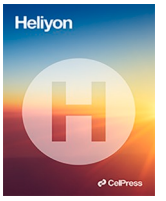
Heliyon 10 (2024) e29086

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Understanding China's CO_2 emission drivers: Insights from random forest analysis and remote sensing data

Qingsheng Lei [a], Hongwei Yu [b], Zixiang Lin [b,*]

[a] State Grid Hubei Electric Power Co., Ltd. Economic and Technological Research Institute, Wuhan, 430077, PR China
[b] Institute of Quality Development Strategy, Wuhan University, Wuhan, 430072, PR China

ARTICLE INFO

Keywords:
CO_2 emissions
Random forest
Environmental impact
Emission reduction strategies

ABSTRACT

China has become the world's largest emitter of carbon dioxide, putting significant pressure on the government to reduce emissions. This study analyzes the driving factors of carbon emissions in 281 prefecture-level cities in China from 2003 to 2019, based on carbon emission data matched with the locations of thermal power stations and nighttime light data. Firstly, we compare the accuracy of multivariate linear regression and random forest models, finding that the random forest regression yields superior results. Then, we rank the impact of various factors using the random forest method, revealing that population, economic development, and industrialization are the top three influencing factors. The interaction between population and economic development explains 68.5% of carbon emissions, with regional variations in the ranking of influencing factors. The main policy implications of this study are as follows: firstly, there is no need to overly concern about the impact of population growth on carbon emissions, and policies regarding fertility can be adjusted flexibly; secondly, controlling urbanization to a certain extent is conducive to achieving efficient low-carbon cities; thirdly, during the process of industrialization, carbon emissions inevitably increase, and it is advisable to accelerate industrialization to reach a turning point as soon as possible.

1. Introduction

The Paris Agreement outlines the target of "restricting the increase in the worldwide mean temperature to less than two degrees higher than pre-industrial levels, with endeavors to confine it to 1.5 °C above pre-industrial levels [1]". As a responsible developing country, China promptly established the aim of "reaching the apex in carbon releases by 2030 and accomplishing carbon neutrality by 2060." In 2021, China's greenhouse gas discharges constituted about one-third of the total global greenhouse gas emissions, and the transformation from the peak of carbon emissions to carbon neutrality spans just 30 years, which is only half the time frame seen in Western developed nations [2,3]. China faces considerable pressure to reduce greenhouse gas emissions. Since 2020, the Chinese administration has introduced a series of "1+N" strategies for curtailing carbon emissions, and local authorities have devised their own strategies for reducing emissions [4]. A comprehensive grasp of the motivating elements influencing regional greenhouse gas emissions can supply beneficial insights for the government to formulate more precise policies and further the realization of China's objectives for reducing greenhouse gas emissions [5].

Prior research on regional carbon emissions has primarily concentrated on factors such as the level of economic development [6],

* Corresponding author.
E-mail addresses: leiqs4@hb.sgcc.com.cn (Q. Lei), hongwei_yu@whu.edu.cn (H. Yu), 2015301750052@whu.edu.cn (Z. Lin).

https://doi.org/10.1016/j.heliyon.2024.e29086
Received 27 December 2023; Received in revised form 13 March 2024; Accepted 29 March 2024
Available online 4 April 2024

population size [7], energy intensity [8], industrial composition [9–11], degree of urbanization [12,13], foreign investment [14,15], environmental regulations [16,17], and others. However, these studies have not reached a consensus regarding the influence of these factors. Some studies assert that economic scale is the most crucial factor. They contend that economic scale currently plays the most significant role in driving CO_2 emissions and that this driving impact diminishes with economic development [18]. On the contrary, some research posits that population size is the most vital factor with a negative influence [19]. These studies argue that the growth in energy efficiency stemming from population concentration far surpasses the increase in carbon emissions resulting from population expansion [20]. Conversely, other pieces of literature have reached opposing conclusions [21]. Some researchers contend that the level of urbanization surpasses population size and per capita GDP in importance [22]. They propose that variations in energy types and consumption patterns between urban and rural residents lead to an upsurge in carbon emissions during the urbanization process [23,24]. The lack of consensus in existing research findings concerning the most significant factors behind regional carbon emissions can obscure the primary drivers, making it challenging for the government to devise targeted carbon reduction policies. This lack of clarity hinders the realization of carbon neutrality goals.

We believe there may be two reasons for the inconsistent conclusions of existing studies:

The first is the objectivity of carbon emission data. Objective carbon emission data is the basis for studying factors affecting carbon emissions. Currently, the primary carbon emission data are mainly calculated according to the methodology of IPCC [25]. Although the MRV system can guarantee the reliability of carbon emission data to a certain extent, it is based on the subjective report of energy consumption by enterprises in essence [26]. To reduce carbon reduction expenditure, enterprises may falsify energy consumption and even collude with third-party verification institutions to issue false carbon emission reports to reduce carbon reduction expenditure. In March 2022, there was an incident of falsification of carbon emission report data by third-party verification institutions in China [27]. On the other hand, local governments may also relax the verification of data reported by enterprises due to the pressure of economic development or the lack of verification ability. Similar situations often occur in reporting environmental monitoring data [28]. From the technical point of view, the deviation of carbon emission data calculated is within 5%, which is unavoidable [29]. In addition, in the actual statistics, the measurement method is based on the measured primary data of emission sources. In contrast, the emission factor method uses the coefficient of different fuels for estimation. It is not easy to obtain consistent data from the data obtained by the two methods. Currently, some research uses space-based nighttime light data to estimate the spatial extent of fossil fuel CO_2 emissions; these studies are often based on nighttime light data directly to estimate [30,31]. Although this kind of method ruled out carbon emissions data as the possibility of subjective tampered with, continuous optimization algorithm on the data fit achieved good effect. However, the night light data is highly relevant to economic activities [32]. It is impossible to exclude the correlation between influencing factors and light when only relying on the data of carbon emission fitting obtained by night light, and the conclusion is unconvincing. Therefore, it is necessary to seek more objective carbon emission data.

