Power University, Baoding 071003, China*

ABSTRACT

CO_2 emission is increasingly focused by public. Beijing and Tianjin are conceived to be a new economic point of growth in China. However, both of them are suffering serious environmental stress. In order to seek for the effect of socioeconomic factors on the CO_2 emission of this region, a novel methodology –symbolic regression– is adopted to investigate the relationship between CO_2 emission and influential factors of Beijing and Tianjin. Based on this method, CO_2 emission models of Beijing and Tianjin are built respectively. The models results manifested that Beijing and Tianjin own different CO_2 emission indicators. The RMSE of models in Beijing and Tianjin are 255.39 and 603.99, respectively. Further analysis on indicators and forecast trend shows that CO_2 emission of Beijing expresses an inverted-U shaped curve, whilst Tianjin owns a monotonically increasing trend. From analytical results, it could be argued that the diversity rooted in different development orientation and the mixture of different natural and industrial environment. This research further expands the investigation on CO_2 emission of Beijing and Tianjin region, and can be used for reference in the study of carbon emissions in similar regions. Based on the investigation, several policy suggestions are presented.

Keywords: Beijing and Tianjin, CO_2 emission, Driven factors, Genetic programming, Pareto front, Symbolic regression

1. Introduction

The public concern is increasingly focused on the climate change. It has reached a consensus that climate change is primarily driven by the greenhouse gases (GHG) emission in the atmosphere. CO_2 is perceived to be the most notable GHG gas associated with human activities.

Beijing and Tianjin are perceived to be an important economic point of growth in north China. From 1995 to 2015, Gross Domestic Product (GDP) of Beijing and Tianjin reached an average 14.95% growth annually, 69.52% of population growth and 84.90% of average urbanization level. On the contrary, the fast development brings a huge amount of CO_2 emission. During this period, CO_2 emission in Beijing and Tianjin almost doubled, from 183.19 million tons in 1995 to 332.34 million tons in 2015. It could be argued that Beijing and Tianjin are suffering serious environmental stress.

Therefore, it is imperative for Beijing and Tianjin to seek for the effect of economic growth on the environment so as to balance environmental protection against the development process. Thus, this study intends to explore the link between CO_2 emissions and certain indicators in the case of Beijing and Tianjin.

In recent years, many experts are committed to inquiring into CO_2 emission and its influential factors. The research could be classified into two groups: Decomposition methods and regression methods.

On one hand, decomposition methods are commonly applied in studying CO_2 emissions. IPAT, ImPAT, STRIPAT and LMDI are commonly applied among decomposition models.

The IPAT model is a method focusing the impacts of human activities, including population growth, economic growth and technological progress, on pollutant [1]. After its initial presentation, the IPAT identity has been regarded as an easily understandable, widely utilized framework for analyzing the driving forces of environmental change [2]. Waggoner and Ausubel further disaggregated a fourth variable C (the intensity of energy use) into per unit of GDP (A) and impact per unit of consumption (T), thus created ImPACT model [3]. IPAT model allows to explicitly identify the relationship between the driving forces and environmental impacts, but it is also been criticized because IPAT model assumes a proportional relationship between environmental indicators and influen-

Received February 22, 2017 **Accepted** June 18, 2017

† Corresponding author
Email: wenlei0312@126.com
Tel: +86-1593-3901836 Fax: +86-0312-7525115

tial factors [4]. In addition, although ImPACT advanced the IPAT model in allowing room for diagnostic analysis, both were equations with fixed factors assuming proportionality between the key determinant factors, which limit further application of the models. The STIRPAT model is one of the most popular measures used in studies on CO_2 emissions. It was proposed by Dietz and Rosa [5-7]. It gives a chance to introduce more variables during analysis and it is much more flexible to test the impacts of each factor on environmental pressures.

Various researches on the CO_2 emission in China using decomposition models are conducted. For instance, Hubacek et al [2] conducted an IPAT analysis for China and made the conclusion that increase of affluence has been the main driving force for China's CO_2 emissions since the late 1970s. Ang [8] and Wang [9] came to the conclusion by using the LMDI method that economic growth is a leading cause of carbon emissions and energy intensity is seen as essential effect if China's carbon emissions are looked forward to being reduced over the long term.

On the other hand, regression methods are also employed to investigate CO_2 emission and influential factors.

Regression methods are widely applied in various fields. From macroscopic point of view, regression methods have been successfully applied in various atmosphere sciences [10-15]. On further research on atmosphere sciences, one of the commonly applied regression methods on CO_2 emission is Environmental Kuznets curves (EKC). EKC is conceived as an inverted U-shaped curve model of the connection among energy consumption, economic growth, and the environment as well [16]. The EKC hypothesis reveals that environmental pollution will increase until reaching a peak and then will start declining over time with economic growth [17]. EKC is commonly adopted to investigate the relationship between environmental degradation and economic growth [18, 19].

In addition, on application of regression methods, various researches, which aim at exploring the relationship between CO_2 emission and socioeconomic factors, have been conducted [20-24]. Based on historical data, these researches quantitatively analyzed the link between CO_2 emission and socioeconomic factors from national perspective, and guaranteed satisfactory results. What's more, these researches further expand the application of CO_2 emission, providing information on relative policy implications.

In artificial intelligence, Genetic Programming (GP) is conceived to be an effective methodology to deal with optimization problems. This algorithm is first proposed by Koza [25]. Essentially, GP is a set of instructions and a fitness function to measure how well a computer has performed a task. The process which targets at producing a computer program linked to a certain data set is also called symbolic regression [26]. Symbolic regression is a common application of genetic programming. Unlike traditional linear or nonlinear regression methods that fit parameters to an equation of a given form, symbolic regression searches both the parameters and the form of equations simultaneously [27]. Without assumed functional forms, symbolic regression method can get insight about the generating systems hidden in various data [28]. It could be argued that symbolic regression using genetic programming is an ideal algorithm for automatically determining an otherwise unknown functional relationship between a set of inputs and outputs.

