J. Geogr. Sci. 2020, 30(5): 743-756
DOI: https://doi.org/10.1007/s11442-020-1753-4

Identification of the key factors affecting Chinese carbon intensity and their historical trends using random forest algorithm

TANG Zhipeng[1,2], MEI Ziao[1,2], LIU Weidong[1,2], [*]XIA Yan[3]

1. Key Laboratory of Regional Sustainable Development Modeling, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;
2. College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China;
3. Institutes of Science and Development, CAS, Beijing 100190, China

Abstract: The Chinese government ratified the Paris Climate Agreement in 2016. Accordingly, China aims to reduce carbon dioxide emissions per unit of gross domestic product (carbon intensity) to 60%–65% of 2005 levels by 2030. However, since numerous factors influence carbon intensity in China, it is critical to assess their relative importance to determine the most important factors. As traditional methods are inadequate for identifying key factors from a range of factors acting in concert, machine learning was applied in this study. Specifically, random forest algorithm, which is based on decision tree theory, was employed because it is insensitive to multicollinearity, is robust to missing and unbalanced data, and provides reasonable predictive results. We identified the key factors affecting carbon intensity in China using random forest algorithm and analyzed the evolution in the key factors from 1980 to 2017. The dominant factors affecting carbon intensity in China from 1980 to 1991 included the scale and proportion of energy-intensive industry, the proportion of fossil fuel-based energy, and technological progress. The Chinese economy developed rapidly between 1992 and 2007; during this time, the effects of the proportion of service industry, price of fossil fuel, and traditional residential consumption on carbon intensity increased. Subsequently, the Chinese economy entered a period of structural adjustment after the 2008 global financial crisis; during this period, reductions in emissions and the availability of new energy types began to have effects on carbon intensity, and the importance of residential consumption increased. The results suggest that optimizing the energy and industrial structures, promoting technological advancement, increasing green consumption, and reducing emissions are keys to decreasing carbon intensity within China in the future. These approaches will help achieve the goal of reducing carbon intensity to 60%–65% of the 2005 level by 2030.

Keywords: machine learning; random forest; carbon intensity; key factors; China

Received: 2019-12-22 **Accepted:** 2020-02-20
Foundation: National Natural Science Foundation of China, No.41771135
Author: Tang Zhipeng (1978–), PhD and Associate Professor, specialized in economic geography and regional sustainable development. E-mail: tangzp@igsnrr.ac.cn
[*]**Corresponding author:** Xia Yan (1981–), Associate Professor. E-mail: xiayan@casipm.ac.cn

1 Introduction

With the formal signing of the Paris Climate Agreement by the Chinese government in 2016, China has committed to reaching peak carbon emissions around 2030 and reducing carbon dioxide emissions per unit of gross domestic product (GDP), or carbon intensity, to 60%–65% of the 2005 level by 2030. The best way to achieve these goals remains a major issue faced by the Chinese government and an area of concern for Chinese society and the international community. A study on the effects of economic development on carbon intensity in China suggested that improving the energy structure is beneficial for reducing carbon intensity, whereas growing household consumption has the opposite effect (Zhang, 2010). Thus, to achieve the carbon intensity target, the Chinese government should adopt policies that will promote economic reconstruction to significantly reduce energy use (Stern *et al*., 2010; Yuan *et al*., 2012).

The total carbon emissions of a country or region depend on its GDP and carbon intensity. Since China is still a developing country, long-term economic growth is necessary to ensure national prosperity and strength. Therefore, reaching peak carbon emissions by 2030 cannot be achieved by reducing economic growth; instead, it must be accomplished by decreasing carbon intensity. Previous studies on carbon intensity have mainly focused on the influential factors from the perspectives of energy and industrial structure, technological progress, consumption, and land use. The results of previous works suggest that optimizing the energy structure can directly reduce energy intensity, thereby reducing carbon intensity (Wang *et al*., 2013; Peng *et al*., 2016; Li *et al*., 2012; Fan *et al*., 2007). Adjusting the industrial structure to increase the proportion of low-carbon industry is also expected to help reduce carbon intensity (Zhang, 2009; Feng, 2017; Xu and Wang, 2016), as is the use of advanced technologies to improve energy efficiency (Yan *et al*., 2017; Huang and Ding, 2014). The level of household consumption has been shown to be strongly related to carbon emissions (Dong *et al*., 2018; Tong *et al*., 2017; Fan *et al*., 2013). There is no general agreement between different land intensive use types and carbon emission efficiency with certain relationships (Zhu *et al*., 2016; You and Wu, 2014; Zhang *et al*., 2016). Most carbon emissions are generated by fossil energy combustion via human activities. Thus, carbon emissions are directly related to the energy structure and technological progress. GDP refers to the value of all final products and services produced by a country or region in a certain period of time. GDP is directly related to the industrial structure and human consumption. Carbon intensity, which refers to the ratio of carbon emissions to GDP, is affected by numerous factors such as energy structure, industrial structure, technological progress, and human consumption. These factors interact with each other to determine carbon intensity, and they could be expressed by specific factor indicators, for example, the factor of energy structure could be expressed by specific factor indicators such as the consumption proportion of coal, the consumption proportion of gas and so on; thus, a comprehensive index system of the factors affecting carbon intensity is constructed. In general, such an index system should be constructed based on the principles of integrity, hierarchy, and relevance of these indicators. The principle of relevance requires that the interrelated factors should be differentiated to avoid overlap resulting from the duplication of indicators (Zhu *et al*., 2015). Undoubtedly, the different factors are influenced by societal forces and do not have constant effects on carbon intensity; instead, their effects on carbon intensity will vary based on the stage of economic develop-

ment. Therefore, an in-depth analysis of the historical trends in the key factors affecting carbon intensity in China is important to inform future decision-making and meet China's goals for carbon emissions and carbon intensity.

