Environmental Science and Pollution Research (2021) 28:19260–19271
https://doi.org/10.1007/s11356-020-12294-7

RESEARCH ARTICLE

Modeling and predicting city-level CO_2 emissions using open access data and machine learning

Ying Li[1] · Yanwei Sun[2]

Received: 24 September 2020 / Accepted: 29 December 2020 / Published online: 4 January 2021

Abstract

Globally, urban has been the major contributor to greenhouse gas (GHG) emissions and thus plays an increasingly important role in its efforts to reduce CO_2 emissions. However, quantifying city-level CO_2 emissions is generally a difficult task due to lacking or lower quality of energy-related statistics data, especially for some underdeveloped areas. To address this issue, this study used a set of open access data and machine learning methods to estimate and predict city-level CO_2 emissions across China. Two feature selection technologies including Recursive Feature Elimination and Boruta were used to extract the important critical variables and input parameters for modeling CO_2 emissions. Finally, 18 out of 31 predictor variables were selected to establish prediction models of CO_2 emissions. We found that the statistical indicators of urban environment pollution (such as industrial SO_2 and dust emissions per capita) are the most important variables for predicting the city-level CO_2 emissions in China. The XGBoost models obtained the highest estimation accuracy with $R^2 > 0.98$ and lower relative error (about 0.8%) than other methods. The CO_2 emissions predictive accuracy can be improved modestly by combing geospatial and meteorological interpolation predictor variables (e.g., DEM, annual average precipitation, and air temperature). We also observed an S-shape relationship between urban CO_2 emissions per capita and urban economic growth when the rest variables were held constant, rather than a U-shaped one. The findings presented herein provide a first proof of concept that easily available socioeconomic statistical records and geospatial data at urban areas have the potential to accurately predict city-level CO_2 emissions with the aid of machine learning algorithms. Our approach can be used to generate carbon footprint maps frequently for the undeveloped regions with scarce detailed energy-related statistical data, to assist policy-makers in designing specific measures of reducing and allocating carbon emissions reduction goal.

Keywords City-level CO_2 emissions · Machine learning algorithms · Socioeconomic statistical information · Geospatial dataset

Introduction

Climate change characterized by global warming is a common environmental challenge (Ballantyne et al. 2016). It is well-known that the increased concentrations of greenhouse gases are a major contributor to climate change, and carbon dioxide (CO_2) accounts for more than 75% (IPCC 2014). Meanwhile, global cities release 71–76% of the carbon dioxide emissions (Seto et al. 2014). Rising CO_2 emissions have become a serious global problem. Many studies on CO_2 emission have conducted to understand CO_2 emission mechanisms and future trends, which plays a vital role in the global climate change mitigation. More recently, research on CO_2 emission mainly focuses on the relationship between energy consumption, CO_2 emission, and economic growth; carbon emission measurement and prediction; analysis of carbon emission influencing factors; spatial and temporal distribution characteristics of carbon emission; and so on. Wasti and Zaidi (2020) disclosed that from GDP to CO_2 emissions and energy consumption to trade liberalization, the causality was unidirectional in Kuwait. However, there is a bi-directional causal relationship between CO_2 emissions and energy consumption. Hu et al. (2020) investigated the temporal and spatial evolution and driving factors of CO_2 emission decoupling from 1991 to 2016 in 57 Belt and Road Initiative

Responsible Editor: Marcus Schulz

✉ Yanwei Sun
sunyanwei@nbu.edu.cn

Ying Li
2020800236@usth.edu.cn

[1] School of Mining Engineering, Heilongjiang University of Science and Technology, Harbin 150022, People's Republic of China

[2] Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo 315211, People's Republic of China

countries. Jeong et al. (2018) used the optimized gene expression programming-harmony search optimization model to forecast CO_2 emission of South Korea in 2030. Bayar et al. (2020) found financial sector development, and primary energy consumption had a positive impact on CO_2 emissions in post-transition European Union countries.

Due to lack of related energy statistic data, researches at the city level are scarce. However, cities account for more than 80% of national energy consumption and carbon emissions (Dhakal 2009). Therefore, understanding and predicting CO_2 emissions in urban areas is particularly important for formulating effective mitigation policies and decomposing greenhouse gas emission reduction targets. Some scholars have proposed statistical models to predict CO_2 emissions at different spatial scales, which could make up for lack of related energy statistic data. For example, Fang et al. (2018) proposed an improved Gaussian process regression method for CO_2 emission prediction and successfully tested the model by using the total CO_2 emission data of the USA, China, and Japan from 1980 to 2012. Huang et al. (2019) used the gray correlation analysis method to identify the influencing factors and established the long- and short-term memory (LSTM) method to predict CO_2 emissions in China. Among these methods, machine learning algorithms have become useful tools for analyzing potential influencing factors and predicting carbon emissions. Linear regression models and artificial neural networks can predict energy demand for heating and cooling, energy consumption, and CO_2 emissions in Chilean office buildings (Pino-Mejías et al. 2017). The hybrid model combining random forest and limit learning machine can be used for CO_2 emission prediction at provincial scale (Sun et al. 2018).

Generally, previous literature mainly focused on the prediction of time evolution characteristics of CO_2 emissions for one or more specific cities. In terms of spatial dimension, modeling city-level CO_2 emissions still encountered significant uncertainty. Thus, accurately prediction framework should be further explored under the restriction of data availability. In our study, we took 182 Chinese cities as examples and applied four mainstream machine learning algorithms to predict CO_2 emissions of the other cities in China. Our research can not only supplement lacking data at the city level but also evaluate the performance of the four mainstream machine learning algorithms. The research methods and results can be widely used in the study of CO_2 emissions in other cities without data in the world. This is very meaningful and helpful for government decision-makers, because only by understanding the spatial distribution of CO_2 emissions within urban agglomerations, and clarifying the spatial correlation and spatial structure of individual CO_2 emissions in cities, can they formulate feasible CO_2 emission reduction measures and achieve coordinated reduction.

China has become the world's largest energy consumer and oil importer since 2009. In order to slow down the current global warming, the Chinese government promised in the "National Independent Emission Reduction Contributions" document that China will no longer increase carbon emissions by 2030, and the carbon emission intensity per unit of GDP needs to be reduced by 60~65% compared to 2005. The "Chinese Dream" is an "urban dream" (Taylor 2015). At the United Nations General Assembly in September this year, China propose that it will strive to achieve carbon neutral by 2060. However, with the rapid development of urbanization and industrialization, a substantial increase in total CO_2 emissions in Chinese cities is an inevitable trend in the near future. However, due to the lack of data on a smaller scale in China, there are few studies on urban carbon emissions (Meng et al. 2014). And studies at the urban level are limited by energy-related statistics, especially in less developed in the mid-western region of China. In this study, the primary objectives are to: (1) compare and evaluate the prediction performance of 4 mainstream machine learning algorithms, namely, Gradient Boosting Machine (GBM), Support Vector Machines (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), in modeling city-level CO_2 emissions with sparse socioeconomic statistical information and geospatial data; (2) conduct the empirical study by a cross-sectional dataset of CO_2 emissions of 182 Chinese cities in 2010 and predict the CO_2 emissions from all other 284 prefecture-level cities in China according to the cross-sectional dataset of CO_2 emissions of 182 Chinese cities in 2010; and (3) test the relative influence of socioeconomic and climatic factors and urban form on carbon emissions metrics in China. The results can provide key information for helping to prepare future CO_2 emission reduction policies in China.