The commonly applied models are widely applied and frequently proved by former researchers. However, as for Beijing and Tianjin, the traditional models usually adopt fixed models and parameters, which limited further application and investigation of the CO_2 emission research. What's more, it is limited to select parameters by personal experience from the complex socioeconomic system. Sometimes, the drawbacks of traditional models could induce contradiction in results. This phenomenon is obviously expressed in the research based on the EKC model. As Lau et al. [29] stated that there exists an EKC model in Malaysia during 1970-2008, whilst Azlina et al. [30] argued that there was no EKC model in Malaysia from 1975 to 2011.

In this paper, symbolic regression method is adopted to further investigate the relationship between CO_2 emissions and its influential factors of Beijing and Tianjin. The application of symbolic regression in this paper wouldn't adopt fixed parameters and equations like traditional methods, but automatically discover the hidden relationship between CO_2 emission and socioeconomic factors of Beijing and Tianjin. This method aims at avoiding the drawbacks of traditional models and attempts to explore new equations for CO_2 emissions, and make forecast and analysis based on the discovered models.

2. Methods

2.1. Methodology

The advantage of symbolic regression lies in its ability to automatically discover the hidden functional relationship without domain knowledge. Symbolic regression would determine the parameters and structures simultaneously other than traditional regression method, which must be predefined a certain function form, such as liner, quadratic equation, natural logarithm and so on. Due to the characteristics of symbolic regression, it is accepted as the "Robot scientist" for automated knowledge discovery [31]. It can be argued that the symbolic regression is successfully applied as a novel automatic discovery method in modeling and optimization problems. Schmidt and Lipson [27] conducted symbolic regression and discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. Yang, et al. [17] applied symbolic regression and discovered four models, including the inverted N-shaped, M-shaped, inverted U-shaped and monotonically increasing without domain experts' intervention, and the newly discovered M-shaped model has received little attention in previous studies but exhibits promising performance. Bahrami [32] conducted GP as a novel method for modeling the Recovery Factor (RF) and the Net Present Value (NPV) in Surfactant–Polymer (SP) flooding, and achieved satisfactory optimization results. Palancz et al. [33] conducted symbolic regression to treat the problem of geoid correction based on GPS ellipsoidal height measurements, the result proposed SR method could reduce the average error to a level of 1-2 cm.

Symbolic regression, based on genetic programming, is applied in this paper as an evolutionary function discovery method. It is intended that the relationship between certain factors and CO_2 emission is automatically discovered by the symbolic regression. Based on the result of this method, the validation of the result

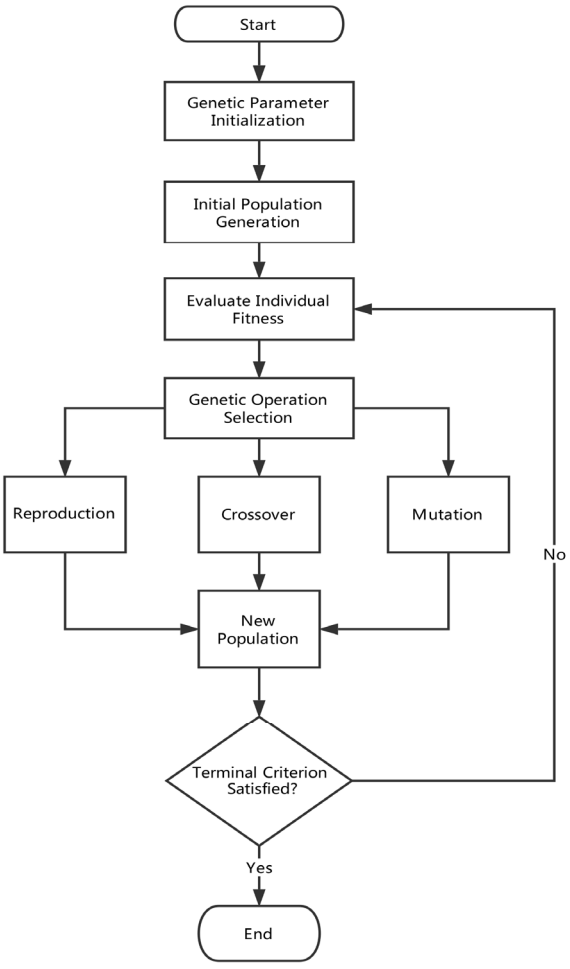

Fig. 1. Flowchart of genetic programming.

is conducted. Further exploration of Beijing and Tianjin is conducted to investigate the relationship between CO_2 emission and certain factors, which aims to check if there exists a general model which performs satisfactory result.

The flow chart of genetic programming is shown in Fig. 1. The function forms in genetic programming are conducted with the syntax tree. Fig. 2(a) illustrates an example of syntax tree. Two types of nodes in a syntax tree are included: Functional nodes and terminal nodes. Functional nodes include functional symbols like numerical operators (+, -, ×, /, sin, cos, etc.), logistical operators, etc. Terminal nodes include terminal symbols such as input variables and constants.

Three basic genetic operations are adopted to implement symbolic regression, which are reproduction, crossover and mutation. Reproduction operation could keep the better individuals survive into the next generation; crossover operation, equivalent to the sexual reproduction procedure in nature, combines two individuals (syntax tree) and creates two new individuals; mutation operation mutates a select node or part of syntax tree, which could play an important role in increasing the individual diversity and avoiding local optimal solution. The detailed description of these operations is shown in Fig. 2(b).