Currently, the main analytical methods applied to assess the factors affecting carbon intensity are path analysis (Wang and Yu, 2013), the panel data model (Peng and Cui, 2016), the spatial econometric model (Feng *et al*., 2017), the adaptive weighted Divisia index method (Greening *et al*., 2001), Laspeyres decomposition model (Ebohon and Ikeme, 2006), and Kaya identity decomposition (Gingrich *et al*., 2011). Path analysis, the panel data model, and the spatial econometric model can be considered as methods of parameter statistical analysis, whereas the adaptive weighted Divisia index method, Laspeyres decomposition model, and Kaya identity decomposition can be categorized as factor decomposition analysis. In parameter statistical analysis, certain statistical assumptions generally need to be satisfied. For example, non-strong collinearity among factors is required to ensure minimum variance in parameter estimation and guarantee that the estimator is effective. In factor decomposition analysis, the need to satisfy the identity relationship tends to weaken the meanings of the decomposed factors or even result in unilateral explanations. Because these traditional methods are limited by classical statistical assumptions, the considered numbers of factors affecting carbon intensity are small. In terms of policy guidance, including only a small number of factors means that all aspects of the economy and society will not be considered in policy development. Global climate change is a complex, cross-cutting natural and societal problem that requires the consideration of a large number of impact factors. However, increasing the number of factors can easily lead to problems of dimensionality during analysis. Machine learning approaches can be used to overcome these dimensionality problems by mining large amounts of data. Machine learning refers to approaches that use computer algorithms to simulate human beings. Machine learning methods evaluate the structure of existing data to then make predictions based on the model of the data (Kohavi and Foster, 1998). Machine learning includes supervised learning, which requires training datasets, and unsupervised learning, which does not. The traditional statistical analysis method relies on the sample data, according to the fixed algorithm flow to calculate the parameters. This type of method does not involve continuous self-learning to improve the operation results. Machine learning methods not only participate in the sample data operating and mining information, but also adapt to the change in optimal strategies through continuous self-learning, and continuously obtain new improved results. Thus, it can be said that machine learning itself is a process of constantly improving the operation results. In supervised learning, integrated learning can be achieved by combining multiple weak supervised models into a more comprehensive and stronger supervised model; this can effectively improve the reliability of the decision-making results. At present, integrated learning remains a popular research area in the field of machine learning. Integrated ensemble learning approaches include bagging, boosting, and random forest algorithms (Chen and Zhu, 2007). Among them, random forest algorithm has the advantages of high classification accuracy, fast operation speed, robust operation results, and strong generalization ability. Thus, random forests have been widely used as classification algorithms. In this study, we applied random forest algorithm to identify the key factors affecting carbon intensity in China and analyzed the historical trends in these key factors to provide a basis for decision-making to reduce carbon intensity in China.

2 Data sources and research methods

2.1 Data sources

This study focused on carbon intensity and its influencing factors from 1961 to 2017 based on data from the Maddison project database, IEA database, China Energy Statistics Yearbook, China Industrial Statistics Yearbook, and China Statistics Yearbook. Based on existing studies and avoiding overlap between factors as much as possible (Zhu *et al*., 2015), we selected 56 factors related to energy structure, industrial structure, technological progress, and household consumption. The selected factors can be classified into nine categories: proportion of fossil energy, price of fossil energy, proportion of non-fossil energy, proportion of new energy, scale or proportion of energy-intensive industry, proportion of service industry, technological progress, traditional consumption of residents, and new consumption of residents (Table 1). Data from China were based on the China Statistical Yearbook, and data from other countries were used as reference data. Some missing data were interpolated.

Table 1 Categorization of factors influencing carbon intensity in China

Category	Factor (unit)
Proportion of fossil energy	Proportion of coal (%) Proportion of oil (%) Proportion of natural gas (%)
Price of fossil energy	Producer price index of coal industry Producer price index of oil and gas industry
Proportion of renewable energy (hydropower and biogas)	Proportion of hydroelectric energy (%) Proportion of biogas energy (%)
Proportion of new energy	Proportion of wind energy (%) Proportion of nuclear energy (%) Proportion of photovoltaic energy (%) Proportion of photothermal energy (%) Proportion of geothermal energy (%) Proportion of liquid biofuel energy (%)
Proportion of service industry	Proportion of financial industry (%) Proportion of information transmission, computer services, and software industry (%) Proportion of education industry (%) Proportion of social welfare in health and social security industry (%) Proportion of culture, sports, and entertainment industry (%) Proportion of science and technology industry (%)
Scale or proportion of energy-intensive industry	Soda ash (10000 tons) Caustic soda (10000 tons) Ethylene (10000 tons) Synthetic ammonia (10000 tons) Cement (10000 tons) Plain glass (10000 weight cases) Crude steel (10000 tons) Finished steel (10000 tons) Proportion of construction industry (%) Proportion of transportation, warehousing, and postal industry (%)
Technological progress	Total labor productivity (10000 yuan per person) Conversion efficiency of electricity generation and heating by power stations (%) Conversion efficiency of coking (%) Conversion efficiency of petroleum refining (%) Standard coal consumption for power generation (g/kWh) Standard coal consumption for power supply (g/kWh) Power plant line loss rate (%) Comprehensive energy consumption per unit of crude steel industry (ton standard coal/ton) Comprehensive energy consumption per unit of cement industry (kg standard coal/ton)