Materials and methods

Modeling and predicting framework

A modeling flowchart for city-level CO_2 emissions is shown in Fig. 1. The whole framework is divided into five steps: Firstly, the dataset of city-level CO_2 emissions and predictors variables (socioeconomic statistical records and geospatial data) was collected and pre-processed from various public data sources; secondly, two feature selection technologies including Recursive Feature Elimination and Boruta were used to extract the critical input parameters for modeling CO_2 emissions; thirdly, we trained 4 categories of machine learning algorithm (including GBM, SVM, RF, and XGBoost) and optimized algorithm parameters by tenfold cross-validation (CV); fourthly, we evaluated and compared the performance of the final acquired models through three statistical indicators, including overall adjusted (R^2) and Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE); finally, the CO_2 emissions prediction result of 284 prefecture-level

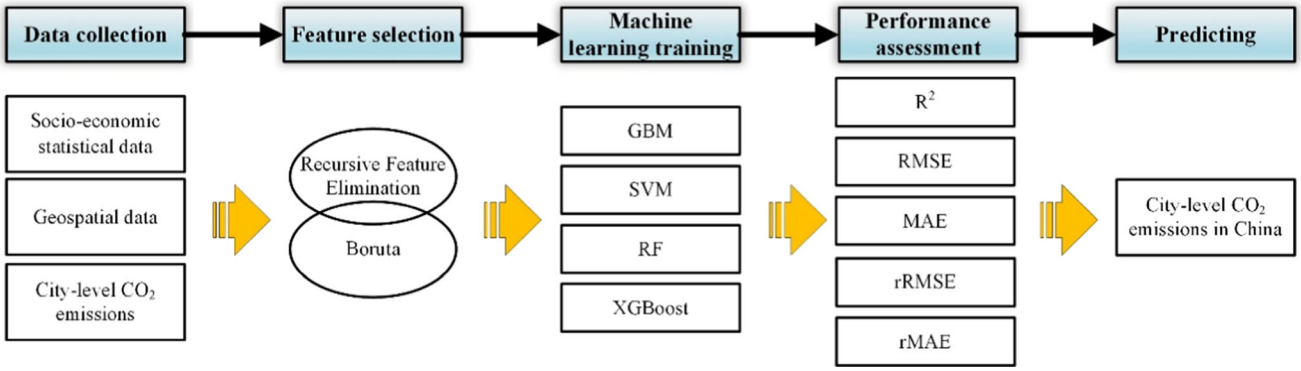

Fig. 1 The modeling flowchart of city-level CO_2 emissions

cities in China was obtained by the optimized machine learning algorithm.

City-level CO_2 emissions

A cross-sectional dataset of city-level CO_2 emissions used in this study is derived from China Emission Accounts and Datasets (http://www.ceads.net). This dataset was constructed by the territorial scope and the Intergovernmental Panel on Climate Change (IPCC) calculation method, which provided the CO_2 emission inventories for 182 Chinese cities in 2010 by 17 fossil fuels and 46 socioeconomic sectors (Shan et al. 2018). The selection of case cities was based on the data availability and covered more than 62% of China's population as well as 77% of GDP. As shown in Fig. 2, the average CO_2 emission among the 182 Chinese cities was 41.81 million tons with the standard deviation of Std. = 37.03 million ton. In 2010, the cities with carbon emissions lower than 50 million tons accounted for 71% of all selected cases. Comprehensive and consistent inventories of city-level emissions could provide robust data support for China's emission control and the carbon emission reduction related researches. More details of this dataset were presented in the study of Shan et al. (2018). The per capita CO_2 emission for each city was used as the dependent variable in subsequent modeling.

Predictor variables

Previous case studies demonstrate that urban greenhouse gas emissions are driven by socioeconomic, climatic, and urban form–specific characteristics (Baiocchi et al. 2015; Huang et al. 2019). According to the existing literature, a total of 31 predictor variables were selected (see Table 1).

To simplify the predicting process of city-level CO_2 emissions, all data involved in the analysis is publicly available through national statistics offices and data sharing website. The detailed descriptions on the variables are as following:

Socioeconomic variables: All demographic and economic indicators of each city in China are collected from the China City Statistics Yearbook (2011). The size of population, GDP, administrative region, and built-up were used as predictor variables. In addition, the statistical-based variables such as transportation, energy consumption, and waste emissions also have been considered for this analysis.

Geospatial variables: Four geospatial variables were chosen as the supplementary predictor of city-level CO_2 emissions. The nighttime light images were widely used as a proxy for economic growth and human-caused CO_2 emissions (Wang et al. 2017; Zhao et al. 2019). Yearly Defense Meteorological Satellite Program's Operational

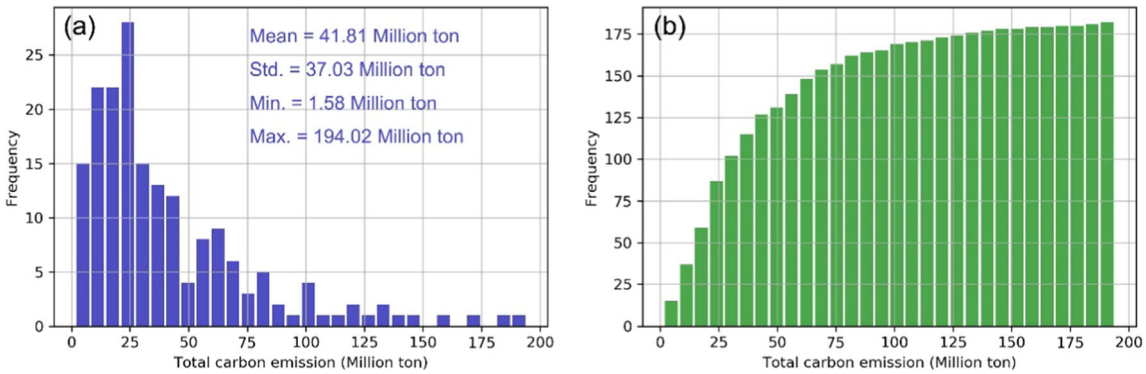

Fig. 2 Histograms of total carbon emission data for 182 Chinese cities in 2010: (**a**) frequency histogram and (**b**) cumulative frequency histogram