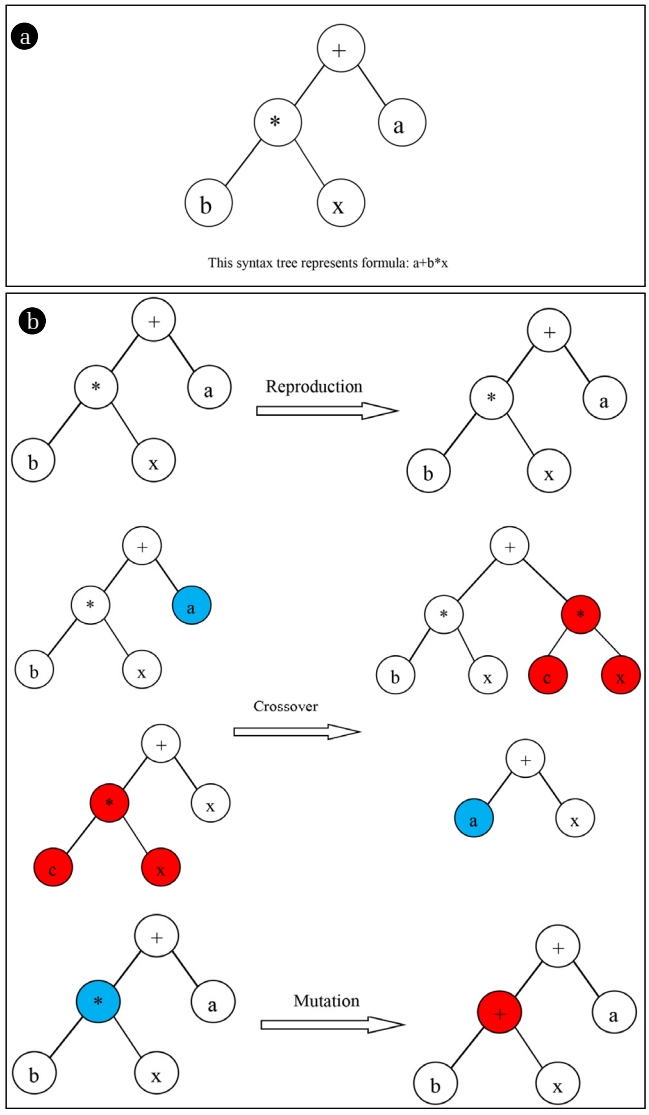

Fig. 2. Detailed illustrations of syntax tree and genetic operations.

Generally speaking, it follows three major steps to conduct symbolic regression.

(1) Parameters initialization. Primitive functions and variables used for exploring the hidden relationships should be pre-defined before conducting the symbolic regression.

(2) Modeling operation. This step includes modeling the variables and parameters. The application of genetic programming, including the genetic operations such as reproduction, crossover and mutation, is also included.

(3) Fitness evaluation. The direction of genetic operation was guided by the fitness evaluation procedure. Individuals with better performance will have more chance to survive and individuals with poor performance will gradually diminish. In addition, fitness evaluation could play an important role in evaluation of terminal criterion.

As for genetic programming, the promising individuals will have more chance to survive to the next generation, whilst the poor-per-

formance individuals will gradually diminish. Complexity and fitness are conflicting features leading to a multi objective problem [34]. According to the Occam's Razor, if there are two models with the same accuracy, the model with less complexity is preferred. In this paper, two objectives were used to evaluate the models. One is the fitness, which is considered to evaluate if this model fits the data well, the other is the complexity, which is calculated by the node count of the tree. A Pareto front could be built based on the fitness and complexity of the models. It can be used to detect the best solutions among thousands of millions of candidates. In this paper, the mean absolute error (MAE) was used to measure the fitness.

In summary, two principles guide our method:

(1) The important factors and functions will frequently emerge;

(2) Models on the Pareto front will be selected.

2.2. Data

In order to conduct the symbolic regression, the panel data form of Beijing and Tianjin is selected. The data covered 21 y from 1995 to 2015. 19 y data from 1995 to 2013 is used as the training data, the data of year 2014 and 2015 is used as validation data. The original data was collected from China Energy Statistical Yearbook and the Statistical Yearbook of Beijing and Tianjin.

Taking into account the factors applied in former researchers [35-37], the driving factors of energy requirements and carbon emission can reduce to economy growth, industrial structure, population and urbanization, technological improvement and innovation, energy structure, living standard improvement and energy-saving policies. Living standard improvement factor and energy-saving policies factor are not involved in modeling for the purpose of difficulty in quantification. From the macroscopic perspective, taking into account the efficiency while conducting the symbolic regression, the variables are defined as follows in Table 1 to investigate the influence on the CO_2 emission.

The carbon dioxide emission data is calculated via the algorithm produced by IPCC. This strategy is a top-down approach, using a country's energy supply data to calculate the emissions of CO_2 from combustion of mainly fossil fuels [38]. Following the IPCC guidelines and former research [39], the CO_2 emission from energy consumption is calculated as follows:

$$E_{CO_2} = \sum_{fuels} \frac{A_i \times NCV \times CC_i \times O_i \times 44}{12} \quad (1)$$

Where E_{CO_2} is the total consumption of CO_2, A_i is the amount of fuel i consumption, NCV is the net calorific value of fuel i, CC_i is the carbon content of fuel i, O_i is the carbon oxidation factor of fuel i. In this paper, the carbon oxidation factor was selected default value 1. Fuels considered in this paper are widely accepted by researchers: Coal, coke, crude oil, gasoline, kerosene, diesel, fuel oil and natural gas. The net calorific value and carbon content of these fuels are shown in Table 2.

Table 1. The Factors and Indicators Applied in This Paper

Factors	Variables	Indicators	Unit
CO_2 emission	y	the amount of CO_2 emission	10,000 kg
Economic growth	x1	Gross Domestic Product (GDP)	100 million RMB
Total population	x2	Total population	10,000 persons
Industrial structure	x3	the industrial share of GDP	none
Energy structure	x4	the coal consumption share of TEC	none
Technology and innovation	x5	Energy intensity	TEC/10,000 RMB
Urbanization	x6	the urban residents' share of total population	none

Table 2. NCV and CC of the Fuels

Fuel	Net calorific value (kJ/kg)	Carbon content (kg/GJ)
Coal	20,908	26.8
Coke	28,435	29.2
Crude oil	41,816	20
Gasoline	43,070	20.2
Kerosene	43,070	19.6
Diesel	42,652	20.2
Fuel oil	41,816	21.1
Natural gas	38,931	15.3

a. GB/T 2859-2008 General principles for calculation of the comprehensive energy consumption

b. 2006 IPCC Guidelines for National Greenhouse Gas Inventories

3. Results

In this paper, we choose the commonly used symbols that appear in the models detected by former researchers: + (addition), - (subtraction), × (multiplication), / (division), exponential, natural logarithm, power and square root. The models will be deeper investigated to identify the factors which will appear in the optimal models.