The second is related to methods of identifying key influencing factors of carbon emissions. Currently, many related studies use methods based on econometric regression or factor decomposition analysis [33]. The method of econometric regression assumes that the influencing factors are independent of each other, which is effective in analyzing a single element but challenging to solve the problem of multicollinearity in the face of many factors. However, the method of factor decomposition needs to comply with the preset theoretical framework, and decomposed factors must follow the specific theoretical logic, which may lead to a lack of inclusiveness in reality [34]. For example, in the Kaya identity used by LMDI, energy consumption and economic aggregate do not correspond in practice, and the economic implications of driving factors after model expansion are unclear [35]. The differences and shortcomings of these methods may also lead to inconsistent research conclusions, making the analysis results poor policy reference. Hence, ranking the significance of influencing factors proves challenging with the existing two methods, emphasizing the need for a more inclusive approach.

To address the shortcomings of existing research in terms of data objectivity and identification methods, this study utilizes the Open-source Data Inventory for Anthropogenic CO_2 (ODIAC) database [36,37]. This database integrates spatial nighttime light data with individual power plant emission/location profiles to estimate the global spatial distribution of fossil fuel CO_2 emissions. It is well known that nighttime lights are highly correlated with human economic activity. However, most existing studies overly rely on the relationship between nighttime lights and carbon emissions, resulting in data that lack convincing power. To overcome these limitations, the database used in this study not only relies on nighttime lights but also considers the distribution of thermal power plants. Thermal power plants are typically located in suburban areas due to factors such as land constraints and pollution, which are often distant from the actual economic centers. Therefore, including thermal power plants in the analysis can partially overcome the drawbacks of existing studies' over-reliance on the relationship between nighttime lights and carbon emissions. Methodologically, this study employs the random forest algorithm in machine learning. Machine learning algorithms have advantages such as not requiring prior assumptions and being able to handle massive amounts of data more finely. They can break away from the many assumptions of traditional statistical methods used in data processing and analysis. In particular, random forests, among many machine learning algorithms, introduce randomness, making overfitting less likely, are adept at handling high-dimensional data, and can rank the importance of various influencing factors. Compared to many existing methods, they provide a more inclusive perspective on the magnitude of influencing factors.

This study, for the first time, utilizes more convincing nighttime light-retrieved carbon emission data. Within the scope of obtainable data, it investigates the influencing factors of carbon emissions using the random forest algorithm in 281 prefecture-level cities in China. The main contributions of this research are as follows: firstly, it identifies the inadequacies in data objectivity in existing studies, most of which use data obtained from potentially subjective "bottom-up" or insufficiently convincing "top-down" approaches, providing a more objective basis for future research. Secondly, the study employs the random forest method, which effectively addresses the limitations of traditional statistical methods when faced with high-dimensional data, providing a new approach for future

research. Finally, the study analyzes the causal factors of carbon emissions, ranks their impact, and discusses their regional differences, potentially providing a basis for local governments to formulate targeted carbon reduction policies.

2. Methods

2.1. Random forest algorithm

The principal objective of this research is to investigate the determinants of greenhouse gas emissions at the provincial level and evaluate the magnitude of each determinant. Concerning the data utilized, the dataset exhibits relatively high dimensions and has some minor gaps due to data availability. Regarding the response curve, the factors influencing greenhouse gas emissions are influenced by numerous variables. The relationship is intricate, and their interconnection is often not a straightforward linear one. In terms of functions to be implemented, this paper needs to rank the impact of each factor. Random forest may be a good choice, effectively meeting the above three requirements. Firstly, the traditional standard linear models require the data to meet certain assumptions or conditions, such as the linear relationship between independent and independent variables [38]. The random forest based on machine learning does not require a priori assumptions and builds the model based on the characteristics of the data. It can deal with massive and high-dimensional data more finely and improve prediction accuracy without significantly increasing the computational burden. Moreover, it is insensitive to multivariate collinearity, and the calculation results on the data with missing or unbalanced data can pass the robustness test well [39].

Secondly, the partial dependency plot produced by the random forest illustrates the marginal impact of a predictor variable in the machine learning model on the forecast outcomes of the previously established model, providing a method for interpreting the results of machine learning. The partial dependency function is obtained by fixing a variable and calculating the average value of all combinations of prediction functions of other variables [40]. We can understand the complex corresponding relationship between a single variable and the explanatory variable through the partial dependency plot.

Finally, rank the importance of independent variables under the two measurement methods of the random forest model. IncMSE refers to the increase of the error of random forest model estimation after the random value of the variable compared with the original error. While IncNodePurity refers to the importance degree of the variable to each decision tree node [41]. The larger the value, the more influential the variable is. The random forest model can screen the features of multiple variables through these two indicators and then give the importance ranking of variables, which effectively meets the requirements of this study.

The principle of random forest used in this paper is as follows:

In terms of the method itself, RandomForest (RF) is a machine learning algorithm based on a classification regression tree, built on multiple decision trees, a typical ensemble learning model. Its principle is often the bootstrap and back on the sampling in the original samples for multiple random samples to form new training samples, through these self-help samples to construct the corresponding decision tree. Furthermore, the combination of all the decision tree building is a “forest.” The accuracy of the calculation model and through many times to the average accuracy of the assessment model. The average value of multiple trees is the final result by voting. The undrawn data will form an Out of Bagging (OOB) dataset, which will be used as test samples to evaluate the accuracy of the random forest model and the importance of each feature value. The implementation process of random forest in this paper is shown in Supplementary Material S1.

The random forest model is versatile, capable of handling both classification and regression tasks, depending on the nature of the data [42]. A forest primarily composed of classification trees is known as a random forest classifier, whereas a forest predominantly composed of regression trees is referred to as a random forest regressor. When addressing classification problems, the classification error rate is assessed, while for regression problems, the variance of the residuals is examined [43]. In this study, we utilize a random forest regression model to investigate the key determinants of greenhouse gas emissions. Utilizing a dataset comprised of 4125 matched samples, we construct a random forest model (depicted in S1), with the determinants as independent variables and greenhouse gas emissions as the dependent variables. The dataset is subsequently divided into a training set and a test set in a 3:1 ratio. The training set is used to train the random forest model, and the model’s accuracy on the test set is employed to assess the impact of determinants on greenhouse gas emissions.

Table 1
Description of variables used in the analysis for the period 2003–2019.