(*To be continued on the next page*)

(*Continued*)

Category	Factor (unit)	Category	Factor (unit)
Traditional consumption of residents	Private cars per 100 urban households	Technological progress	Comprehensive energy consumption per unit of ethylene industry (ton standard coal/ton)
	Motorcycles per 100 urban households		Comprehensive energy consumption per unit of synthetic ammonia industry (ton standard coal/ton)
	Motorcycles per 100 rural households		Proportion of science and technology appropriation to total fiscal expenditure (%)
	Refrigerators per 100 urban households		Energy consumption per unit area of public buildings (kg standard coal/m^2)
	Refrigerators per 100 rural households		Energy consumption per unit area of urban residential buildings (kg standard coal/m^2)
	TV sets per 100 urban households		Energy consumption per unit area of rural residential buildings (kg standard coal/m^2)
	TV sets per 100 rural households	New consumption of residents	Internet broadband users (10000 households)
	Washing machines per 100 urban households		Mileage of high-speed rail (km)
	Washing machine per 100 rural households		Electronic commerce transaction volume (trillions of transactions)

2.2 Random forest algorithm

Random forest is an integrated learning method proposed by Breiman in 2001 based on decision trees. The "randomness" of the random forest is reflected in the training of each tree. A range of the same number of elements of datasets are randomly selected from all training samples to train the algorithm. This data acquisition method is considered as bootstrap sampling. In each branch node variable of building the tree, several subsets of all features are randomly selected to obtain the best segmentation method of subset feature partition by purity calculation (e.g., information gain, information gain rate, and Gini coefficient). The "forest" is reflected in the full growth of each tree without pruning, and the number of trees affects the final decision value.

The size of the forest composed of trees is determined as the number of trees for which increasing the number of trees increases the computational load but does not significantly change the final decision value. Random forests are insensitive to multiple collinearities and robust to missing and unbalanced data; thus, they provide reasonable prediction results (Iverson *et al*., 2008) and are one of the best machine learning algorithms for processing high-dimensional attribute data. The random forest algorithm mainly consists of bootstrap sampling and a classification and regression trees (CART) algorithm.

2.2.1 Bootstrap sampling

In the random forest algorithm, bootstrap sampling is used to extract multiple samples from the original sample. A decision tree is then constructed from each bootstrap sample, and the decision trees are combined to obtain the final result using the voting score rule (Breiman, 2001). Bootstrap sampling (Breiman, 1996) generates a new training sample set by randomly extracting N samples with playback from the training set of original sample size N. Independent bootstrap sample sets are generated by n iterations of independent sampling. As a self-help sampling method, bootstrap sampling works well for small samples. Bootstrap sampling generates a large number of new samples and enlarges the scale of the data by

sampling the initial data with playback and then estimating the overall distribution of the data. Strictly speaking, bootstrap sampling is not a training algorithm; rather, it is a non-parametric method that uses small sample datasets to estimate the entire dataset.

In the bootstrap sampling process, the probability P_i of each sample not being extracted is given by

$$P_i = (1-1/N)^N \, i = 1, 2, \cdots, N. \tag{1}$$

When the value of N is sufficiently large, the probability of each sample not being extracted is given by

$$lim_{N\to\infty}(1-1/N)^N = \underset{(U=-N\to\infty)}{lim}[(1+1/U)^U]^{-1} = e^{(-1)} \approx 0.37 \tag{2}$$

Equation (2) indicates that approximately 37% of the original sample set T does not appear in the bootstrap samples; these data are referred to as out-of-bag (OOB) data. The random forest algorithm generates multiple training sample sets using bootstrap sampling and then constructs multiple classifiers to form a "forest" classifier. Bootstrap sampling adopts random and independent sampling with playback, which can avoid information loss caused by random sampling to a large extent. Bootstrap sampling also overcomes the negative effects of sample class imbalance and improves the reliability of the algorithm.

2.2.2 Decision tree CART algorithm

A decision tree includes a root node, intermediate node, and leaf node, with each node representing the attributes of the object. The path from the root node through the intermediate node to the leaf node represents a decision rule. The generation of a decision tree is usually done recursively starting from the root node. A root node is divided into two subtrees. Then, starting from the subtree, it continues to produce the new root node and again the new root node produce the left and right subtrees. Each of root nodes continues to generate new subtrees recursively until leaf nodes are generated.

Many algorithms exist for generating decision trees, including CLS, ID3, C4.5, and CART node-splitting algorithms (Cao, 2014). In the CLS algorithm, the process of node splitting is random, and the number of attribute fields corresponds to the number of branches. The algorithm does not terminate until the leaf node is generated. This CLS algorithm leads to different results because of selecting different test attributes. The CLS algorithm does not specify which test attributes to use, leading to uncertainty in the algorithm. Thus, the ID3 algorithm obeys the following rule: the maximum information gain of attributes is selected as the test attribute by introducing information entropy to calculate and compare the purity of attributes. To solve the problem of information gain deviation in the ID3 algorithm, the C4.5 algorithm introduces the split information ratio index and calculates the information gain ratio to make the selected attributes more uniform and avoid bias.