Take Beijing for example, the panel data from 1995 to 2013 was selected and used to run symbolic regression. According to the thousands of candidate models, the Pareto front was built, it is widely accepted that the most promising models lies on the Pareto front, and it is suitable for simultaneously balancing fitting accuracy and model complexity [40]. The Pareto front of Beijing in one trail is shown in Fig. 3. After the construction of Pareto front, we should concentrate on the models on the Pareto front. In order to investigate the important factors and models, the symbolic regression procedure is repeatedly conducted. It is obvious that the model which fits the result best should be selected. In addition, the models which frequently emerge on the Pareto front could be conceived to be more likely to attach to the authentic relationship.

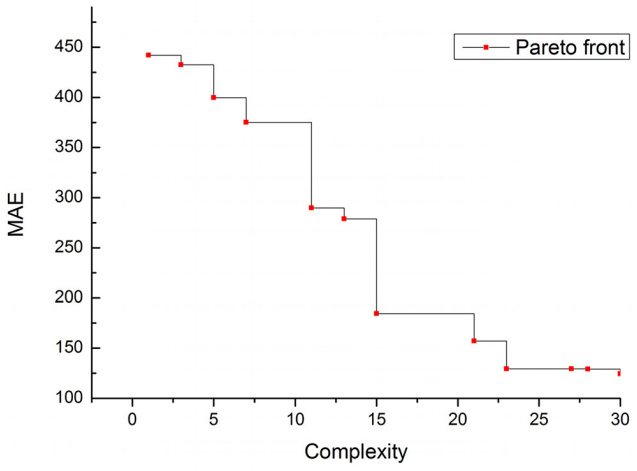

Fig. 3. The Pareto front of Beijing.

Take Beijing for example, the models returned from the symbolic regression are illustrated in Eq. (2) to Eq. (7). The model which fits the given data best is demonstrated in Eq. (2). It is obvious that this model fits the data best, but it also owns the highest complexity. The more complex candidate model usually expresses better fitting performance, but it also suffers from a higher risk of over-fitting [17]. In order to check if these models are sensitive to the present data sets and keep the risk of over-fitting under control, Leave-One-Out Cross Validation (LOOCV) method is conducted. Cross-validation is a measurement of assessing the performance of a predictive model, and statistical analysis will be generalized to an independent dataset [41]. What is generally accepted is that 3 commonly adopted cross validation methods are conducted by researchers: Hold-out Cross Validation, K-fold Cross Validation and LOOCV. LOOCV not only fully utilizes the available data, but also eliminates the influence of choices of random pairing. The generalization error of LOOCV is nearly unbiased, thus this could make a reliable result of estimation.

$$y = a + b^*x1^*x6 + c^*x2^*x3 + d^*x1^*x3 + e^*x1^*x2 + f^*x1^2 - g^*x1 - h^*x2 - i^*x3 - j^*x6 - k^*x1^*x2^*x3 - l^*x6^*x1^2 \quad (2)$$

$$y = a^*x3 + b^*x1 - c - d^*x1^*x3 - e^*x1^2 - f^*x3^2 \quad (3)$$

$$y = a + b^*x5 + c^*x2 + d^*x1 - e^*x2^*x5 - f^*x1^2 \quad (4)$$

$$y = a^*x6 + b^*x5 + c^*x1 + d^*x5^*x1^2 - e^*1000 - f^*x1^*x5 - g^*x1^2 \quad (5)$$

$$y = a^*x6 + b^*x5 - c \quad (6)$$

$$y = a^*x6 - b \quad (7)$$

In this paper, the cross validation is conducted and the mean RMSE of each model is calculated to evaluate the performance. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})}{n}} \quad (8)$$

y_i stands for the predicted value of year i. $\hat{y}$ stands for the truth value of CO_2 emission. The results of cross validation are listed in Table 3. It could be obviously observed that the models with lower complexity usually express poorer performance, whilst the models with higher complexity perform better, but too complex models would have higher risk of over-fitting. Such as the model No. 2 (Eq. (2)) fits the given data most, but its RMSE not performs best. It can be argued that model No. 2 expresses over-fitting. Model No. 1 (Eq. (3)) has the lowest RMSE and relatively low complexity, therefore, the model No. 1 can be considered to be the best model. Similarly, the selected models and cross validation results of Tianjin are also listed in Table 3.

As is shown in Table 3, the best models selected by symbolic regression and cross validation of Beijing and Tianjin are listed as follows.

$$Beijing: y = a^*x3 + b^*x1 - c - d^*x1^*x3 - e^*x1^2 - f^*x3^2 \quad (9)$$

$$Tianjin: y = a + b^*x1 + c^*x6^2 + d^*x4^2 - e^*x2 - f^*x4 - g^*x6 - h^*x1^*x6 \quad (10)$$

It could be concluded from the formulas that the carbon emission of Beijing is mainly associated with GDP and the industrial share of GDP. CO_2 emission in Tianjin is related to GDP, total population, the coal consumption share of TEC, the urban residents' share of total population.

Table 3. Models Returned from Symbolic Regression

- Beijing

Index	Complexity	RMSE	Model
1	27	255.38802	$y=a*x3+b*x1-c-d*x1*x3-e*x1^2-f*x3^2$
2	63	353.06759	$y=a+b*x1*x6+c*x2*x3+d*x1*x3+e*x1*x2+f*x1^2-g*x1-h*x2-i*x3-j*x6-k*x1*x2*x3-l*x6*x1^2$
3	25	357.64344	$y=a+b*x5+c*x2+d*x1-e*x2*x5-f*x1^2$
4	34	381.70980	$y=a*x6+b*x5+c*x1+d*x5*x1^2-e*1000-f*x1*x5-g*x1^2$
5	9	623.66761	$y=a*x6+b*x5-c$
6	5	625.48895	$y=a*x6-b$