Table 1 Predictor variables used in machine learning models

Abbreviation	Variable description
Socioeconomic variables	
ToEmp	Total employees [person per capita]
PopDen	Population density [people per km^2]
PerGDP	Per capita GDP [Yuan]
GroGDP	Annual growth rate of GDP [%]
Primary	Proportion of primary industry [%]
Second	Proportion of secondary industry [%]
Tertiary	Proportion of tertiary industry [%]
ToEnter	Total number of industrial enterprises [unit per 10^4 capita]
GroOut	Gross value of industrial output [Yuan per capita]
FixInve	Fixed asset investment [Yuan per capita]
ToPasse	Total volume of passengers [person per capita]
ToFreig	Total volume of freight [ton per capita]
ToWater	Total amount of water supply [ton per capita]
ToElect	Total amount of electricity consumption [kwh per capita]
ToGas	Total amount of gas supply [m^3 per capita]
ToLPG	Total amount of LPG supply [ton per capita]
ToBus	Total number of buses in operation [unit per 10^4 capita]
ToBusPass	Total volume of passengers with bus [person per capita]
ToTaxi	Total number of taxi [unit per 10^4 capita]
ToOwnBus	Total number of owning bus per 10^4 person
PerRoad	Per capita road surface area [m^2 per capita]
ToWaste	Total volume of industrial waste water discharge [ton per capita]
ToSO2	Total volume of industrial SO_2 emissions [ton per capita]
ToDust	Total volume of industrial dust emissions [ton per capita]
PopMun	Population proportion of municipal districts [%]
Geospatial variables	
MeanLight	Mean value of nighttime lights intensity for each city
SumLight	Sum value of nighttime lights intensity for each city
DEM	Mean value of DEM for each city [m]
ToUrbanB	Total area of urban built-up per capita [km^2 per capita]
Climatic background variables	
AvePrecip	Average annual precipitation [mm]
AveTemp	Average annual temperature [°C]

Linescan System (DMSP-OLS) composite data was downloaded from the National Oceanic and Atmospheric Administration's National Geophysical Data Center (NOAA/NGDC) https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html). We calculated the sum and average DN value of DEMP-OLS data at city level as the explanatory indicator of CO_2 emissions. The spatial predictors including the average altitude and urban built-up area were chosen to represent geomorphology and urban form features.

Climatic factors: In this study, the surface meteorological parameters data in 2010 (air temperature and precipitation) were downloaded from the National Meteorological Information Center (http://data.cma.cn/). The inverse distance weighted method in ArcGIS 10.2 was used to interpolate the annual mean values of the two meteorological parameters into a resolution of 1 km^2, and then the regional statistical function was used to calculate the mean values of meteorological factors of each city.

Feature selection

To avoid information redundancy generated by large number of features, we integrated two feature selection technologies

(Recursive Feature Elimination and Boruta) to identify the critical features and speed up the training process. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is an efficient feature selection technique, which can select a fixed number of high-ranking features for subsequent analyzes (Guyon et al. 2002). As Fig. 3 shown, the prediction accuracy was close to the highest when the 18 high-ranked variables were selected. The Boruta algorithm was further applied to filter the control variables of city-level CO_2 emissions. This method proposed by Kursa and Rudnicki (2010) is an all-relevant feature selection wrapper algorithm based on the RF algorithm. It was widely used in the preliminary selection of model variables by adding randomness to the system and collecting results from the ensemble of randomized samples. This process can reduce the misleading impact of random fluctuations and their collinearity with phenomena of interest. Figure 4 shows the importance ranking of predictor variables for each forest parameter. According to the Boruta algorithm, 11 of 31 variables were rejected with the red boxes, while 17 variables were confirmed by the green boxes. The blue boxes indicate the minimum, mean, and maximum Z-scores of the shadow feature in the Boruta algorithm. The rest of 3 variables with yellow boxes denote tentative parameters. Based on the results of the two feature selection technologies, we finally chose 18 high-ranking variables as the input features to model city-level CO_2 emissions.

Modeling techniques

The model concludes Gradient Boosting Machine (GBM), Support Vector Machines (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost).

GBM is a stepwise additive regression model (Friedman 2001, 2002), which is unique in that it attempts to build a new basis based on the previously assembled tree that has the greatest negative gradient correlation with the loss function (Natekin and Knoll 2013).

SVM is a supervised machine learning technique that is used for both classification and regression purposes (Cortes and Vapnik 1995). This method has a better learning capability in solving small sample, nonlinear problem. In our study, the hyperparameters of SVM model were determined by the grid search.

RF is the ensemble learning algorithm developed by Breiman (2001), which can handle noisy and highly correlated predictor variables, thus providing more accurate regression prediction. During the model construction, two important parameters, the number of trees in the forest (ntree) and number of variables randomly sampled (mtry), must be defined.

The XGBoost algorithm proposed by Chen and Guestrin (2016) is a novel implementation method that is developed from gradient boosting. The XGBoost can help to reduce overfitting and perform a better prediction accuracy (Fan et al. 2018).

All predictor variables were log-normalized before modeling. To obtain optimal parameters for four machine learning models, we conducted a grid search using tenfold cross-validation. Table 2 provides a short overview of optimal parameters for each model. All processing was implemented within the R software environment using multiple contributed packages (R Development Core Team 2008).

Evaluation and comparison of model performance

Evaluation of model performance is an important step to select the final optimal model. In this study, leave one out cross-validation (LOOCV) was used to assess and compare the prediction performance of each modeling technique. This step is repeated 10 times to stabilize each model evaluation. Several statistical data indices, including the coefficient of determination (R^2), root mean squared error (RMSE), and Mean Absolute Error (MAE), were then calculated for each cross-validation to further characterize the prediction error. Generally, the model with higher R^2 and lower RMSE/MAE is considered to be the more accurate model. Finally, we also

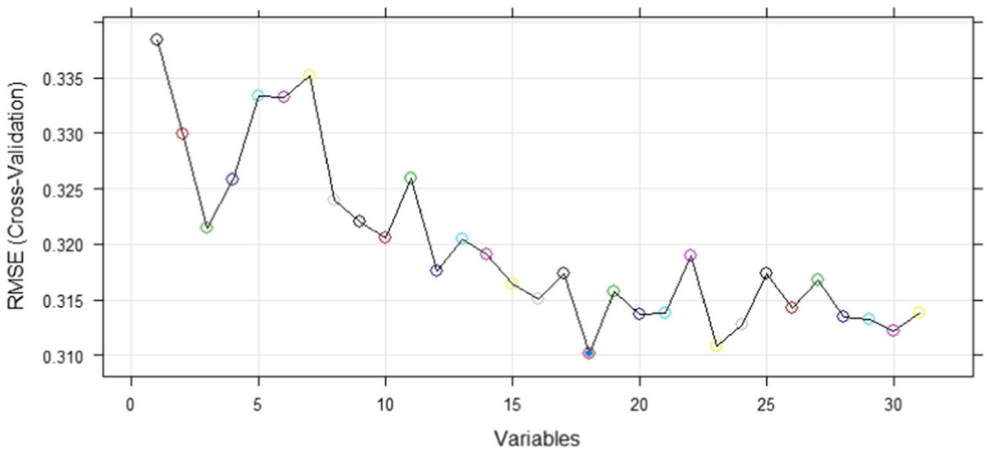

Fig. 3 Relationship between the number of features and prediction accuracy using SVM-RFE algorithm with tenfold cross-validation