The CART algorithm, which was proposed in 1984 (Breiman *et al*., 1984), is also based on information entropy theory and recursively constructs decision trees to generate decision rules, as in the CLS algorithm. The CART algorithm differs from the ID3 and C4.5 algorithms in that the partition is based on the Gini coefficient. As the Gini coefficient decreases, the purity of the attribute partition increases; thus, the partition with the smallest Gini coefficient is selected to construct the decision tree.

The CART algorithm is commonly used to construct decision trees in the random forest algorithm. The steps in the CART algorithm are detailed below (Cao, 2014):

1) Continuous characteristic variables are discretized. There are N continuous features A

for N samples. Arrangement of continuous characteristic variables from small to large is a_1, a_2,⋯, a_N, take median of two adjacent samples and totally take $N-1$. The partition point of i (S_i) is expressed as

$$S_i = (a_i + a_{i+1})/2. \tag{3}$$

2) Each partition's Gini coefficient Gini(P) is calculated. If there are K categories, the probability of occurrence of category i is P_i, and the Gini coefficient is expressed as

$$Gini(P) = \sum_{(i=1)}^{K} P_i(1-P_i) \ = 1 - \sum_{(i=1)}^{K} P_i^2. \tag{4}$$

For a given sample T, suppose there are K categories. The number of categories i is T_i, and the Gini coefficient of sample T is expressed as

$$Gini(P) = 1 - \sum_{(i=1)} (T_i / T)^2. \tag{5}$$

According to value a of characteristic A, T is divided into T_1 and T_2. Under the condition of characteristic A, the Gini coefficient of T is expressed as

$$Gini(T, A) = \frac{|T_1|}{|T|} Gini(T_1) + \frac{|T_2|}{|T|} Gini(T_2). \tag{6}$$

3) For the dataset of current nodes, the nodes are split according to the principle of minimum Gini coefficient, and the decision tree is constructed recursively.

2.2.3 Random forest algorithm

The random forest algorithm is based on bootstrap sampling and the CART algorithm. The process of the random forest algorithm is divided into the following steps (Cao, 2014):

1) Generate training sets. Each tree corresponds to a training set. To construct N decision trees, N training sets need to be generated. Each training set is sampled by bootstrap sampling to generate a subset of training sets. The results are predicted by integrating all subsets of training sets (i.e., bootstrap aggregation, also known as bagging technology). The large differences between the sub-training sets results in high diversity and guarantees robust results.

2) Construct each decision tree. Node splitting is mainly carried out from the root node through the intermediate node and to the leaf node based on the node splitting rule, which maximizes information gain and information gain rate while minimizing the Gini coefficient. Generally, the minimum Gini coefficient in the CART algorithm is used for node splitting. The node splitting process is repeated. This specific process allows to randomly select several attributes from all attributes according to a certain probability distribution to participate in the process of node splitting. As F characteristic variables are randomly extracted, instead of putting all M characteristic variables into the node splitting [$F(F \leqslant M)$], usually is given by $\log_2 M+1$.

3) Form the forest. The above two steps are repeated to build a large number of decision trees, and the random forest is generated. Each tree in the forest is used to classify the samples in the OOB data. One occurrence of the category is denoted as one vote accordingly, the votes for each category are counted, and the category with the largest number of votes is considered the sample category. The proportion of samples not correctly classified in the OOB data is the error rate for the OOB data.

As N increases in the series of decision trees $k(x_1)$, $k(x_2)$,⋯, $k(x_N)$ in a random forest, for

almost all random variable sequences $\theta_1, \theta_2, \cdots, \theta_k$, the generalization error GE^* converges as follows (Gingrich *et al*., 2011):

$$\lim_{N\to\infty} GE^* = P_{xy}\left\{P_\theta\left[k(x,\theta)=Y\right] - \max_{j\neq Y} P_\theta(k(x,\theta)=j) < 0\right\}. \quad (7)$$

This shows that the generalization error GE^* is close to a probability upper limit. Overfitting does not occur as the number of decision trees increases, indicating good robustness.

3 Empirical analysis

3.1 Identification of key factors affecting carbon intensity in China

Obtaining reliable results using machine learning often requires a sufficient number of samples. Thus, in this study, the sample size every 20 years as a dataset from 1961 to 2017, the first dataset of the year 1980 includes the sample from 1961 to 1980, the second dataset of the year 1981 includes the sample from 1962 to 1981 and slide backwards year by year to obtain the annual dataset of 1980–2017. Each dataset has thousands of indicators to ensure a sufficient number of samples. Based on bootstrap sampling, a training set was created for each dataset. Then, based on the CART algorithm, the attribute partition of each training set was created by splitting different nodes. The node splitting continued until the Gini coefficient was minimized [Eq. (6)], ensuring that all partitions were of the highest purity; each partition with the highest purity was a decision tree. By repeating the previous sampling, training, and node partitioning, multiple decision trees were established, and the random forest was generated. The above process was primarily implemented by programming in R software.

In the completed dataset of partitioning attributes of node splitting, the Gini coefficient reflects the purity of the attribute partition in the decision tree, with a smaller Gini coefficient indicating greater purity. When the reduction in Gini coefficient was large, the average purity of all decision trees in the forest increased substantially, indicating that the node variable had a large effect on the forest. Therefore, we calculated the reduction in Gini coefficient for different indicators in each dataset; based on the results, the importance of each factor affecting carbon intensity in China from 1980 to 2017 was evaluated. Because Gini coefficients in random forests are calculated based on information entropy according to the additivity of the entropy value, the sum of the reductions in Gini coefficient for all factors represents the importance of all factors in the index system. Therefore, to uniformly set the quantitative threshold for key factors affecting carbon intensity, we sorted the reductions in Gini coefficient from large to small for each factor from 1980 to 2017 and then calculated the average for each factor from 1980 to 2017 to obtain the corresponding relationships between carbon intensity index number and average reduction in Gini coefficient (Table 2).