Variables	Symbol	Definition measuring method	Unit of measurement
CO_2 emission	I	CO_2 emissions	10^6 Ton
Population size	POP	Registered population	Thousand people
Energy intensity	EI	Energy consumption to divided by GDP	Tons of standard coal per 10^6 Yuan
Economic level	AF	GDP per capita	Yuan
Industry proportion	ID	The percentage of the second industry in the total GDP	%
Urbanization level	UR	The percentage of the urban population in the total population	%
Environmental regulations	ER	Provincial environmental regulations	Item
Foreign investment	FDI	Amount of foreign direct investment	10^6 Yuan

2.2. Explanation of variables and data sources

Carbon emissions are influenced by a multitude of factors, as detailed in the reference literature. This study observes that, based on a comprehensive review of existing research, seven primary directions emerge when selecting these factors. Our selection includes population size, energy intensity, economic development level, industrialization level, urbanization, and foreign investment. These seven factors are frequently the focal points in existing studies.

Table 1 provides details about the variables employed in this study, in addition to their respective explanations. The data related to the size of the population, the level of economic development, the industrial composition, the level of urbanization, and foreign direct investment is sourced from the China City Statistical Yearbook. Information concerning energy intensity is derived from the China Energy Statistical Yearbook within the framework of this research. It is worth noting that data on carbon emissions for Chinese cities are not directly measured and are not included in the statistical yearbook.

A majority of the greenhouse gas emission data for provincial-level cities utilized in previous research is approximated via remote sensing techniques relying on nighttime illumination data and refined through iterative algorithm enhancements. This method, which places significant reliance on the correlation between nighttime illumination and economic activity, often yields inconclusive results. In our study, the data regarding the emission of greenhouse gases is derived from the ODIAC database [37], which is a high-resolution global gridded dataset of monthly emissions of carbon dioxide resulting from the burning of fossil fuels. This dataset provides a spatial resolution of 1000 m and is the first to incorporate spatially-based nighttime illumination data along with individual power plant emission profiles and location specifics to approximate the global spatial distribution of CO_2 emissions originating from sources of fossil fuels. This approach somewhat mitigates the over-dependence on nighttime illumination fitting. When compared to similar databases like EDGAR, ODIAC data is more current, with the latest data available up to 2019. It also boasts higher resolution and is more user-friendly for data processing. The precision of greenhouse gas emission data specifically extracted from subnational areas (e.g., provincial, municipal, or county-level) heavily relies on the accuracy of illumination data and power plant distribution data. Despite this limitation, this dataset remains one of the primary sources of greenhouse gas emission information for secondary administrative regions and plays a crucial role as a reference dataset for calculating greenhouse gas emissions in such areas. Due to changes in administrative divisions, the number of prefecture-level cities in China varies annually. The sample period of this study spans from 2003 to 2019. After matching carbon emission data with statistical yearbook data, a total of 281 prefecture-level cities with relatively complete data were obtained.

3. Results

3.1. Applicability evaluation of random forest

To compare the goodness of fit between random forest and linear regression, we initially conducted a linear regression analysis on carbon emissions and the seven influencing factors, presenting the findings in Table 2. A noteworthy observation from Table 2 is the varying correlation coefficients between China's carbon emissions and each influencing factor. Significantly, six out of the seven factors selected in this paper exhibit substantial correlation. Computation of the standardized correlation coefficient reveals a positive and strong correlation between carbon emissions and population, economic development level, secondary industry proportion, urbanization level, environmental regulation, and foreign investment. Notably, there is no statistically significant correlation between carbon emissions and energy intensity. Following the correlation analysis, we selected the factors that passed the significance test as independent variables for multiple linear regression, and the equation is as follows:

$$I = 0.477POP + 0.424AF + 0.069ID + 0.041UR + 0.034ER + 0.209FDI \quad (\mathrm{n} = 4215, \mathrm{R}^2 = 0.63, \mathrm{P} < 0.01) \tag{1}$$

In the equation, due to the standardized regression coefficient, there is no constant term; I represent carbon dioxide emissions, POP, AF, ID, UR, ER, and FDI are population size, economic development level, industrialization level, urbanization level, environmental regulation, and foreign investment, respectively. Equation (1) shows that population size, economic development level, urbanization level, environmental regulation, and foreign investment are positively correlated with carbon emissions. Among them, the standardized regression coefficient of population size is the largest, indicating that it has the most significant impact on carbon emissions, followed by the level of economic development and foreign investment.

Table 2
Correlation coefficient and standard regression coefficient between CO_2 emissions and each influencing factor.

Variables	Symbol	Correlation coefficient	standard regression coefficient
Population size	POP	0.004	0.479**
Energy intensity	EI	0.011	0.010
Economic level	AF	0.000	0.425**
Industry proportion	ID	11.452	0.067**
Urbanization level	UR	5.801	0.041**
Environmental regulations	ER	0.011	0.035**
Foreign investment	FDI	0.004	0.209**

Notes: ** indicates statistical significance at the 1% level.

All the influencing factors were utilized as independent variables to perform random forest regression on carbon emissions. We computed the coefficient of determination (R^2) and Root Mean Squared Error (RMSE) between the fitted values obtained by the two regression methods and the measured carbon emissions. As illustrated in Fig. 1 (a, b), the R^2 for multiple linear regression is 0.63, which is notably lower than the R^2 of random forest ($R^2 = 0.86$). Additionally, the RMSE for multiple linear regression exceeds that of the random forest method, indicating that the random forest's predicted values exhibit smaller deviation errors compared to multiple linear regression. Furthermore, the predicted values of the multiple linear regression include negative values, which contradict the actual significance of carbon emissions. In contrast, there are no negative values in the fitted values produced by the random forest. Consequently, the random forest method boasts a higher R^2 than multiple linear regression, offering greater practical significance and improved predictive accuracy. The random forest approach is well-suited for **analyzing the influencing factors of carbon emissions.**

3.2. *Response of influencing factors*

In this study, we utilize feature importance ranking and partial dependence diagrams to enhance the interpretability of the random forest model. As shown in Fig. 2, we present the importance ranking of variables obtained through the IncMSE and IncNodePurity methods. The results from both importance ranking methods exhibit similarities: Population, Economic level, Industry proportion, and Foreign investment are identified as more influential than Urbanization level, Energy intensity, and Environmental regulations. In the IncMSE method, Population emerges as the most critical factor, whereas the IncNodePurity method places Foreign investment at the forefront, with Environmental regulations being the least influential. The factors with the highest importance order identified by the IncMSE method tend to align with the multiple linear regression standardized coefficients, although some variations exist.