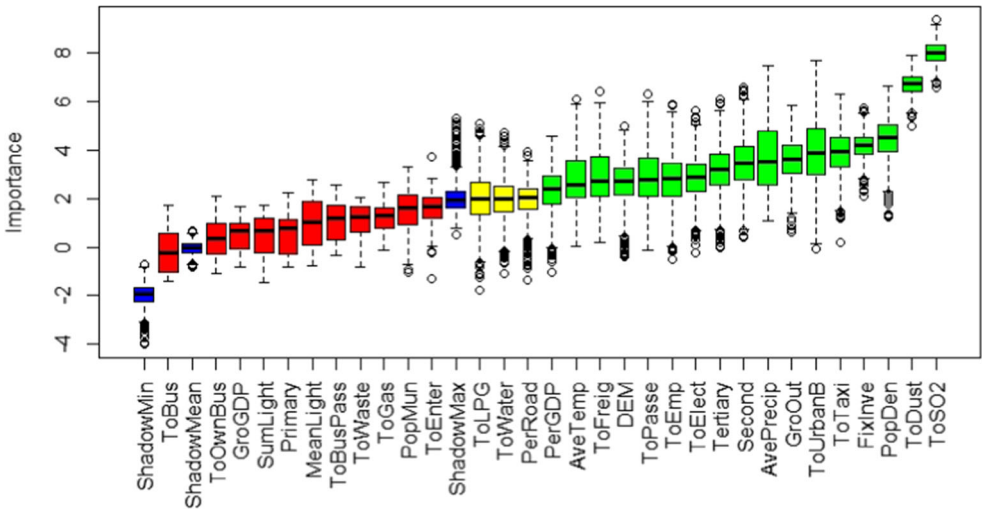

Fig. 4 Contribution ranking of predictor variables derived from the Boruta algorithm

calculated the relative metrics as relative Mean Absolute Error (rMAE) and relative Root Mean Squared Error (rRMSE) through dividing the absolute metrics by the mean of observed values.

Prediction and driving analysis

All predictors for 284 prefecture-level cities over China were collected. Finally, the amount of CO_2 emissions in 2010 for the other prefecture-level cities could be predicted by the optimized machine learning algorithm.

To explore relationships between the various socioeconomic and geographical variables and per capita CO_2 emission variation, we calculated relative variable importance and graphed partial dependency plots (PDPs) from the random forest models. Relative variable importance can be used to assess the contribution of each variable to model fit and averaging this contribution across all trees. The PDP plots demonstrate the influence of a change in the value of an independent variable on the variation of CO_2 emission, while all other variables are held constant.

Results and discussion

Model performance

The comparison of the overall performance of estimation model is given in Fig. 5. Though measuring both absolute and relative metrics, the best estimation accuracy for city-level CO_2 emission is obtained with the XGBoost algorithm. This model shows the lowest estimation error with MAE = 0.27 ton (rMAE ≈ 0.7%). The SVM model performed slightly better than the RF model in terms of absolute metrics, with MAEs of 6.71 and 7.05, respectively. However, both SVM and RF models showed relative larger errors for the cities with

Table 2 Optimal parameters of four machine learning techniques as result of the grid search

Methods	Parameters	Optimum values
GBM	Minimum number of observations in the terminal nodes of the trees (n.minobsinnode)	10
	Total number of trees to fit (n.trees)	50
	Maximum depth of each tree (interaction.depth)	3
	Shrinkage parameter applied to each tree in the expansion (shrinkage)	0.1
SVM	Cost of constraints violation (cost)	0.25
	Regularized L1-loss support vector regression (loss)	L1
RF	Number of trees (ntree)	8
	Number of variables randomly sampled at each split (mtry)	706
XGBoost	Step size shrinkage used in update to prevents overfitting (eta)	0.3
	The max number of iterations (nrounds)	50
	L2 regularization term on weights (lambda)	0.1
	L1 regularization term on weights (alpha)	0.1

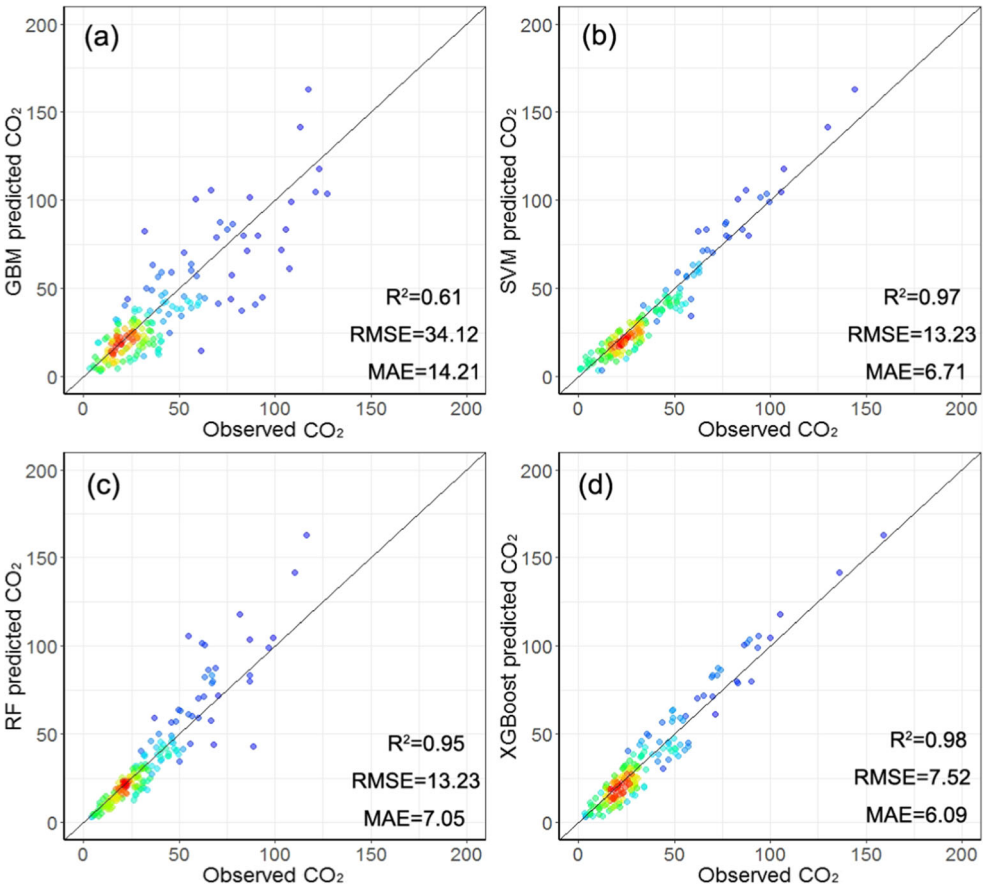

Fig. 5 Predicted vs. observed city-level CO_2 emissions derived from Gradient Boosting Machine (GBM) (**a**), Support Vector Machines (SVM) (**b**), Random Forest (RF) (**c**), and eXtreme Gradient Boosting (XGBoost) (**d**)