If the number of key factors is too large, the significance of identifying key factors will be lost. However, if the number of key factors is too small, it will be difficult to grasp the importance of the entire carbon intensity factor index system. We want proper index numbers to reflect the importance of the entire carbon intensity factor index system. Therefore, according to the 2/3 principle, that is, the set threshold of the number of indicators can cover more than 2/3 of the importance of the entire carbon intensity factor index system, which can be regarded as a key factor. As shown in Figure 1, the cumulative importance of factor increased from 5.7% to 100.0% as the index number increased from 1 to 56.

Table 2 Carbon intensity indicator numbers and corresponding average reductions in Gini coefficient

Carbon intensity index number	1	2	3	4	5	6	7	8
Gini coefficient reduction	0.701	0.613	0.577	0.576	0.572	0.571	0.562	0.560
Carbon intensity index number	9	10	11	12	13	14	15	16
Gini coefficient reductions	0.504	0.462	0.449	0.380	0.376	0.372	0.371	0.365
Carbon intensity index number	17	18	19	20	21	22	23	24
Gini coefficient reductions	0.362	0.341	0.327	0.283	0.241	0.185	0.177	0.176
Carbon intensity index number	25	26	27	28	29	30	31	32
Gini coefficient reductions	0.170	0.165	0.156	0.152	0.151	0.151	0.140	0.103
Carbon intensity index number	33	34	35	36	37	38	39	40
Gini coefficient reductions	0.075	0.069	0.069	0.067	0.067	0.066	0.064	0.050
Carbon intensity index number	41	42	43	44	45	46	47	48
Gini coefficient reductions	0.050	0.049	0.035	0.034	0.034	0.034	0.034	0.033
Carbon intensity index number	49	50	51	52	53	54	55	56
Gini coefficient reductions	0.031	0.029	0.028	0.027	0.026	0.025	0.020	0.011

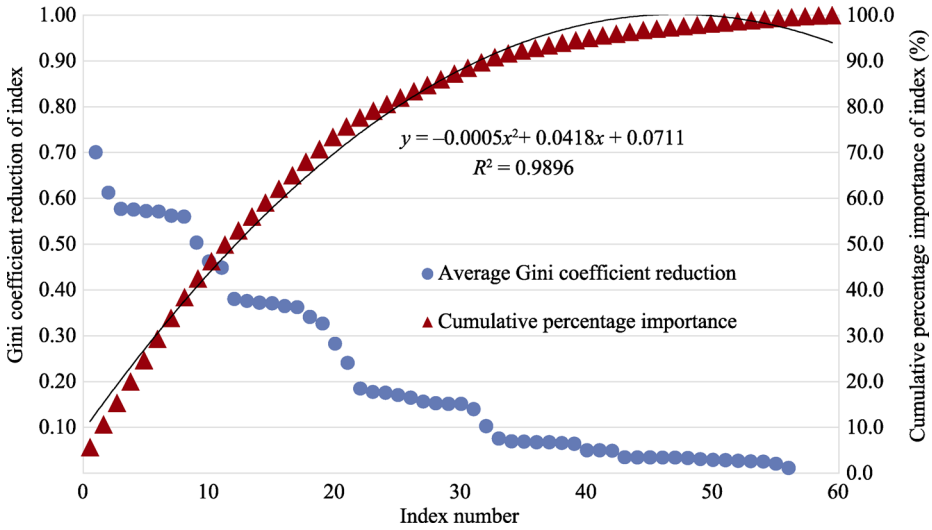

Figure 1 Average reductions in Gini coefficient and the corresponding cumulative percentage importance as a function of carbon intensity index number

Figure 1 shows that the first 22 factors cumulatively accounted for approximately 80% of the importance of the entire carbon intensity factor index system. Less than half of the indicators accounted for more than 2/3 of the importance of the entire index system. Therefore, the number of key factors identified in this study was 22, and the 22 factors with the largest annual reductions in Gini coefficient were identified as the key influencing factors from 1980–2017.

In 1980, the five factors affecting Chinese carbon intensity with the largest reductions in

Gini coefficient were the proportion of natural gas, standard coal consumption for power generation, synthetic ammonia, caustic soda, and comprehensive energy consumption per unit of ethylene industry. In 1981, the top five factors were proportion of natural gas, standard coal consumption for power supply, total labor productivity, comprehensive energy consumption per unit of cement industry, and comprehensive energy consumption per unit of ethylene industry. In 1991, the top five factors were proportion of oil, proportion of coal, motorcycles per 100 urban households, proportion of natural gas, and synthetic ammonia. In 2000, the top five factors were proportion of construction industry, private cars per 100 urban households, comprehensive energy consumption per unit of rough steel industry, power plant line loss rate, and comprehensive energy consumption per unit of ethylene industry. In 2010, the top five factors were proportion of coal, proportion of information transmission, computer services, and software industry, proportion of science and technology appropriation to total fiscal expenditure, proportion of geothermal, and TV sets per 100 rural households. In 2017, the top five factors were proportion of coal, proportion of hydroelectric, conversion efficiency of coking, washing machines per 100 urban households, and energy consumption per unit area of public buildings. The top 22 factors affecting carbon intensity in China were identified for each year from 1980 to 2017 and classified as indicated in Table 1; the number of factors in each category in different years are shown in Table 3.

