



ARTICLE



<https://doi.org/10.1057/s41599-025-05150-x>

OPEN

Manufacturing agglomeration and carbon emissions: an ensemble learning approach with evidence from South Korea

Zhen Wu¹, Su-Han Woo², Pairach Piboonrungrroj³✉ & Po-Lin Lai²✉

This study calculates two key manufacturing agglomeration indices, namely the specialization index and the diversification index, across different regions of South Korea. By leveraging random forest and gradient boosting decision tree models, it addresses a critical gap in understanding how the interplay between manufacturing agglomeration patterns affects carbon emissions. The findings indicate that specialized agglomeration significantly increases carbon emissions, whereas diversified agglomeration has a clear emission reduction effect, particularly in its early stages. Notably, a trade-off exists between the two agglomeration patterns, exerting a complex and evolving influence on carbon emissions. Regions with high specialization and low diversification tend to exhibit higher carbon emissions, while regions with high diversification and low specialization are associated with lower emissions. The study further reveals that introducing a specialized agglomeration development model in regions already characterized by diversified agglomeration leads to a significant increase in carbon emissions, especially during the early stages of diversification. Conversely, introducing a diversified agglomeration model in regions dominated by specialization reduces carbon emissions; however, as specialization intensifies, the emission reduction effects of diversification diminish considerably. Additionally, the influence of the interaction between these two agglomeration patterns on carbon emissions gradually weakens over time. This research provides new insights into the relationship between manufacturing agglomeration and carbon emissions, offering a valuable framework for evaluating policies and guiding sustainable industrial planning.

¹ Department of Business Administration, Honam University, Gwangju, Republic of Korea. ² Department of International Logistics, College of Business and Economics, Chung-Ang University, Seoul, Republic of Korea. ³ Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand. ✉email: pairach.p@cmu.ac.th; polin@cau.ac.kr

Introduction

The interplay between manufacturing agglomeration and carbon emissions has become a focal point in academic research, given its dual implications for economic development and environmental health. At its core, manufacturing agglomeration—defined by the spatial concentration of industrial activities—creates a paradox: while it drives economic productivity, its environmental consequences are far from straightforward (Lazăr et al., 2018). On the negative side, the sheer scale of production in densely industrialized zones leads to heightened fossil fuel consumption, particularly coal and oil, exacerbating pollution levels. Yet, counterintuitively, industrial clustering may also foster environmental benefits. Drawing on agglomeration theory (Krugman et al., 1998) and new economic geography (Enrenfeld et al., 2003), researchers argue that concentrated industrial activity can accelerate the adoption of clean technologies, spur efficiency gains in resource use, and ultimately curb carbon emissions. This duality is further complicated by the distinction between *specialized* and *diversified* agglomeration—a conceptual framework introduced by Hoover (1937). Specialized agglomeration, where firms within the same sector cluster together, generates industry-specific externalities. Diversified agglomeration, in contrast, arises from the co-location of firms across different sectors, fostering cross-industry spillovers. The environmental impacts of these two forms of agglomeration remain a contested topic. Some scholars, like Enrenfeld et al. (2003), argue that diversified agglomeration enhances regional resource circulation and utilization efficiency, leading to lower emissions. Capello et al. (2015) support this view, noting that cities with diversified industrial bases tend to share infrastructure more effectively, reducing waste and energy use. However, Lan et al. (2021) present a contrasting finding: in their study, only specialized agglomeration had a measurable effect on regional carbon emissions, while diversified agglomeration showed no significant impact. Adding another layer of complexity, recent research suggests that these relationships are not linear but instead follow an inverted U-shaped trajectory. Han et al. (2018) observed this pattern in China, where both specialized and diversified agglomeration initially increased emissions but eventually helped reduce them after reaching a certain threshold. A similar dynamic was documented in South Korea by Wu et al. (2024), reinforcing the idea that agglomeration's environmental effects are highly context-dependent and nonlinear. This evolving body of research underscores a key insight: the environmental outcomes of industrial clustering are neither uniform nor predictable. They hinge on the type of agglomeration, regional economic structures, and the stage of industrial development—making this an area ripe for further investigation.

While existing studies have examined the impact of manufacturing agglomeration on carbon emissions, they have largely focused on the isolated effects of individual agglomeration patterns, neglecting the interactions between different types of agglomeration patterns and their influence on carbon emissions. This leaves a critical research gap in understanding the joint impacts of specialized and diversified agglomeration on carbon emissions. Understanding the dual and potentially conflicting impacts of these agglomeration patterns is critical for formulating effective environmental policies. Moreover, there is currently a lack of appropriate models that reflect the joint effects of these two types of agglomeration on emission reduction.

Based on the above, this study aims to investigate the mechanisms underlying the joint impact of manufacturing specialized and diversified agglomeration on carbon emissions and presents the following research question: Do the two agglomeration patterns interact and influence each other? How does this interaction affect carbon emissions? Does the impact of this

interaction on carbon emissions evolve over time? To tackle this question, this paper introduces two key agglomeration indices—specialization and diversification—and employs ensemble learning methods, specifically Random Forest (RF) and Gradient Boosting Decision Trees (GBDT), to build a comprehensive prediction model for carbon emissions. The study also utilizes partial dependence plots to visualize the joint impacts of these agglomeration patterns and illustrate how carbon emissions evolve in response to their interactions. This approach offers a novel framework for analyzing the joint impacts of agglomeration patterns, going beyond traditional econometric models that only focus on single-variable impacts.