higher CO_2 emission per capita (see Fig. 5c). The GBM model showed the worst performance with estimation error reaching MAE = 34.12 ton. Consequently, the XGBoost algorithm can be actually considered unbiased estimators for city-level CO_2 emission based on cross-sectional data.

Figure 5 also demonstrate the overestimation of total amount of CO_2 emission for energy production and heavy industry cities with higher emission levels (> 50 million tons), such as Tangshan and ordos. Apparently, the bias of Random Forest model is larger than the corresponding SVM and XGBoost models. Comparing the two best performing techniques, the RF model provides more accuracy estimations at lower emission levels (< 25 million tons), especially for service and high technology-typed cities. Overall, these results suggest that the selection of optimal modeling technique depend largely on the level of CO_2 emissions of the city. The proposed machine learning models can be very useful for estimating the city-level CO_2 emission with various available parameters.

Driving factors analysis

As shown in Fig. 4, the explanatory variables were ranked according to their contribution to model fitting. Among chosen 18 high-ranking variables, the industrial SO_2 and dust emissions per capita (ToSO2 and ToDust) are the most important indicators for predicting the city-level CO_2 emissions in China. The industrial sectors are CO_2 emission-intensive industries (Li et al. 2020), and industrial agglomeration increases CO_2 emissions (Han et al. 2018). The remaining social-economical statistical indices, like population density (PopDen), fixed asset investment (FixInve), industrial structure (Grout, Second, and Tertiary), traffic condition (ToTaxi, ToPasse, ToFreig, PerRoad), and energy consumption (ToElect), show reasonably high relative influence on variation of city-level CO_2 emissions. There are several reasons as follows: Firstly, in terms of population density, Zhang et al. (2017) and Shi et al. (2020) found that the relationship between population and CO_2 emissions was significantly positively correlated with larger population size leading to more CO_2 emissions. Secondly, investment-driven growth model leads to the increase of CO_2 emissions (Lin and Wang 2020). Thirdly, for industrial structure, some researches show that industrial structure and CO_2 emissions are closely related, and the optimization and upgrading of manufacturing industry can promote CO_2 emission reduction (Kunnas et al. 2009; Hammond and Norman 2012). Fourthly, for transport, it has become the second largest CO_2 emission in China, and the importance of road transport is very high as an intervening area for carbon emission reduction (Lin et al. 2019; Alam

et al. 2020). In particular, taxi is so inefficient that its emissions are about 20 times higher than those of public transport in China (He et al. 2011). Finally, energy consumption is an important factor affecting CO_2 emission (Heun et al. 2015), and energy efficiency influences CO_2 emissions, which is worthy of worth further exploration (Chen et al. 2021). At the same time, renewable energies serve as a feasible policy tool to abate CO_2 emissions (Jebli et al. 2016).

For geospatial indices, the per capita urban built-up area (ToUrbanB) and regional mean altitude (DEM) rank in 6 and 14 of all 18 variables, respectively. This is because urbanization means the continuous agglomeration of population and resources to cities, and carbon that has accumulated over long time periods has been released because of urban development (Milnar and Ramaswami 2020). It is worthy to note that two of nighttime lights intensity indices (SumLight and MeanLight) were not chosen as modeling variables by the Boruta algorithm. In terms of climatic background variables, annual mean precipitation presents more important influence on CO_2 emissions than the annual mean temperature.

Partial dependency plots (PDPs) were graphed for four selected variables to display the dependent relationships between CO_2 emissions and the variables of our interest, while the rest variables are held constant (Fig. 6). On the prefecture-level, per capita CO_2 emissions increased as industrial SO_2 emissions per capita, proportion of secondary industry, and per capita GDP increased (Fig. 6 a, b, and c). Specifically, per capita CO_2 emissions was highest at the industrial SO_2 emissions per capita > 0.28 ton, proportion of secondary industry > 50%, and per capita GDP > 7.5×10^4 Yuan. This implies an S-shaped curve between economic growth and CO_2 emissions in China, and when per capita GDP of cities reach up to or over 7.5×10^4 Yuan, per capita CO_2 emissions will remain stable.

On the contrary, the PDPs indicated that per capita CO_2 emissions tend to decrease with the increase of population density in cities. Figure 6 d shows that when population density in cities was larger than 50 person/km^2, per capita CO_2 emissions will be close to be stable.

Prediction of city-level CO_2 emissions across China

After the training and validation process of machine learning model, the prediction of city-level CO_2 emissions in 2010 was executed for 284 prefecture-level cities over China. The statistics results were displayed in Fig. 7. The average value of city-level CO_2 emissions in 2010 was around 38.69 ton per capita, whereas there was significantly regional difference between three economy-geographic regions in China. The western region is found to demonstrate the highest value (about 50.18 ton) and follow by central region (33.99 ton) and eastern region (32.65 ton). This is mainly related to the economic development, population, and industrial structure optimization of the region. On the one hand, the western region is in the middle and low stage of economic development, which needs to consume a large amount of fossil energy, resulting in a large amount of CO_2 emissions. On the other hand, the highly developed economy in the eastern region promotes technological progress, thus reducing CO_2 emissions. Moreover, in economically developed regions, the service