The partial dependence plots provide insights into the relationship between CO_2 emissions and influencing factors. On the whole, Population, Economic level, Industry proportion, Foreign investment, Urbanization level, and Environmental regulations exhibit a positive correlation with CO_2 emissions. Meanwhile, the impact of Energy intensity on CO_2 emissions demonstrates significant variability across different stages.

The population is recognized as the most significant factor. The partial dependence plot (Fig. 3f) illustrates that when the population experiences growth, greenhouse gas emissions also exhibit an upward trend. The impact of population is most notable when it exceeds a population of ten million, after which it stabilizes. This suggests that during the development of a city, an initial surge in population leads to a substantial rise in greenhouse gas emissions. However, as the city matures and the industrial structure improves, subsequent population growth has a diminished effect on greenhouse gas emissions.

Concerning the level of economic development (Fig. 3a), there are four distinct phases in the correlation between carbon emissions and per capita GDP. In the early phases of income growth, there is a significant upsurge in carbon emissions. When per capita GDP surpasses a specific threshold, the pace of carbon emissions growth decelerates. However, when per capita GDP surpasses another threshold, carbon emissions rise rapidly again before stabilizing after a particular GDP level.

The industrialization level (Fig. 3e) exhibits a complex pattern with phases of decline, increase, and stabilization. When the industrialization level is below 0.4, carbon emissions generally decrease. Between an industrialization level of 0.4 and 0.7, carbon emissions increase rapidly with industrialization improvement. Once the industrialization level exceeds 0.7, carbon emissions decline gradually. The early stages of industrialization are marked by improved production modes and increased efficiency, resulting in reduced emissions from the primary production mode. As industrialization advances, the expansion of industrial scale leads to a sharp increase in carbon emissions. Upon reaching full industrialization, incremental carbon emissions from industrial scale expansion decrease, while the reduction of carbon emissions from improved production efficiency takes precedence, causing a gradual decline in

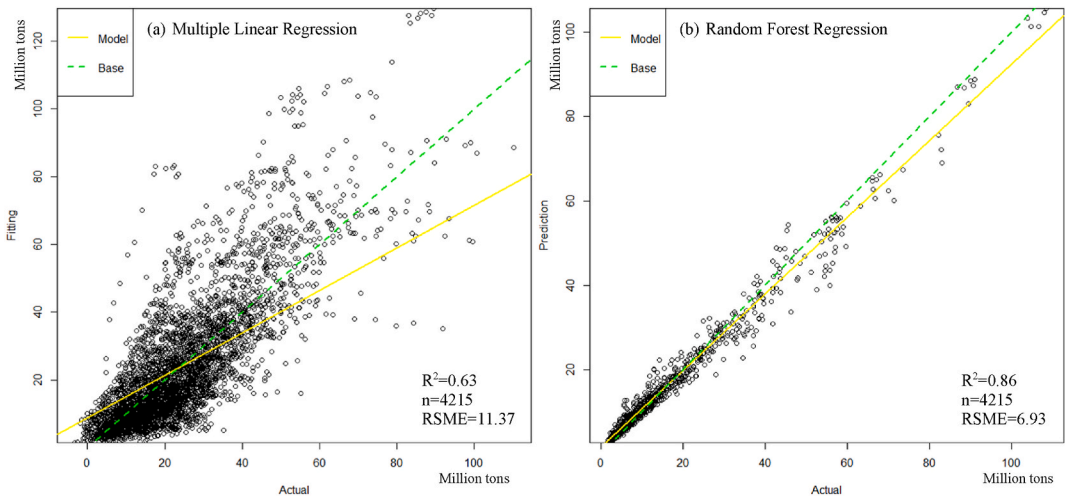


Fig. 1. Scatter plot of CO_2 emissions verification in Multiple Linear Regression (a) and Random Forest Regression (b).

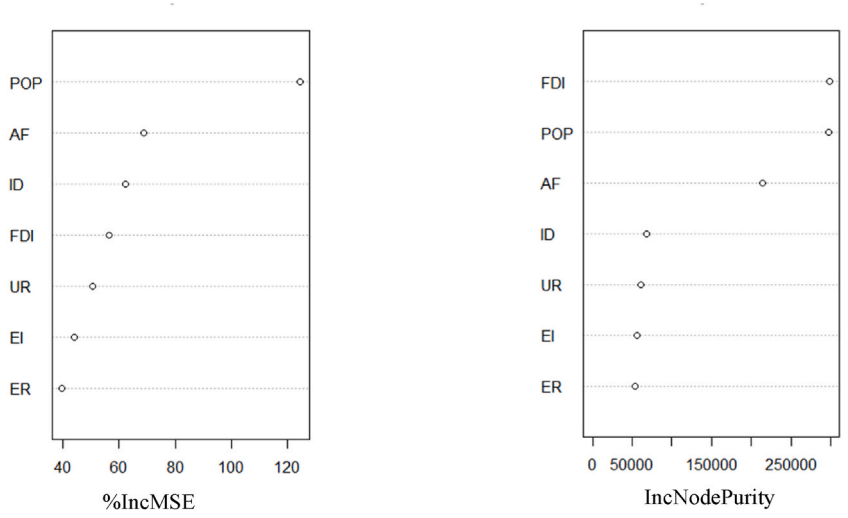


Fig. 2. Importance ranking of influencing factors. Note: IncMSE signifies the average increase in inaccuracy, indicating the rise in the estimation error of the random forest model when a variable is randomly removed, compared to the original error. The greater the IncMSE value, the more significant the variable. IncNodePurity represents the average reduction in node impurity, demonstrating the influence of this variable on each decision tree node. A higher IncNodePurity value implies a more vital variable.