- Tianjin

Index	Complexity	RMSE	Model
1	35	602.99536	$y=a+b*x1+c*x6^2+d*x4^2-e*x2-f*x4-g*x6-h*x1*x6$
2	63	610.83346	$y=a*x2+b*x1*x3+c*x3^2+d*x1^2+e*x1^3-f-g*x1-h*x3-i*x2*x3-j*x1^4-k*x3*x1^2$
3	35	612.71188	$y=a*x6+b*x1+c*x6^3+d/(x6-e)-f-g*x1*x6-h*x6^2$
4	32	631.54156	$y=a+b*x1+c*x6^2+-d/(x6-e)-f*x2-g*x6-h*x1*x6$
5	25	631.54232	$y=a+b*x1+c*x6^2-d*x2-e*x6-f*x1*x6$
6	15	761.05129	$y=a*x2+b*x1-c-d*x1*x2$
7	9	993.24015	$y=a*x3+b*x2-c$
8	9	1,036.4123	$y=a+b*x1-c*x5$
9	5	1,176.3559	$y=a+b*x1$

4. Discussion

According to the CO_2 emission value, which takes 1995 as the base year, the CO_2 emission trend was demonstrated in Fig. 4(a). It can be easily concluded that Beijing and Tianjin have experienced an increase in CO_2 emission, whilst the trend of the two cities were quite different. Carbon emission increase in Tianjin is relatively mild, but it surpassed Beijing at year 2005 and experienced a maximum of 177.01% increase from year 1995 to year 2015. In Beijing, the condition was quite different. Beijing experienced an increase of 40.5% of CO_2 emission in the year 2010, much lower than that of Tianjin. After that, the CO_2 emission started to decrease. From 2010 to 2015, the carbon emission even experienced a sharp decrease of 14.94%.

From Fig. 4(a), it is easy to conclude that during the past 21 y, CO_2 emission in Tianjin kept a steady growth, Beijing experienced less increase. In the recent 4 y, the CO_2 emission was obviously under control, Beijing experienced an obvious steady decrease.

Drawing from the results returned from symbolic regression, the CO_2 emission forecasting model could be constructed. In this paper, the data necessarily for constructing the model from 1995 to 2013 is used to build the model. Based on the models, the CO_2 emission of Beijing and Tianjin could be conducted. The comparisons of truth value and predicted value of the year 2014 and 2015 are illustrated in Table 4.

Table 4. CO_2 Emission Forecast Result

	Beijing		Tianjin	
	2014	2015	2014	2015
Truth value	12,825.62	12,341.20	21,362.84	20,893.03
Predicted value	12,380.15	12,077.53	21,740.89	23,419.10
Relative error	3.47%	2.14%	1.77%	12.09%

In this paper, the future trend of indicators adopted in the models is predicted by time series prediction method. The CO_2 emission trend of 2014-2020 is demonstrated in Fig. 4(b).

Based on the result of CO_2 emission forecast, it seems that during the forecast period, Beijing and Tianjin have different features. CO_2 emission of Beijing expressed an inverted-U shaped curve as time goes on. The forecast result illustrates that Beijing has already reached the peak in 2010 and starts to decrease. As for Tianjin, CO_2 emission will continue to increase till 2020. This result meets with results from former researchers [42]. It is recommended that CO_2 emission of Beijing and Tianjin will not decrease until 2020. However, in this paper, it is more optimistic for Beijing. In order to further investigate the CO_2 emission of Beijing and Tianjin, the analysis based on the models and forecast result is conducted.

It could be concluded from the model that the CO_2 emission in Beijing and Tianjin is related to GDP. This conclusion is consistent

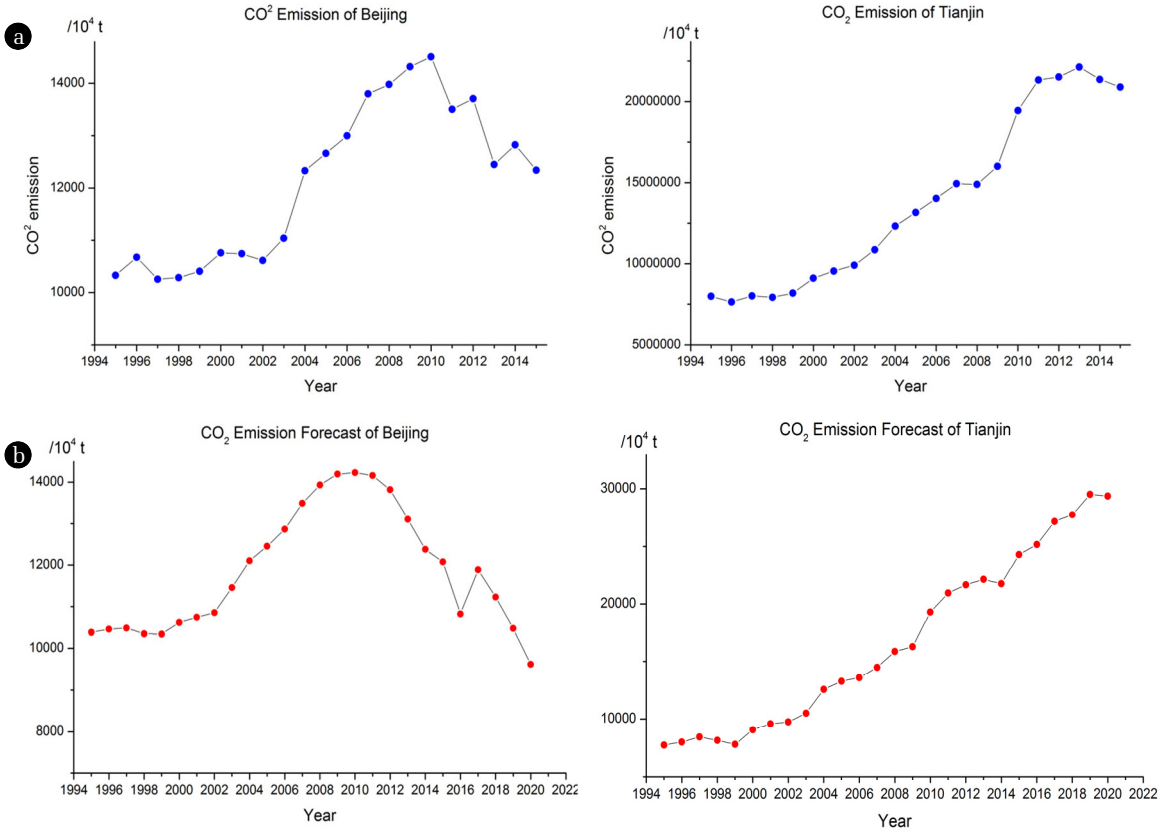

Fig. 4. CO_2 emission and forecast result of Beijing and Tianjin.