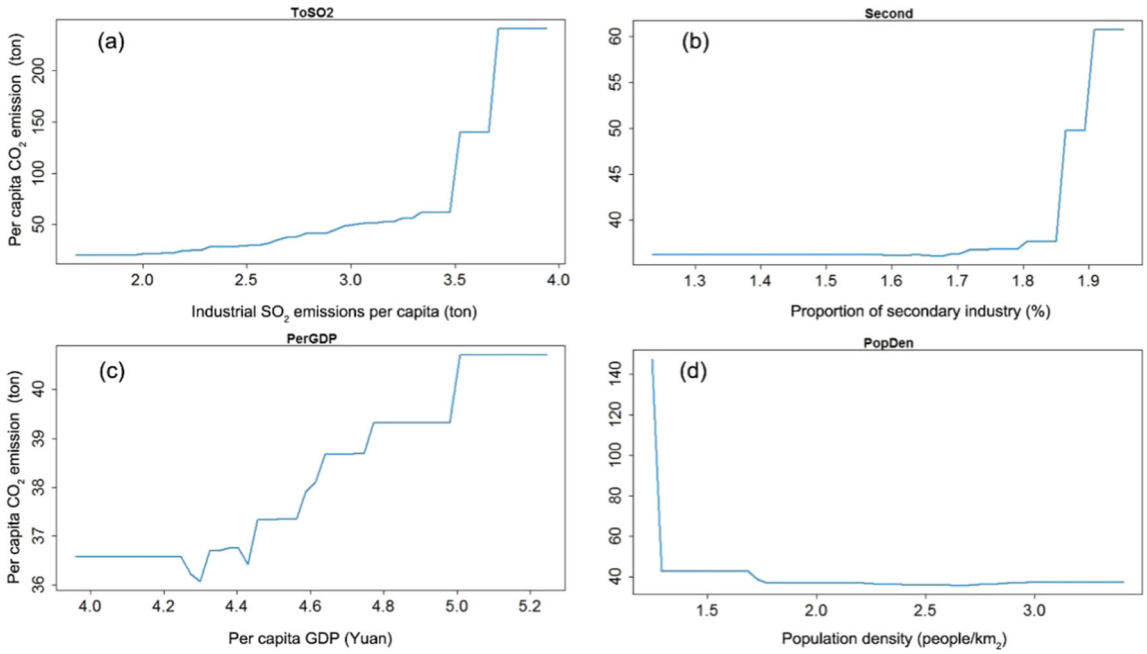

Fig. 6 Partial dependency plots for the four selected variables for the industrial SO_2 emissions per capita (**a**), proportion of secondary industry (**b**), per capita GDP (**c**), and population density (**d**)

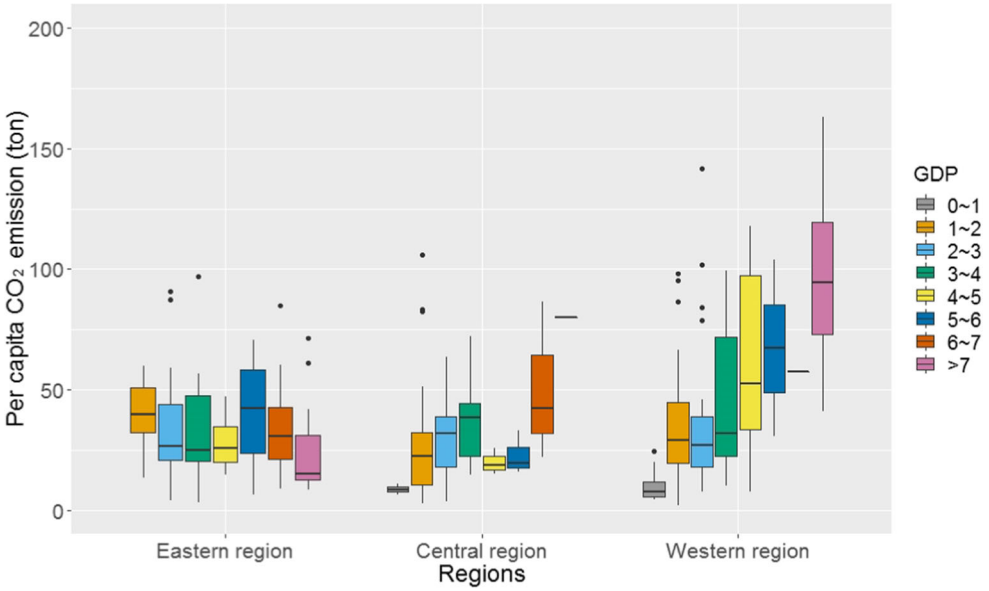

Fig. 7 Boxplots of predicted city-level CO_2 emissions for different per capita GDP class (unit: 10^4 Yuan) across three regions of China

industry is relatively developed. Economic growth driven by the service industry can reduce the demand for heavy industrial products, thus reducing CO_2 emissions (Munir et al. 2018). And the central region lies between the western and eastern regions.

Figure 7 shows that there is no significant difference in distribution tendencies among different income level eastern cities. Per capita CO_2 emissions tended to be the lowest when per capita GDP was higher than 7×10^4 Yuan. Compared to the eastern cities in China, the central and western cities existed large differences between low-, middle-, and high-income cities. We found that per capita CO_2 emissions generally increased with the growth of income-level cities for both central and western regions as descripted in Fig. 7.

The regional differences in industrialization and urbanization can explain this result. On one hand, the eastern region is dominated by technology-intensive industries with relatively mature industrialization. For example, in the developed regions, such as Jiangsu and Zhejiang province, the industrial structure and energy consumption structure have been adjusted, and the technology is advanced. On the other hand, the eastern region is in the later stage of urbanization, and the intensification effect of urbanization is gradually reflected. Therefore, with the rapid economic growth, the growth of carbon emissions has been effectively controlled, and there is no significant difference between eastern cities.

However, there are great differences between industrialization and urbanization in western China. The per capita CO_2 emissions in the more developed regions are relatively high, which is due to the fact that the urbanization level is still in the stage of promoting carbon emissions, a large amount of infrastructure construction consumes a large amount of energy and promotes the CO_2 emissions (Krantz et al. 2015), and their economic development is more dependent on energy-intensive industries. However, the urbanization level in the western low-income areas is very low, and there are few enterprises. Therefore, there are significant differences between different income cities in the central and western regions, especially in the western regions.

Conclusion and policy implications

Understanding and anticipating city-level CO_2 emissions is still a major challenge owing to scarce related data in undeveloped regions. This paper evaluated the abilities of GBM, SVM, RF, and XGBoost algorithms to predict the city-level CO_2 emissions for remote or lacking of energy statistical records cites using a set of open access data. We considered three categories features associated with CO_2 emissions including the socioeconomic and climatic background and urban form–specific characteristics. As verified by RMSE and MAE metrics, the XGBoost model outperformed the other machine learning algorithm (GBM, SVM, and RF). This model demonstrated excellent estimation accuracy with rRMSE and rMAE 0.7~0.8%. Thus, a machine learning-based approach can be used to capture the city-level CO_2 emissions using variables, which are easily collected from official statistics website and open access geospatial data.

The performance of the proposed models depends on the nature of cities' CO_2 emissions and the number of predictors. The amount of CO_2 emissions for energy production and heavy industry cities is always overestimated, while that for high technology and service typed cities were more accurate with RF model. We demonstrated that the industrial atmospheric pollutants emissions (such as SO_2 and dust emissions) are the most important indicators for predicting the city-level CO_2 emissions in China. The population density, industrial

structure, traffic statistics, and climatic background were also recognized as slightly relative influence on variation of city-level CO_2 emissions. Compared to the previous studies, two of nighttime lights intensity indices were excluded from input variables by the Boruta algorithm.