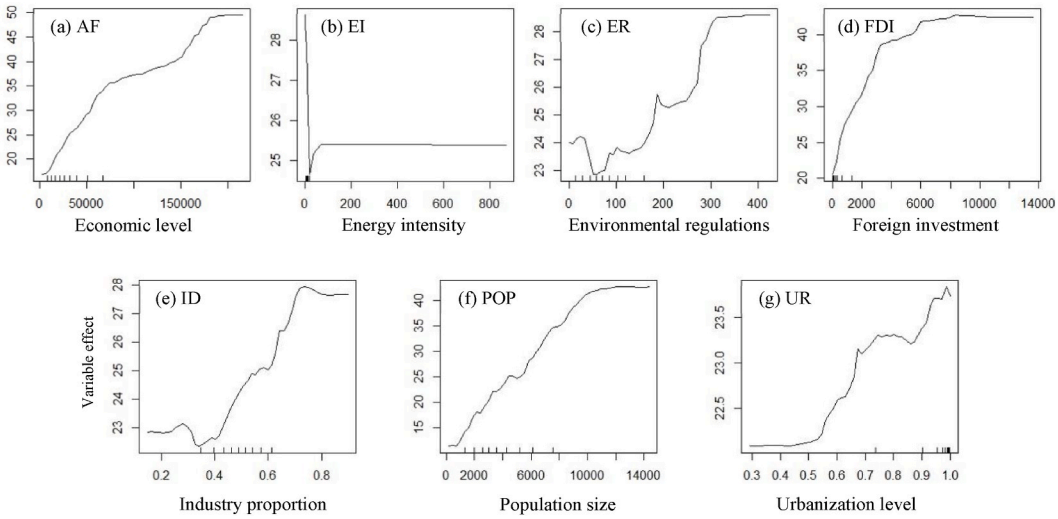


Fig. 3. Partial dependency plots for the factors in the RF model predicting CO_2 annual emissions.

carbon emissions.

Regarding foreign investment (Fig. 3d), carbon emissions increase rapidly with the increase in investment. The growth rate starts to slow when the foreign investment reaches 3 billion. When foreign investment reaches 6 billion, the impact on carbon emissions reaches the maximum and remains stable. This phenomenon suggests that early at the beginning of the foreign investment, they tend to invest in companies with high emissions. With the increasing investment and thorough, the smaller carbon footprint of smaller companies, increasing the proportion of high emissions of enterprise transformation to a low-carbon, by the end of the investment, enterprise of low carbon enterprise with foreign investment proportion rising, eventually making increased carbon emissions are no longer.

The connection between the level of urbanization (Fig. 3g) and greenhouse gas emissions reveals an increasing-stable-increasing pattern. When the urbanization level is below 0.7, there is an upward trend in greenhouse gas emissions with the improvement in the level of urbanization. When the urbanization level falls within the range of 0.7–0.85, greenhouse gas emissions remain relatively steady. As the urbanization level exceeds 0.85, greenhouse gas emissions gradually increase. This result implies that in the initial stages of urban development, a considerable fraction of the rural populace relocates to urban regions. The labor force engaged in production and daily living increases rapidly, leading to a swift rise in greenhouse gas emissions. With the ongoing growth of urbanization, production efficiency begins to gradually improve, and the reduction in greenhouse gas emissions due to increased efficiency begins to

take precedence, maintaining greenhouse gas emissions at a specific level of stability and possibly even initiating a decline. As the level of urbanization continues to rise, the marginal impact of reduced greenhouse gas emissions resulting from efficient production diminishes, the size of the city continues to expand, and greenhouse gas emissions stemming from production and daily living start to increase rapidly.

From the perspective of environmental regulations (Fig. 3c), when the number of environmental regulations is less than 50, the carbon emissions show a downward trend; when the environmental regulations are between 50 and 300, the carbon emissions continue to rise with the increase of environmental regulations; when the environmental regulations are more than 300, the carbon emissions remain stable and show a slight downward trend. This result suggests that less environmental regulation can effectively control carbon emissions when there are low emissions. After the sharp increase in carbon emissions, demand for environmental regulation is increasingly urgent, so the government policies and regulations for carbon reduction. Finally, under the action of a growing number of environmental regulations, carbon is contained, and the emissions remain stable.

When the energy intensity (Fig. 3b) continuously decreases to a critical point, the carbon emissions begin to decline, and when the energy intensity decreases to a certain extent, the carbon emissions begin to increase. This result suggests that in the early stage of development, the energy intensity was relatively high. With the continuous progress of technology, the energy intensity declined rapidly and occupied a dominant position. Advances in technology led to declining carbon emissions. When energy intensity drops to a certain level, the marginal effect of technological progress descends. After the increase of GDP dominated, with the rapid rise of GDP, emissions increased sharply.

3.3. The impact of factor interactions on carbon emissions

Pairwise combinations of determinants were used as input variables in the random forest model to evaluate the impact of interactions among determinants on greenhouse gas emissions (as shown in Fig. 4). The interaction between Population and Economic Level was identified to be responsible for 68.5% of greenhouse gas emissions, while the interaction between Population and Foreign Investment accounted for 63.1% of greenhouse gas emissions. These discoveries highlight the substantial joint impact of population size and economic development on greenhouse gas emissions. In essence, the simultaneous expansion of the population and economic development makes a substantial contribution to primary greenhouse gas emissions.

3.4. Regional differences in the influencing factors

To investigate regional disparities in the influencing factors of carbon emissions, we employed random forest regression to model and assess the factors for prefecture-level cities in each economic region of China. We ranked the seven influencing factors based on their importance (Fig. 5). The regression outcomes indicate the R^2 values of carbon emissions for each economic region, all of which surpass 0.85 with minimal deviation errors. This demonstrates the high predictive accuracy of this method and the convincing ranking of the impact factors.