Table 3 Numbers of key factors affecting Chinese carbon intensity per category by year between 1980 and 2017[1]

Category/Year	1980	...	2000	...	2010	...	2016	2017
Proportion of fossil energy	3	...	0	...	1	...	1	2
Price of fossil energy	0	...	0	...	0	...	0	0
Proportion of renewable energy (hydropower and biogas)	0	...	0	...	0	...	0	1
Proportion of new energy	0	...	0	...	1	...	3	2
Scale or proportion of energy-intensive industry	8	...	7	...	6	...	7	4
Proportion of service industry	0	...	1	...	2	...	2	2
Technological progress	8	...	6	...	4	...	6	5
Traditional consumption of residents	3	...	8	...	6	...	2	4
New consumption of residents	0	...	0	...	2	...	1	2
Total	22	...	22	...	22	...	22	22

3.2 Historical evolution analysis of the key factors affecting carbon intensity in China

Based on the numbers of key factors in different categories from 1980 to 2017 (Table 3), the percentages of key factors contained within each category were calculated for each year (Figure 2). A high percentage for a category indicates that the category had a large effect on carbon intensity in the given year.

According to Figure 2, the key factors affecting carbon intensity in China from 1980–2017 generally presented two characteristics. First, some factors had relatively stable effects on carbon intensity with no obvious changes over time. For example, factors in the

[1] Note: Based on length limitations, Table 3 only lists the statistics for 1980, 2000, 2010, 2016, and 2017; please contact the author for additional data.

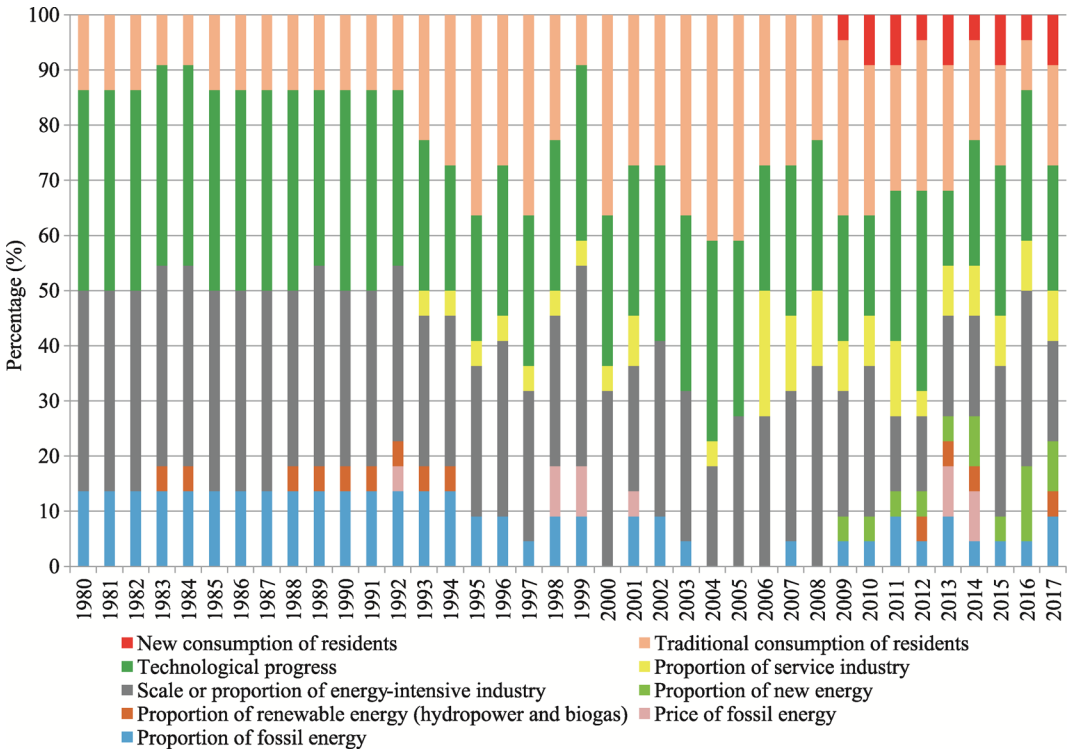

Figure 2 Percentages of factors affecting Chinese carbon intensity in different categories between 1980 and 2017

categories of proportion of fossil energy, the scale or proportion of energy-intensive industry, technological progress, and traditional consumption of residents have always had important effects on carbon intensity. When considering the effects of resident consumption on carbon intensity, no significant difference was observed between rural and urban residents in terms of the consumption of household appliances and household transportation-related consumption. This indicates that green consumption can be practiced by all residents and applies to household energy consumption. And green consumption needs to maintain consistency and perseverance. Second, other factors affecting carbon intensity in China showed obvious temporal characteristics. For example, the proportion of renewable energy (particularly the proportion of hydroelectric energy) had a greater effect on carbon intensity early in the study period. Since 2009, the effects of the proportion of new energy, specifically the proportions of geothermal and photothermal energies, on carbon intensity have increased. In terms of industrial structure, factors in the proportion of service industry category had no obvious effects on carbon intensity during the early years in the study period. Since 1993, the effect of the proportion of financial industry on carbon intensity has increased. Since 2006, the effect of the proportion of information transmission, computer services, and software industry on carbon intensity has remained significant, and the effect of the proportion of scientific research, education, culture, and entertainment industry has increased. In the new consumption of residents category, the effect of Internet broadband users on carbon intensity has increased since 2009, while the effect of electronic commerce transaction volume has increased significantly since 2010. Thus, the variations in the effects of these factors with time should be considered when developing emission reduction measures.