with results obtained by former researchers [43, 44]. The trend of GDP in Beijing and Tianjin is demonstrated in Fig. 5. It could be obviously observed from the trend of data that the basic monotonically increasing trend of economic growth meets with CO_2 emission increasing trend during this period. It can be argued that Beijing and Tianjin have all experienced a relatively fast and steady economic growth. From year 1995 to year 2015, the GDP in Beijing grows from 1,507.7 × 10^8 yuan to 23,014.6 × 10^8 yuan, which is 15.26 times' increase. In Tianjin, GDP experienced 17.74 times' increase. What interests us is that from 2010 to 2012, the economic development in Beijing and Tianjin has experienced a sharp increase. It is consistent with the GDP in Tianjin, whilst the CO_2 emission in Beijing started to decrease during this period. 2010 is the final year of 11th five year plan, and 2011 is the first year of 12th five year plan. It can be argued that it is the 12th five year plan that focuses on the industrial transformation and upgrade, which intends to increase the core competitiveness that makes the government carry out policies to reintegrate the industry in Beijing and Tianjin. However, based on the analysis of GDP, it could hardly give an exactly explanation of the different expression of CO_2 emission in Beijing and Tianjin. Thus, it is essential to further investigate on the relationship between different factors and CO_2 emission.

To make further investigation, the comparative data of Beijing and Tianjin is illustrated in Table 5.

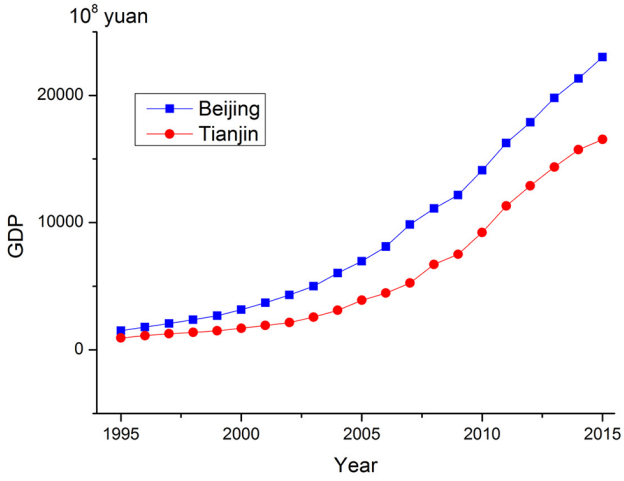

Fig. 5. GDP of Beijing and Tianjin.

As for Beijing, the industrial share of GDP is also conceived to be related to the CO_2 emission. In this paper, industrial structure factor is defined as the industrial share of GDP. It is demonstrated that the industrial share of GDP in Tianjin remains stable during this period, but in Beijing, the trend is obviously different. The industrial share of GDP in Beijing decreased 18.9%. It could be observed that Beijing had taken effective measures to further adjust

Table 5. The Comparison of Beijing and Tianjin during 1995-2014

	GDP(x1)	TP(x2)	ISG(x3)	CCST(x4)	EI(x5)	URST(x6)
Beijing						
1995	1,507.7	1,251.1	35.0	54.42	2.34	75.63
2015	23,014.6	2,170.5	16.1	13.7	0.338	86.51
Tianjin						
1995	931.97	941.83	50.2	67.49	2.58	53.93
2015	16,538.19	1,546.95	42.2	39.23	0.50	82.64

the industrial structure, and adhered to "3-2-1" industrial development situation. During the 11th five year plan and 12th five year plan, Beijing started to remove the industries which beyond the environmental carrying capacity or couldn't match the development orientation of Beijing. In the perspective of city orientation, Beijing, being the capital of China, should be a combination of political center, cultural center, international exchange center and national innovation center. Beijing seeks to cultivate new economic growth point and focuses on the development of tertiary industry. As for the industries in Beijing, it starts to transform from traditional industries to the technology-driven manufacturing, and accelerates the integration of traditional industries and information technology. For example, Shougang group moved to Tangshan in 2008; three new high-tech industrial areas, Zhongguancun, Yizhuang and Shunyi were established during 1995-2011. It could be argued that the implementation of these development policies could explain the high speed economic development when a relatively low CO_2 emission standard was kept.

It can be concluded from the model given by symbolic regression that GDP, total population, the coal consumption share of TEC and the urban residents' share of total population are linked to the CO_2 emission in Tianjin. Compared with Beijing, it could be observed that Tianjin expressed much more complexity in the perspective of carbon emission factors. Tianjin not only owns the richest soil resource with the largest comprehensive harbor, but also lays a solid manufacturing industrial foundation. The advantages of opening and modern industry make Tianjin express a characteristic of the mixture of industrial city and service-type city. As Wang and Yang [44] stated, population growth had an incremental effect on carbon emissions in Beijing and Tianjin. Birdsall [45] considered that there were two aspects of population growth affecting carbon emissions. First, a steadily growing population and increasing incomes create a higher demand for goods and services which consume resources and energy and generate pollutants and greenhouse gases at the same time in every production procedure. Second, a larger number of cultivated lands are occupied. Soil erosion, soil fertility degradation, soil degradation and desertification, environmental deterioration and other issues caused by irrational use of land resources become more serious. In Tianjin, the fast pace of population growth and urbanization experienced a growth of 64.24% and 28.70% respectively, leading the surrounding region. It could be argued that it helped to explain the link between total population, urbanization and CO_2 emission. Furthermore, in this model, the factor coal intensity-x4 is conceived to be an indicator of CO_2 emission. Coal intensity factor helps to explain the industrial transformation conducted in Tianjin. During 1995-2015, the coal intensity experienced a 28.25% growth. The emergence of this factor indicates that Tianjin could have a good achievement in industrial transformation.