As shown in the table, with the exception of the western region, population emerges as the most influential factor contributing to greenhouse gas emissions in each region. In regions like central, eastern, and northeastern China, which have larger populations, population size exerts a more significant impact on greenhouse gas emissions compared to the level of economic development. In the western region, characterized by a smaller population, the boost in economic development assumes a more dominant role in driving

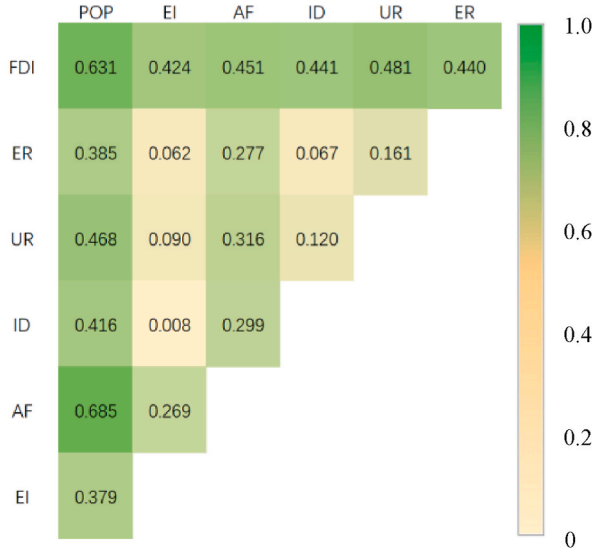


Fig. 4. Spatial influencing magnitude of interactive factors on CO_2 emissions.

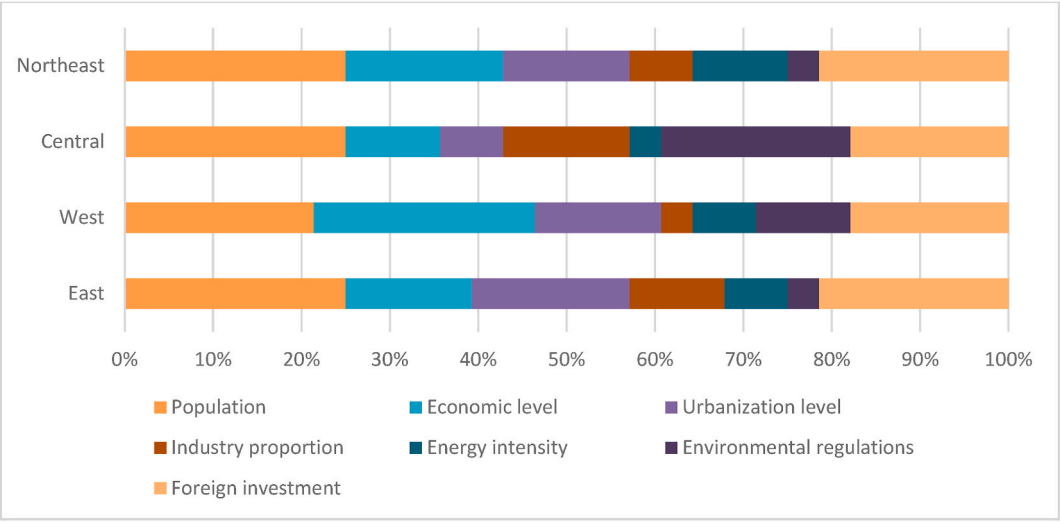


Fig. 5. CO_2 emissions rankings of influencing factors among different regions in China.

the rise in greenhouse gas emissions. Moreover, only in the eastern region does the level of urbanization rank among the top three determinants. This can be attributed to the regional disparities in China; as the eastern region achieves a certain level of urbanization, economies of scale begin to manifest, resulting in a substantial impact on greenhouse gas emissions. Foreign direct investment ranks among the top three influencing factors in all regions. This is attributed to the increased foreign investment across China as part of the country's deepening reform and opening up policies. Foreign investments have led to widespread and large-scale production, stimulating economic development while also contributing to higher carbon emissions. Among all the regions in the country, only the central region ranks environmental regulation within the top three factors. This could be attributed to the central region's role as a national key energy production area, where fossil fuels are the primary energy source. As China's economic development progresses, energy demands rise, leading to increased emissions from energy production. Environmental regulations have been implemented to control carbon emissions as a response to the growing energy production in this region.

4. Discussion

China is confronted with substantial pressure to curtail greenhouse gas emissions. To provide a more practical foundation for the government in formulating carbon reduction policies, we, for the very first time, consider the potential manipulation of data related to emissions of greenhouse gases and examine the impact of crucial factors on greenhouse gas emissions. A majority of previous research emphasizes economic development, population size, energy intensity, industrial composition, urbanization level, foreign investment, environmental regulations, and other determinants. Due to variations in data sources and methodologies, prior research outcomes are not always congruent. Our study fills the gap in objective data aspects and, from a global perspective, reveals the impact of various determinants on greenhouse gas emissions.

Our findings reveal that, generally, the hierarchy of the impact of each determinant on greenhouse gas emissions is as follows: population size > economic development level > industrial composition > foreign investment > level of urbanization > energy intensity > environmental regulations. For instance, Wu [44] investigated the determinants of greenhouse gas emissions in 339 provincial-level cities in China and concluded that, overall, the influence of the level of urbanization was greater than that of industrial composition. Yu [45] analyzed county-level greenhouse gas emissions and found that industrial composition > economic development level. We believe that the disparities between this study and other research conclusions may be attributed to the following reasons: Firstly, this study opted for different CO_2 emission data acquired through a "top-down" approach, which combines nighttime illumination and ground thermal location data, ensuring both the objectivity of research data and enhanced reliability of estimates. Currently, there is no research employing the same data, which may lead to inconsistent results. Secondly, even though both this study and existing research explore the determinants of greenhouse gas emissions in China as a whole, they operate at different scales, with studies encompassing provincial, provincial-level, municipal, and county-level data. Different research scales may lead to varying conclusions. Lastly, most existing studies still rely on traditional measurement methods. While a limited number of studies use the same random forest model, discrepancies in conclusions still arise due to data and other factors.

It is worth noting that the data of 281 prefecture-level cities in China are involved in this paper. Although regional differences are analyzed, the conclusions drawn are still relatively overall, and the actual situation of a prefecture-level city may not be applicable, so more detailed analysis is needed. Although this paper selected some factors that are popular in current research, with the change in China's economic situation and the gradual promotion of carbon reduction policies, such as the vigorous national promotion of energy-saving technology and new energy, other factors like the number of green patents and the use of clean energy are more and more worthy of further exploration. In addition, although this paper uses the latest data we can access, it is still already three years or a more

prolonged time ago. China's economic growth has been diverting in the past three years, and the complex international situation has changed. The timeliness of the conclusion, if can with the updated data for study in recent years, will be more practical significance.