Overall, the historical evolution of key factors affecting carbon intensity can be roughly

divided into three stages:

1) From the early 1980s to 1991, the key factors influencing carbon intensity were in the categories of scale or proportion of energy-intensive industry, proportion of fossil energy, technological progress, and traditional consumption of residents. During this period, the level of economic development in China was low, and extensive development was occurring. The economic structure was dominated by energy-intensive industries, and energy consumption was dominated by coal. Production factors such as standard coal consumption in power generation had a strong effect on carbon intensity during the early stage of Chinese economic development.

2) From 1992 to 2007, China underwent a period of opening up, and the rate of economic growth was high. This was accompanied by an increase in the incomes of Chinese residents along with accelerated consumption of household appliances and transportation resources. China's entry into the World Trade Organization in 2001 further expanded its opening up to the outside world, and the proportion of service industry within the Chinese economy increased steadily with the development of electronic information technology. In 1993, China changed from an oil exporter to a net oil importer for the first time, and demand for energy increased. From 1992 to 2007, the effects of scale or proportion of energy-intensive industry, proportion of fossil energy, technological progress, and traditional consumption of residents on carbon intensity increased significantly, and the proportion of service industry and price of fossil energy began to have important effects. Overall, from 1992 to 2007, the key factors affecting carbon intensity changed from production factors to production and consumption factors; thus, both production and consumption factors should be considered in efforts to reduce carbon intensity.

3) After the global financial crisis in 2008, China began to adjust its economic structure and implement measures to save energy and reduce emissions. The Chinese government supported new energy types and invested in the research and development of various energy-saving and emission-reducing technologies. Meanwhile, the Internet became universally integrated into the lives of Chinese residents, high-speed rail travel greatly shortened commute times, and the prevalence of e-commerce rapidly increased. From 2008 to 2017, the effects of emerging industries and consumption on carbon intensity increased significantly, and the effects of the scale or proportion of energy-intensive industry, proportion of fossil energy, technological progress, traditional consumption of residents, proportion of service industry, price of fossil energy, proportion of new energy, new consumption of residents, and proportion of science and technology appropriation to total fiscal expenditure emerged. This indicates that in addition to the traditional consumption of residents, the effects of new consumption of residents on carbon intensity are also important and should be considered in the development of energy-saving and emission-reduction policies in the future.

4 Conclusions and implications

Global climate change is a complex issue at the intersection of nature and society. Developing a low-carbon economy is the only way to deal with global climate change. As a responsible country, China has promised to reduce carbon intensity to 60%–65% of the 2005 level by 2030. To achieve this goal, it is first necessary to identify the key factors affecting carbon intensity. These key influential factors can then be used to reduce carbon intensity by implementing effective policy measures.

In view of the shortcomings of traditional quantitative methods, this study used a machine learning algorithm to identify the key factors affecting carbon intensity in China and then

analyzed their historical trends. The main conclusions are summarized as follows:

(1) The key factors affecting carbon intensity in China from 1980 to 2017 were identified using random forest algorithm. Evolutionary analysis showed that the key factors and their effects changed over time. Therefore, policies designed to save energy and reduce emissions in China should also be adjusted over time.

The key factors affecting carbon intensity from 1980 to 1991 were mainly the scale or proportion of energy-intensive industry, proportion of fossil energy, and technological progress. After Deng Xiaoping's Speech on the Southern Tour, the Chinese economy entered a period of rapid growth from 1992 to 2007, the proportion of service industry and price of fossil energy began to affect carbon intensity, and the effects of traditional consumption of residents increased. After the Global Financial Crisis in 2008, China entered a period of economic restructuring, and policies to save energy and reduce emissions were enacted. The proportion of new energy and new consumption of residents also began to affect carbon intensity during this time.

Overall, reducing the proportion of fossil energy, reducing the scale and proportion of energy-intensive industry, and promoting technological progress are the main measures that should be taken to reduce emissions. At the same time, China should vigorously develop the service industry, optimize the industrial structure, reduce the use of fossil energy by increasing the price of fossil energies via taxation, and promote the development of new energy resources. Green consumption, which includes the consumption of traditional household appliances, household transportation, and Internet e-commerce, is also important for reducing carbon intensity.

(2) With the widespread attention to the issue of global climate change, economic growth must be balanced with environmentally sustainable development. Carbon intensity is closely related to many aspects of social production and the lives of residents. The key factors affecting carbon intensity are different at different stages of economic development, as demonstrated in the paper.

Traditional statistical analysis cannot be used to assess the many factors affecting carbon intensity because of the need to overcome multicollinearity. Meanwhile, factor decomposition analysis tends to weaken the meaning of the decomposed factors or even explain them unilaterally. In contrast, machine learning has innate advantages when dealing with large datasets. Among machine learning approaches, random forest algorithm has good robustness and generalizability. However, the change of the international geopolitics circumstances along with the emergence of new technologies and industries create uncertainties for future socioeconomic development, and the applicability of stochastic forest algorithm for identifying the key factors affecting carbon intensity in uncertain scenarios requires further analysis.

Nevertheless, considering the rapid socioeconomic development in China, the factors affecting Chinese carbon intensity in the future will likely be closely related to the historical trends. To achieve China's carbon intensity target by 2030, strategic planning and the adoption of strong policy measures are necessary. Therefore, understanding the historical evolution of the key factors affecting carbon intensity in China is important for formulating policy for the future.