What we should put into consideration is that the commonly used factor energy intensity is not applied in this paper. It could be observed that the models involved energy intensity factor x5 appear in the Pareto front, but these models are not selected as the forecast model. One explanation to this phenomenon is that the energy intensity factor is gradually diminishing with the genetic process going on. It could also make the explanation that these models overfit the data, which could not be involved in the forecast models. From applicable perspective, it could be argued that energy intensity is a derived unit, which reflects the compositive effect of economic and social factors, which could not help to forecast the CO_2 emission.

It should state that the relationship between CO_2 emission and economic indicators is investigated under the complex socio-economic system. It seems impossible to control CO_2 emission only directly with a certain set of economic indicators. The significance of this study lies in the discovery of certain indicators that could be significantly related to the CO_2 emission, which could help to provide appropriate policy implications based on models and certain related indicators.

It should also be clarified that all these influential factors are provided to aid the forecast of CO_2 emission and analysis based on the models. The parameters of the forecast model could vary as the time goes on. In addition, the parameters are also affected by new policy implementation or certain circumstances. The application of symbolic regression for a target problem in a specific domain, the factors and criterion should be carefully integrated with specific knowledge or domain experts. This research further expands the investigation on CO_2 emission of Beijing and Tianjin region, and can be used for reference in the study of carbon emissions in similar regions. For the purpose of extending these results in regions with similar climates, the selection of proper factors and criterion should be selected by detailed analysis or seeking for help from domain experts, and proper analysis should be made based on it.

5. Conclusions and Policy Implications

Instead of analyzing the data and assuming a model with fixed formula or functional pattern, a novel approach –symbolic re-

gression– is conducted. This method is able to automatically find the functional forms and the relationship of factors simultaneously. In this paper, we analyzed six factors which are generally accepted by former researchers to investigate the carbon emission: GDP, total population, the industrial share of GDP, the coal consumption share of TEC, energy intensity and the urban residents' share of total population. The 21 y data during the period 1995-2015 are collected. Based on the experimental results and the analysis, the main conclusions are listed as follows:

(1) There is no universal model which is able to fit both Beijing and Tianjin. The empirical results show that distinct models are constructed based on the data of Beijing and Tianjin. The pattern, parameters and influential factors vary from different regions.

(2) Based on the results returned from the symbolic regression, the influential factors of Beijing and Tianjin expressed definitely different features. It could be argued that this phenomenon is rooted in the different development orientation and the mixture of different natural and industrial environment. In addition, the influential factors are also key points to reduce the CO_2 emission and achieve a balanced development.

(3) CO_2 emission of Beijing region has already peaked in 2010. Under current circumstances, the CO_2 emission of Beijing will gradually diminish. Tianjin could keep a relatively steady growth of CO_2 emission.

Based on the findings in this study, the following policy implications are illustrated to balance the CO_2 emission and economic indicators:

(1) There are different relationships between CO_2 emission and socioeconomic indicators in Beijing and Tianjin. It is urgent for the local government of Beijing and Tianjin to fully understand the relationship and choose appropriate models other than blindly choose a model ahead of time. Furthermore, it seems improper for the region that shows an increasing trend of CO_2 emission to copy the experience of well-controlled region for the variance of different relating factors in different regions.

(2) It can be observed that in well-controlled region like Beijing, the future trend of CO_2 emission shows an inverted U-shaped curve, which means the CO_2 emission is about to fall in the coming years. It is proposed that Beijing should continue to implement sustainable development policies, and take lead in improving the CO_2 reduction. As for Tianjin, the CO_2 emission shows a monotonically increasing trend. It tells us that the CO_2 emission will continue to increase in the future if no change in current policies occurs. It is advised that Tianjin could advance more energy efficiency improvement and conduct proper policies related to energy technology application, especially in population, coal intensity and urbanization.

(3) It could be obviously observed that the indicators of Beijing and Tianjin vary dramatically. This phenomenon could help to give us a guide for the policy implementation of coordinate development of Beijing, Tianjin and Hebei, the province which is surrounding Beijing and Tianjin. For instance, Beijing should reduce its industrial share of GDP and improve the share of tertiary industry by moving the industrial companies to Tianjin and Hebei. It is intended that this policy helps the growth of urbanization rate and improving the technology applied in Hebei, which could help the upgrade of industry in Hebei province, and improve the competitiveness of Beijing as well. As for Tianjin, the unique position and industrial structure make it an important supplement for Beijing. The advantage of opening, new strategic industries and high-end equipment manufacturing make it fully utilize the population and technology to be a development engine.

References

1. Ehrlich PR, Holdren JP. Impact of population growth. *Science* 1971;171:1212-1217.
2. Hubacek K, Feng K, Chen B. Changing lifestyles towards a low carbon economy: An IPAT analysis for China. *Energies* 2012;5:22-31.
3. Waggoner PE, Ausubel JH. A framework for sustainability science: A renovated IPAT identity. *P. Natl. Acad. Sci. USA.* 2002;99:7860-7865.
4. York R, Rosa EA, Dietz T. STIRPAT, IPAT and ImPACT: Analytic tools for unpacking the driving forces of environmental impacts. *Ecol. Econ.* 2003;46:351-365.
5. Dietz T, Rosa EA. Effects of population and affluence on CO_2 emissions. *P. Natl. Acad. Sci. USA.* 1997;94:175-179.
6. Dietz T, Rosa EA. Rethinking the environmental impacts of population, affluence and technology. *Hum. Ecol. Rev.* 1994;1:277-300.
7. Shahbaz M, Loganathan N, Muzaffar AT, Ahmed K, Jabran MA. How urbanization affects CO_2 emissions in Malaysia? The application of STIRPAT model. *Renew. Sust. Energ. Rev.* 2016;57:83-93.
8. Ang B, Zhang F, Choi K-H. Factorizing changes in energy and environmental indicators through decomposition. *Energy* 1998;23:489-495.
9. Wang C, Chen J, Zou J. Decomposition of energy-related CO_2 emission in China: 1957-2000. *Energy* 2005;30:73-83.
10. Valipour M, Banihabib ME, Behbahani SMR. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *J. Hydrol.* 2013;476:433-441.
11. Valipour M. Variations of land use and irrigation for next decades under different scenarios. *Braz. J. Irrigation Drainage* 2016;1:262-288.
12. Valipour M. Analysis of potential evapotranspiration using limited weather data. *Appl. Water Sci.* 2017;7:187-197.
13. Valipour M. How much meteorological information is necessary to achieve reliable accuracy for rainfall estimations? *Agriculture* 2016;6:1-9.
14. Valipour M, Sefidkouhi MAG. Temporal analysis of reference evapotranspiration to detect variation factors. *Int. J. Global Warm.* (in press).
15. Valipour M, Sefidkouhi MAG, Raeini-Sarjaz M. Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events. *Agr. Water Manage.* 2017;180:50-60.
16. Azam M, Khan AQ. Testing the Environmental Kuznets Curve hypothesis: A comparative empirical study for low, lower middle, upper middle and high income countries. *Renew. Sust. Energ. Rev.* 2016;63:556-567.