5. Conclusions

Our analysis has unveiled the hierarchical impact of determinants on greenhouse gas emissions in China, highlighting the significance of population size, economic development level, and industrial composition. These key drivers exhibit complex, non-linear relationships, emphasizing the multifaceted nature of emission dynamics in the region.

In light of these findings, policy recommendations are paramount. Shifting policy emphasis from population growth concerns to flexible birth policies and implementing controlled urbanization strategies tailored to specific levels can foster the development of efficient, low-carbon cities. Furthermore, expediting industrialization to reach the inflection point for stable carbon emissions reduction and adopting a comprehensive approach to foreign investment can mitigate emission surges. Timely adaptation of carbon reduction strategies is essential, with a focus on addressing regional variations in determinant impacts through tailored policies.

Looking ahead, future research should explore a broader set of determinants, including emerging factors influencing greenhouse gas emissions. Deepening the application of nonlinear regression models and machine learning methods will provide a more nuanced understanding of determinant responses. Additionally, utilizing credible and recent data sources will enhance the timeliness and reliability of research outcomes, facilitating informed policy interventions and sustainable development strategies.

Data availability statement

The data utilized in this paper has not been deposited into a publicly available database. However, interested parties can request access to the data by contacting the corresponding author via email.

Funding statement

The research received support from "the Fundamental Research Funds for the Central Universities".

CRediT authorship contribution statement

Qingsheng Lei: Writing – review & editing, Validation, Investigation, Conceptualization. **Hongwei Yu:** Writing – review & editing, Resources, Methodology, Data curation. **Zixiang Lin:** Writing – original draft, Software, Project administration, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e29086.

References

[1] J. Rogelj, et al., Paris Agreement climate proposals need a boost to keep warming well below 2 °C, Nature 534 (7609) (2016) 631–639, https://doi.org/10.1038/nature18307.
[2] S. Abbas, et al., Going green: understanding the impacts of economic complexity, clean energy and natural resources on ecological footprint in complex economies, Environ. Dev. Sustain. (2023), https://doi.org/10.1007/s10668-023-04154-4.
[3] B. Dogan, et al., The mitigating effects of economic complexity and renewable energy on carbon emissions in developed countries, Sustain. Dev. 29 (1) (2021) 1–12, https://doi.org/10.1002/sd.2125.
[4] F.V. Bekun, Race to carbon neutrality in South Africa: what role does environmental technological innovation play? Appl. Energy 354 (2024) https://doi.org/10.1016/j.apenergy.2023.122212.
[5] Y. Deng, W.Y. Jiang, Z.Y. Wang, Economic resilience assessment and policy interaction of coal resource oriented cities for the low carbon economy based on AI, Resour. Pol. 82 (2023), https://doi.org/10.1016/j.resourpol.2023.103522.
[6] R. Khalfaoui, et al., Environment-growth nexus and corruption in the MENA region: novel evidence based on method of moments quantile estimations, J. Environ. Manag. 342 (2023), https://doi.org/10.1016/j.jenvman.2023.118146.
[7] S.F. Hong, E.C.M. Hui, Y.Y. Lin, Relationships between carbon emissions and urban population size and density, based on geo-urban scaling analysis: a multi-carbon source empirical study, Urban Clim. 46 (2022), https://doi.org/10.1016/j.uclim.2022.101337.
[8] F. Xiao, et al., Spatial distribution of energy consumption and carbon emission of regional logistics, Sustainability 7 (7) (2015) 9140–9159, https://doi.org/10.3390/su7079140.
[9] B. Dogan, et al., Formulating energy security strategies for a sustainable environment: evidence from the newly industrialized economies, Renew. Sustain. Energy Rev. 184 (2023), https://doi.org/10.1016/j.rser.2023.113551.
[10] S. Ghosh, et al., Harnessing the roles of renewable energy, high tech industries, and financial globalization for environmental sustainability: evidence from newly industrialized economies, Nat. Resour. Forum (2023), https://doi.org/10.1111/1477-8947.12356.