References

Breiman L, 1996. Bagging predictors. *Machine Learning*, 24(2): 123–140.
Breiman L, 2001. Random forests. *Machine Learning*, 45(1): 5–32.

Breiman L, Friedman J, Stone C J *et al*., 1984. Classification and Regression Trees. UK: Chapman & Hall/CRC.
Cao Z F, 2014. Study on optimization of random forests algorithm [D]. Beijing: Capital University of Economics and Business. (in Chinese)
Chen K, Zhu Y, 2007. A summary of machine learning and related algorithms. *Statistics and Information Forum*, 22(5): 105–112. (in Chinese)
Dong M, Xu Z Y, Li C F, 2018. The impact of carbon intensity restriction on welfare of urban and rural residents: An analysis based on CGE model. *China Population, Resources and Environment*, 28(2): 94–105. (in Chinese)
Ebohon O J, Ikeme A J, 2006. Decomposition analysis of CO_2 emission intensity between oil-producing and non-oil-producing sub-Saharan African countries. *Energy Policy*, 34(18): 3599–3611.
Fan J, Liao H, Liang Q M *et al*., 2013. Residential carbon emission evolutions in urban-rural divided China: An end-use and behavior analysis. *Applied Energy*, 101: 323–332.
Fan Y, Liu L C, Wu G, *et al*., 2007. Changes in carbon intensity in China: Empirical findings from 1980–2003. *Ecological Economics*, 62(3/4): 683–691.
Feng Y, Zhu L Y, Zhang D H, 2017. Spatial and econometric analysis of effect of industrial structure adjustment on carbon intensity in China. *Soft Science*, 31(7): 11–15. (in Chinese)
Gingrich S, Kušková P, Steinberger J K, 2011. Long-term changes in CO_2 emissions in Austria and Czechoslovakia: Identifying the drivers of environmental pressures. *Energy Policy*, 39(2): 535–543.
Greening L A, Ting M, Krackler T J, 2001. Effects of changes in residential end-uses and behavior on aggregate carbon intensity: Comparison of 10 OECD countries for the period 1970 through 1993. *Energy Economics*, 23(2): 153–178.
Huang J, Ding G, 2014. A research on the threshold effect of the impact of technical progress on carbon intensity. *Science & Technology Progress and Policy*, 31(18): 22–26. (in Chinese)
Iverson L R, Prasad A M, Matthews S N *et al*., 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management*, 254(3): 390–406.
Kohavi R, Foster P, 1998. Special issue on application of machine learning and the knowledge discovery process. *Journal of Machine learning*, 30(2): 271–274.
Li H, Wang L, Shen L *et al*., 2012. Study of the potential of low carbon energy development and its contribution to realize the reduction target of carbon intensity in China. *Energy Policy*, 41: 393–401.
Peng X, Cui H R, 2016. Research on the effects of energy structure adjustment in China on carbon intensity. *Journal of Dalian University of Technology* (*Social Sciences*), 37(1): 11–16. (in Chinese)
Stern D, Jotzo F, 2010. How ambitious are China and India's emissions intensity targets? *Energy Policy*, 38(11): 6776–6783.
Tong J P, Chen G D, Yang Z Y *et al*., 2017. Threshold effects of household consumption level on residential carbon emissions. *Journal of Arid Land Resources and Environment*, 31(1): 38–43. (in Chinese)
Voigt S, Cian E, Schymura M *et al*., 2014. Energy intensity developments in 40 major economies: Structural change or technology improvement? *Energy Economics*, 41: 47–62.
Wang S H, Yu W Y, 2013. Sensitivity analysis of primary energy consumption structural change and carbon intensity. *Resources Science*, 35(7): 1438–1446. (in Chinese)
Xu H, Wang H, 2016. The impact of industrial structure adjustment on China's carbon intensity goal: The outlook of 2020. *Science and Technology Management Research*, 36(13): 232–236. (in Chinese)
Yan Z M, Deng X L, Yang Z M, 2017. The impact of heterogeneous technological innovation on carbon intensity: A global evidence based on patent statistics. *Journal of Beijing Institute of Technology (Social Sciences Edition)*, 19(1): 20–27. (in Chinese)
You H Y, Wu C F, 2014. Analysis of carbon emission efficiency and optimization of low carbon for agricultural land intensive use. *Transactions of the Chinese Society of Agricultural Engineering*, 30(2): 224–234. (in Chinese)
Yuan J, Hou Y, Xu M, 2012. China's 2020 carbon intensity target: Consistency, implementations, and policy implications. *Renewable and Sustainable Energy Reviews*, 16(7): 4970–4981.
Zhang M, Gan C L, Chen Y R *et al*., 2016. Carbon emission efficiency and optimization of low carbon for construction land development intensity in China according to provincial panel data. *Resources Science*, 38(2): 265–275. (in Chinese)
Zhang Y G, 2009. Structural decomposition analysis of sources of decarbonizing economic development in China: 1992–2006. *Ecological Economics*, 68(8/9): 2399–2405.
Zhang Y G, 2010. Economic development pattern change impact on China's carbon intensity. *Economic Research Journal*, (4): 120–133. (in Chinese)
Zhu F F, Zhou L, Mo L P *et al*., 2015. A scientific study of establishing statistical index system. *Statistical Theory and Practice*, (11): 8–12. (in Chinese)
Zhu Z Y, Miao J J, Cui W, 2016. Analysis of carbon emission efficiency for urban construction land intensive use. *Areal Research and Development*, 35(3): 98–103. (in Chinese)