Partial dependency plots (PDPs) is a useful tool to examine relationships between CO_2 emissions and the variables of our interest, while the rest variables are held constant. Results indicated that there exists an S-shaped curve for the relationship between CO_2 emissions and economic development in China based on city-level cross-sectional data, which is inconsistent with the typical inverse U-shaped environmental Kuznets curve (EKC) hypothesis (Wang and Liu 2017; Fujii et al. 2018). Such results indicated there was an uncertain relationship between per capita CO_2 emissions and urban economic development using cross-sectional or time-series data and various approaches. Our findings suggest that per capita CO_2 increased as per capita GDP value increased and then keep stable at a certain level. In addition, a synergistic effect for urban atmospheric pollutants and CO_2 emissions is observed in our study, indicating that a win-win situation would be obtained by adopting effective air environmental protection measures.

Finally, our study provides firstly try to predict undeveloped cities' per capita CO_2 emissions by combing the open access data and machine learning algorithms. The key prediction variables were also sought to simply the modeling process and reduce data collection difficulty. This framework can be applied to link the socioeconomic, climatic, and urban-form metrics and city-level per capita CO_2. A machine learning-based technology can further contribute to the improvement of predicting and modeling city-level CO_2 emission under the condition of related data scarcity.

Based on the above results, we provide some policy implications for policy-makers to reduce emissions in China. On the one hand, optimizing the industrial structure will help reduce energy consumption and CO_2 emissions. Because compared with the primary and tertiary industries, the secondary industry is energy intensive and carbon intensive. The reduction of the secondary industry can help reduce CO_2 emissions. Therefore, we should vigorously support the development of tertiary industry and high-tech industry, increase the proportion of tertiary industry, improve the production efficiency of heavy industry, and reduce energy consumption. Economically developed eastern regions should give full play to their talents and geographical advantages and strengthen cooperation with less developed central and western regions to help them upgrade low-carbon technologies and optimize industrial structure while maintaining economic growth under the CO_2 emission reduction target. The central and western regions need to introduce technical talents actively, upgrade energy-intensive industries, and develop characteristic emerging industries, so as to achieve economic growth and reduce fossil energy consumption.

On the other hand, optimizing the structure of energy consumption and promoting the development of renewable energy is an effective strategy to reduce CO_2 emissions. We should strengthen the low carbon and high efficiency of traditional fossil energy to realize the low carbonization utilization of high carbon energy. Meanwhile, we should also increase the development and utilization of clean energy such as hydropower, solar energy, wind energy, and geothermal energy. For the central region, abundant water resources are its unique advantage (Xie et al. 2020). A large number of hydropower stations could be built. Making full use of hydropower resources can effectively reduce fossil energy consumption. For the western region, it is rich in natural gas resources. For example, Xinjiang and Sichuan are the provinces with the most natural gas resources. We can rationally develop natural gas resources and encourage industrial enterprises to replace traditional coal with natural gas.

Acknowledgments The authors extend their appreciation to the reviewers who gave constructive comments that helped bring considerable improvements to a previous version of the manuscript.

Authors' contributions Ying Li and Yanwei Sun designed this study; Yanwei Sun revised the paper. All authors read and approved the final manuscript.

Funding This work was sponsored by K.C. Wong Magna Fund in Ningbo University and Philosophical and Social Science Planning Foundation of Zhejiang Province (20NDJC077YB) and National Natural Science Foundation of China (41571018 and 41871024).

Data availability Data are available from the authors upon request.

Compliance with ethical standards

Competing interests The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate Not applicable.

Consent to publish All authors read and approved the final manuscript.

References

Alam S, Kumar A, Dawes L (2020) Roughness optimization of road networks: an option for carbon emission reduction by 2030. J Transp Eng B Pavements 146(4):04020062. https://doi.org/10.1061/JPEODX.0000203

Baiocchi G, Creutzig F, Minx J, Pichler P (2015) A spatial typology of human settlements and their CO_2 emissions in England. Glob Environ Chang 34:13–21. https://doi.org/10.1016/j.gloenvcha.2015.06.001

Ballantyne AG, Wibeck V, Neset TS (2016) Images of climate change–a pilot study of young people's perceptions of ICT-based climate