[11] C.M. Li, et al., Low-carbon strategy, entrepreneurial activity, and industrial structure change: evidence from a quasi-natural experiment, J. Clean. Prod. 427 (2023), https://doi.org/10.1016/j.jclepro.2023.139183.
[12] G.F. Wang, M.L. Liao, J. Jiang, Research on agricultural carbon emissions and regional carbon emissions reduction strategies in China, Sustainability 12 (7) (2020), https://doi.org/10.3390/su12072627.
[13] C.M. Li, et al., The impact of smart cities on entrepreneurial activity: evidence from a quasi-natural experiment in China, Resour. Pol. 81 (2023), https://doi.org/10.1016/j.resourpol.2023.103333.
[14] B. Dogan, D. Balsalobre-Lorente, M.A. Nasir, European commitment to COP21 and the role of energy consumption, FDI, trade and economic complexity in sustaining economic growth, J. Environ. Manag. 273 (2020), https://doi.org/10.1016/j.jenvman.2020.111146.
[15] C.M. Li, et al., Digital finance and enterprise financing constraints: structural characteristics and mechanism identification, J. Bus. Res. 165 (2023), https://doi.org/10.1016/j.jbusres.2023.114074.
[16] B. Dogan, et al., How environmental taxes and carbon emissions are related in the G7 economies? Renew. Energy 187 (2022) 645–656, https://doi.org/10.1016/j.renene.2022.01.077.
[17] B. Dogan, et al., Impacts of export quality on environmental degradation: does income matter? Environ. Sci. Pollut. Control Ser. 27 (12) (2020) 13735–13772, https://doi.org/10.1007/s11356-019-07371-5.
[18] B. Dogan, et al., What do we learn from Nexus between trade diversification and structural change: informing the future about climate action and Sustainability, Environ. Sci. Pollut. Control Ser. 30 (40) (2023) 92162–92181, https://doi.org/10.1007/s11356-023-28770-9.
[19] C.G. Zhang, Z. Tan, The relationships between population factors and China's carbon emissions: does population aging matter? Renew. Sustain. Energy Rev. 65 (2016) 1018–1025, https://doi.org/10.1016/j.rser.2016.06.083.
[20] W. Guo, T. Sun, H.J. Dai, Effect of population structure change on carbon emission in China, Sustainability 8 (3) (2016), https://doi.org/10.3390/su8030225.
[21] F.H. Wen, Z.L. Sun, Y. Luo, Population structure and local carbon emission reduction: evidence from guangdong, China, Sustainability 15 (5) (2023), https://doi.org/10.3390/su15054079.
[22] Y.C. Yi, J. Qi, D. Chen, Impact of population agglomeration in big cities on carbon emissions, Environ. Sci. Pollut. Control Ser. 29 (57) (2022) 86692–86706, https://doi.org/10.1007/s11356-022-21722-9.
[23] D.D. Li, et al., The clean energy development path and sustainable development of the ecological environment driven by big data for mining projects, J. Environ. Manag. 348 (2023), https://doi.org/10.1016/j.jenvman.2023.119426.
[24] J.E. Payne, et al., The effect of economic complexity and energy security on measures of energy efficiency: evidence from panel quantile analysis, Energy Pol. 177 (2023), https://doi.org/10.1016/j.enpol.2023.113547.
[25] M. Cellura, M.A. Cusenza, S. Longo, Energy-related GHG emissions balances: IPCC versus LCA, Sci. Total Environ. 628–629 (2018) 1328–1339, https://doi.org/10.1016/j.scitotenv.2018.02.145.
[26] G. Panagakos, et al., Monitoring the carbon footprint of dry bulk shipping in the EU: an early assessment of the MRV regulation, Sustainability 11 (18) (2019), https://doi.org/10.3390/su11185133.
[27] M.o.E.a. Environment, in: S.M. Liu (Ed.), Tampering Detection Reports and Verifying through the Motions, the Ministry of Ecology and Environment Reported Four Institutions for Falsifying Carbon Emission Report Data, China Times, 2022.
[28] R.M. Bao, W.R. Jiang, Research and analysis of data mining in indoor environmental monitoring, in: International Conference on Environmental Remote Sensing and Big Data (ERSBD 2021), 2021.
[29] T. Oda, et al., Errors and uncertainties in a gridded carbon dioxide emissions inventory, Mitig. Adapt. Strategies Glob. Change 24 (6) (2019) 1007–1050, https://doi.org/10.1007/s11027-019-09877-2.
[30] X.L. Cui, et al., Mapping spatiotemporal variations of CO_2 (carbon dioxide) emissions using nighttime light data in Guangdong Province, Phys. Chem. Earth 110 (2019) 89–98, https://doi.org/10.1016/j.pce.2019.01.007.
[31] W.S. Zhang, et al., Effects of urbanization on airport CO_2 emissions: a geographically weighted approach using nighttime light data in China, Resour. Conserv. Recycl. 150 (2019), https://doi.org/10.1016/j.resconrec.2019.104454.
[32] C.N.H. Doll, J.P. Muller, J.G. Morley, Mapping regional economic activity from night-time light satellite imagery, Ecol. Econ. 57 (1) (2006) 75–92, https://doi.org/10.1016/j.ecolecon.2005.03.007.
[33] Q.Q. Shen, T. Sun, K.M. Luo, Research on China's tourism carbon emission efficiency and its influencing factors, Fresenius Environ. Bull. 28 (9) (2019) 6380–6388.
[34] H. Zhang, et al., Spatial planning factors that influence CO_2 emissions: a systematic literature review, Urban Clim. 36 (2021), https://doi.org/10.1016/j.uclim.2021.100809.
[35] W.X. Wang, D.Q. Zhao, Y.Q. Kuang, Decomposition analysis on influence factors of direct household energy-related carbon emission in Guangdong provinceBased on extended Kaya identity, Environ. Prog. Sustain. Energy 35 (1) (2016) 298–307, https://doi.org/10.1002/ep.12219.
[36] T. Oda, S. Maksyutov, A very high-resolution (1 km×1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights, Atmos. Chem. Phys. 11 (2) (2011) 543–556, https://doi.org/10.5194/acp-11-543-2011.
[37] T. Oda, S. Maksyutov, R.J. Andres, The Open-source Data Inventory for Anthropogenic CO_2, version 2016 (ODIAC2016): a global monthly fossil fuel CO_2 gridded emissions data product for tracer transport simulations and surface flux inversions, Earth Syst. Sci. Data 10 (1) (2018) 87–107, https://doi.org/10.5194/essd-10-87-2018.
[38] Z.Y. Wang, et al., Risk prediction and credibility detection of network public opinion using blockchain technology, Technol. Forecast. Soc. Change 187 (2023), https://doi.org/10.1016/j.techfore.2022.122177.
[39] Z.Y. Wang, et al., Achieving sustainable development goal 9: a study of enterprise resource optimization based on artificial intelligence algorithms, Resour. Pol. 80 (2023), https://doi.org/10.1016/j.resourpol.2022.103212.
[40] J.Y. Lin, et al., Analyzing the impact of three-dimensional building structure on CO_2 emissions based on random forest regression, Energy 236 (2021), https://doi.org/10.1016/j.energy.2021.121502.
[41] W. Sun, Y.W. Wang, C.C. Zhang, Forecasting CO_2 emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine, Environ. Sci. Pollut. Control Ser. 25 (29) (2018) 28985–28997, https://doi.org/10.1007/s11356-018-2738-z.
[42] L. Wen, X.Y. Yuan, Forecasting CO_2 emissions in Chinas commercial department, through BP neural network based on random forest and PSO, Sci. Total Environ. (2020) 718, https://doi.org/10.1016/j.scitotenv.2020.137194.
[43] H. Zhang, et al., Use of random forest based on the effects of urban governance elements to forecast CO_2 emissions in Chinese cities, Heliyon 9 (6) (2023) e16693, https://doi.org/10.1016/j.heliyon.2023.e16693.
[44] J.S. Wu, City scale analysis of China's carbon emissions and influencing factors, Environ. Sci. (2023) 1–13.
[45] W.M. Yu, T.T. Zhang, D.J. Shen, Evolution analysis of carbon emission intensity pattern and influencing factors in Chinese county based on Random Forest Model, Environ. Sci. China 42 (6) (2022) 2788–2798.