17. Yang G, Sun T, Wang J, Li X. Modeling the nexus between carbon dioxide emissions and economic growth. *Energ. Policy* 2015;86:104-117.
18. Dinda S. Environmental Kuznets Curve hypothesis: A survey. *Ecol. Econ.* 2004;49:431-455.
19. Kaika D, Zervas E. The Environmental Kuznets Curve (EKC) theory – Part A: Concept, causes and the CO_2 emissions case. *Energ. Policy* 2013;62:1392-1402.
20. Asumadu-Sarkodie S, Owusu PA. Energy use, carbon dioxide emissions, GDP, industrialization, financial development, and population, a causal nexus in Sri Lanka: With a subsequent prediction of energy use using neural network. *Energ. Source. Part B.* 2016;11:889-899.
21. Asumadu-Sarkodie S, Owusu PA. Forecasting Nigeria's energy-use by 2030, an econometric approach. *Energ. Source. Part B.* 2016;11:990-997.
22. Asumadu-Sarkodie S, Owusu PA. The impact of energy, agriculture, macroeconomic and human-induced indicators on environmental pollution: Evidence from Ghana. *Environ. Sci. Pollut. Res*. 2017;24:6622-6633.
23. Asumadu-Sarkodie S, Owusu PA. A multivariate analysis of carbon dioxide emissions, electricity consumption, economic growth, financial development, industrialization and urbanization in Senegal. *Energ. Source. Part B.* 2016;12:77-84.
24. Asumadu-Sarkodie S, Owusu PA. Carbon dioxide emissions, GDP, energy use and population growth: A multivariate and causality analysis for Ghana, 1971-2013. *Environ. Sci. Pollut. Res.* 2016;23:13508-13520.
25. Koza JR. Genetic programming: On the programming of computers by means of natural selection. London: The MIT Press; 1992.
26. Burke EK, Kendall G. Search methodologies: Introductory tutorials in optimization and decision support techniques. 2nd ed. New York: Springer; 2014.
27. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. *Science* 2009;324:81-85.
28. Chattopadhyay I, Kuchina A, Süel GM, Lipson H. Inverse Gillespie for inferring stochastic reaction mechanisms from intermittent samples. *P. Natl. Acad. Sci. USA.* 2013;110: 12990-12995.
29. Lau LS, Choong CK, Eng YK. Investigation of the environmental Kuznets curve for carbon emissions in Malaysia: DO foreign direct investment and trade matter? *Energ. Policy* 2014;68: 490-497.
30. Azlina AA, Law SH, Nik Mustapha NH. Dynamic linkages among transport energy consumption, income and CO_2 emission in Malaysia. *Energ. Policy* 2014;73:598-606.
31. Khu ST, Liong SY, Babovic V, Madsen H, Muttil N. Genetic programming and its application in real-time runoff forecasting. *J. Am. Water Resour. Assoc.* 2001;37:439-451.
32. Bahrami P, Kazemi P, Mahdavi S, Ghobadi H. A novel approach for modeling and optimization of surfactant/polymer flooding based on Genetic Programming evolutionary algorithm. *Fuel* 2016;179:289-298.
33. Paláncz B, Awange J, Völgyesi L. Correction of gravimetric geoid using symbolic regression. *Math Geosci.* 2015;47:867-883.
34. Smits G, Kotanchek M. Pareto-front exploitation in symbolic regression. *Genet. Program. Theory Pract. II.* 2004;8:283-299.
35. 2050 CEACER. 2050 China energy and CO_2 emissions report. Beijing: Science Press; 2009.
36. Ru G, Xiaojing C, Fengting L. 2050 Shanghai energy CO_2 emissions. Shanghai: Tongji University Press; 2011.
37. Xuena W. Study on estimation method of carbon emission to energy carbon sources in China [thesis]. China: Beijing Forestry University; 2006.
38. IPCC. 2006 IPCC guidelines for national greenhouse gas inventories. 2006.
39. Geng Y, Tian M, Zhu Q, Zhang J, Peng C. Quantification of provincial-level carbon emissions from energy consumption in China. *Renew. Sust. Energ. Rev.* 2011;15:3658-3668.
40. Jin Y, Sendhoff B. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Trans. Syst. Man Cybern. C. Appl. Rev.* 2008;38:397-415.
41. Jiang P, Chen J. Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. *Neurocomputing* 2016;198:40-47.
42. Wang Z, Liu X, Zhu YB, Huang R. Prediction on Beijing's, Tianjin's and Hebei's carbon emission. *Geogr. Geo-Inform. Sci.* 2012;28:84-89.
43. Zhang M, Mu H, Ning Y, Song Y. Decomposition of energy-related CO_2 emission over 1991-2006 in China. *Ecol. Econ.* 2009;68:2122-2128.
44. Wang Z, Yang L. Delinking indicators on regional industry development and carbon emissions: Beijing-Tianjin-Hebei economic band case. *Ecol. Indic.* 2015;48:41-48.
45. Birdsall N. Another look at population and global warming. Policy Research Working Papers; no. WPS 1020. Population, health, and nutrition. Washington D.C.: World Bank; c1992. Available from: http://documents.worldbank.org/curated/en/985961468766195689/Another-look-at-population-and-global-warming.