visualization. Clim Chang 134:73–85. https://doi.org/10.1007/s10584-015-1533-9
Bayar Y, Diaconu L, Maxim A (2020) Financial development and CO_2 emissions in post-transition European Union countries. Sustainability 12:2640. https://doi.org/10.3390/su12072640
Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. ArXiv e-prints785–794
Chen X, Zhang S, Ruan SM (2021) Polycentric structure and carbon dioxide emissions: Empirical analysis from provincial data in China. J Clean Prod 278:123411. https://doi.org/10.1016/j.jclepro.2020.123411
China City Statistical Yearbook (2011) China Statistical Press, Beijing, China (in Chinese)
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
Dhakal S (2009) Urban energy use and carbon emissions from cities in China and policy implications. Energ Policy 37:4208–4219. https://doi.org/10.1016/j.enpol.2009.05.020
Fan JL, Wang XK, Wu LF, Zhou HM, Zhang FC, Yu X, Lu XH, Xiang YZ (2018) Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers Manag 164:102–111. https://doi.org/10.1016/j.rser.2018.10.018
Fang DB, Zhang XL, Yu Q, Jin TC, Tian L (2018) A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression. J Clean Prod 173:143–150. https://doi.org/10.1016/j.jclepro.2017.05.102
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232. https://doi.org/10.1214/aos/1013203451
Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378. https://doi.org/10.1016/S0167-9473(01)00065-2
Fujii H, Iwata K, Chapman A, Kagawa S, Managi S (2018) An analysis of urban environmental Kuznets curve of CO_2 emissions: empirical analysis of 276 global metropolitan areas. Appl Energy 228:1561–1568. https://doi.org/10.1016/j.apenergy.2018.06.158
Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422. https://doi.org/10.1023/A:1012487302797
Hammond GP, Norman JB (2012) Decomposition analysis of energy-related carbon emission from UK manufacturing. Energy 41(1):220–227. https://doi.org/10.1016/j.energy.2011.06.035
Han F, Xie R, Lu Y, Fang JY, Liu Y (2018) The effects of urban agglomeration economies on carbon emissions: evidence from Chinese cities. J Clean Prod 172:1096–1110. https://doi.org/10.1016/j.jclepro.2017.09.273
He DQ, Meng F, Wang MQ, He KB (2011) Impacts of urban transportation mode split on CO_2 emissions in Jinan. China. Energies 4:685–699. https://doi.org/10.3390/en4040685
Heun MK, Carbajales-Dale M, Haney BR (2015) Beyond GDP, lectures notes in energy. Springer, Cham. https://doi.org/10.1007/978-3-319-12820-7
Hu M, Li R, You W, Liu YB, Lee CC (2020) Spatiotemporal evolution of decoupling and driving forces of CO_2 emissions on economic growth along the Belt and Road. J Clean Prod 277:123272. https://doi.org/10.1016/j.jclepro.2020.123272
Huang YS, Shen L, Liu H (2019) Grey relational analysis, principal component analysis and forecasting of carbon emissions based on long short-term memory in China. J Clean Prod 209:415–423. https://doi.org/10.1016/j.jclepro.2018.10.128
IPCC (2014) Climate change 2014: synthesis report. Contribution of working groups I, II, III to the fifth assessment report of the intergovernmental panel on climate change. IPCC, Geneva, p 151
Jebli MB, Youssef SB, Ozturk I (2016) Testing environmental Kuznets curve hypothesis: the role of renewable and non-renewable energy consumption and trade in OECD countries. Ecol Indic 60:824–831. https://doi.org/10.1016/j.ecolind.2015.08.031
Jeong K, Hong T, Kim J (2018) Development of a CO_2 emission benchmark for achieving the national CO_2 emission reduction target by 2030. Energy Build 158:86–94. https://doi.org/10.1016/j.enbuild.2017.10.015
Krantz J, Larsson J, Lu W, Olofsson T (2015) Assessing embodied energy and greenhouse gas emissions in infrastructure projects. Buildings 5(4):1156–1170. https://doi.org/10.3390/buildings5041156
Kunnas, Jan, Myllyntaus (2009) Postponed leap in carbon dioxide emissions: the impact of energy efficiency, fuel choices and industrial structure on the Finnish Energy Economy, 1800–2005. Glob Environ 2(3):154–189. https://doi.org/10.3197/ge.2009.020307
Kursa MB, Rudnicki WR (2010) Feature selection with the Boruta package. J Stat Softw 36:1–13. https://doi.org/10.18637/jss.v036.i11
Li Y, Wei YG, Dong Z (2020) Will China achieve its ambitious goal? — forecasting the CO_2 emission intensity of China towards 2030. Energies 13:2924. https://doi.org/10.3390/en13112924
Lin BQ, Wang M (2020) The role of socio-economic factors in China's CO_2 emissions from production activities. Sustain Prod Consump. https://doi.org/10.1016/j.spc.2020.10.029
Lin DT, Zhang LY, Chen C, Lin YY, Wang JK, Qiu RZ, Hu XS (2019) Understanding driving patterns of carbon emissions from the transport sector in China: evidence from an analysis of panel models. Clean Technol Environ 21(6):1307–1322. https://doi.org/10.1007/s10098-019-01707-y
Meng L, Graus W, Worrell E, Huang B (2014) Estimating CO_2 (carbon dioxide) emissions at urban scales by DMSP/OLS (Defense Meteorological Satellite Program's Operational Linescan System) nighttime light imagery: methodological challenges and a case study for China. Energy 71:468–478. https://doi.org/10.1016/j.energy.2014.04.103
Milnar M, Ramaswami A (2020) Impact of urban expansion and in situ greenery on community-wide carbon emissions: method development and insights from 11 US cities. Environ Sci Technol 20. https://doi.org/10.1021/acs.est.0c02723
Munir Q, Lean HH, Smyth R (2018) CO_2 emissions, energy consumption and economic growth in the ASEAN-5 countries: A cross-sectional dependence approach. Energy Econ 85:104571. https://doi.org/10.1016/j.eneco.2019.104571
Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21. https://doi.org/10.3389/fnbot.2013.00021
Pino-Mejías R, Pérez-Fargallo A, Rubio-Bellido C, Pulido-Arcas JA (2017) Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO_2 emissions. Energy 118:24–36. https://doi.org/10.1016/j.energy.2016.12.022
R Development Core Team (2008) R: a language and environment for statistical computing. Retrieved from. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org. Accessed 25 July 2020
Seto KC, Dhakal S, Bigio A, Blanco H, Delgado GC, Dewar D, Huang L, Inaba A, Kansal A, Lwasa S, McMahon J, Mueller D, Murakami J, Nagendra H, Ramaswami A (2014) Human settlements, infrastructure and spatial planning, climate change 2014: mitigation of climate change. In: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
Shan YL, Guan DB, Klaus H et al (2018) City-level climate change mitigation in China. Sci Adv 4(6):0390. https://doi.org/10.1126/sciadv.aaq0390
Shi KF, Xu T, Li YQ, Chen ZQ, Gong WK, Wu JP, Yu BL (2020) Effects of urban forms on CO_2 emissions in China from a multi-perspective

analysis. J Environ Manag 262:110300. https://doi.org/10.1016/j.jenvman.2020.110300

Sun W, Wang YW, Zhang CC (2018) Forecasting CO_2 emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine. Environ Sci Pol 25: 28985–28997. https://doi.org/10.1007/s11356-018-2738-z

Taylor JR (2015) The China dream is an urban dream: assessing the CPC's national new-type urbanization plan. J Chin Polit Sci 20: 107–120. https://doi.org/10.1007/s11366-015-9341-7

Wang S, Liu X (2017) China's city-level energy-related CO_2 emissions: spatiotemporal patterns and driving forces. Appl Energy 200:204–214. https://doi.org/10.1016/j.apenergy.2017.05.085

Wang SJ, Liu XP, Zhou CS, Hu JC, Ou JP (2017) Examining the impacts of socioeconomic factors, urban form, and transportation networks on CO_2 emissions in China's megacities. Appl Energy 185:189–200. https://doi.org/10.1016/j.apenergy.2016.10.052

Wasti SKA, Zaidi SW (2020) An empirical investigation between CO_2 emission, energy consumption, trade liberalization and economic growth: a case of Kuwait. J Build Eng 28:101104. https://doi.org/10.1016/j.jobe.2019.101104

Xie X, Jiang X, Zhang T, Huang Z (2020) Study on impact of electricity production on regional water resource in China by water footprint. Renew Energy 152:165–178. https://doi.org/10.1016/j.renene.2020.01.025

Zhang N, Yu K, Chen Z (2017) How does urbanization affect carbon dioxide emissions? A cross-country panel data analysis. Energ Policy 107:678–687. https://doi.org/10.1016/j.enpol.2017.03.072

Zhao JC, Ji GX, Yue YL, Lai ZZ, Chen YL, Yang DY, Yang X, Wang Z (2019) Spatio-temporal dynamics of urban residential CO_2 emissions and their driving forces in China using the integrated two nighttime light datasets. Appl Energy 235:612–624. https://doi.org/10.1016/j.apenergy.2018.09.180