This study selects South Korea as a case study. South Korea provides an ideal case for examining the research question due to its rapid industrialization, energy-intensive manufacturing sector, and distinct geographic clustering of industries. Since the 1970s, South Korea has developed prominent manufacturing agglomerations, such as the automobile hub in Ulsan, the steel base in Pohang, and the electronics industry in Gumi. These clusters have been critical drivers of economic growth, but have also contributed significantly to the country's greenhouse gas emissions. In 2017, South Korea ranked fifth among OECD countries in carbon emissions, with manufacturing responsible for most of the carbon emissions (Jung and Park, 2000; Kim et al., 2010). The South Korean government has pledged to reduce greenhouse gas emissions by 728 million tons by 2050 to achieve a carbon-neutral economy (Lee, 2021). Given this challenge, South Korea offers a compelling context to explore the nuanced effects of manufacturing agglomeration on carbon emissions. The insights from South Korea's case study have broader applicability to other regions and countries, particularly those undergoing industrialization or grappling with the trade-offs between economic development and environmental sustainability.

The subsequent sections of this manuscript are organized as follows: The "Literature review" section reviews pertinent theoretical frameworks and empirical research. The "Methodology and data" section delineates the methodology and data employed in this investigation. The "Results" section articulates the empirical findings. The "Conclusion" section synthesizes the conclusions of the study, while the "Discussion" section examines the implications of the findings and delineates the limitations of the research.

Literature review

The relationship between industrial agglomeration and carbon emissions. The relationship between manufacturing agglomeration and carbon emissions has attracted significant academic attention due to its implications for economic growth and environmental sustainability. Manufacturing agglomeration, characterized by the geographic clustering of industrial activities, produces both positive and negative environmental externalities (Lazăr et al., 2018). On the negative side, industrial agglomeration leads to the massive use of fossil fuels, such as coal and oil, thereby contributing to increased carbon emissions and severe environmental pollution (Verhoef and Nijkamp, 2002; Duc, 2007). However, agglomeration theories, such as traditional agglomeration theory (Krugman et al., 1998) and new economic geography theory (Enrenfeld et al., 2003), argue that industrial clustering promotes scale effects, technological spillovers, and clean technology adoption, which improve resource efficiency and reduce carbon emissions.

Hoover (1937) divides agglomeration economies into specialized agglomeration (enterprises in the same industry) and diversified agglomeration (enterprises across different industries).

These two types of agglomeration have distinct and dynamic environmental impacts. Specialized agglomeration tends to focus on resource-intensive industries, which significantly increase regional carbon emissions (Lan et al., 2021). Conversely, diversified agglomeration promotes resource recycling, enhances resource utilization efficiency, and reduces emissions by creating symbiotic relationships among firms (Enrenfeld et al., 2003; Han et al., 2018). Diversified clusters are also better positioned to share low-carbon infrastructure, improving overall energy efficiency and reducing emissions (Capello et al., 2015).

The connection between industrial agglomeration and carbon emissions rarely follows a simple, straight-line trajectory—instead, research increasingly points to complex, nonlinear dynamics. Take, for instance, the work of Han et al. (2018) and Wu et al. (2024), who uncovered an inverted U-shaped pattern in China and South Korea. Their findings reveal a critical tipping point: both specialized and diversified agglomeration initially drive emissions upward, only to eventually help pull them down after crossing a certain density threshold. This phenomenon isn't limited to industrial clustering alone. Li (2014) observed a similar delayed-benefit effect in marketization processes—where weak market structures amplify agglomeration's environmental harm, but beyond a certain level of economic maturity, the relationship flips, yielding net positive outcomes. Even urban spatial structures exhibit this nonlinearity. Qin and Wu (2015) demonstrated how city clusters follow an inverted U-curve with emissions, while Xia et al. (2022)'s analysis of China's fiscal policies revealed a U-shaped pattern, suggesting that decentralization's environmental impact worsens before improving. What emerges from these studies is a clear, if counterintuitive, insight: agglomeration's environmental effects often get worse before they get better. The key question for policymakers, then, isn't just whether to encourage industrial clustering—but how to navigate that precarious transition phase where the benefits finally outweigh the costs.

Industrial agglomeration presents a fascinating paradox in environmental economics. On one hand, the geographic concentration of industries creates powerful positive externalities. As Marshall (2009) first articulated, these benefits emerge through three key channels: pooled labor markets, shared input suppliers, and—perhaps most crucially—knowledge spillovers. Subsequent research by Han et al. (2014) has shown how these mechanisms collectively boost energy efficiency and accelerate the adoption of clean technologies. The environmental advantages appear particularly pronounced in diversified agglomerations, where Capello (2007) and Enrenfeld et al. (2003) documented enhanced resource recycling and more efficient shared infrastructure for emission control.

Yet this bright picture has its shadows. As Poumanyvong and Kaneko (2010) caution, excessive clustering can trigger negative externalities like congestion effects and resource depletion. The environmental impact becomes even more complex when we consider spatial spillovers. Carbon emissions notoriously respect no administrative boundaries, spreading through industrial transfers, economic linkages, and even atmospheric circulation (Han et al., 2018). Recent work by Lan et al. (2021) reveals an intriguing spatial dichotomy: while specialized agglomeration mainly affects local emissions, diversified clustering generates beneficial spillovers that reduce emissions in neighboring regions. This spatial dimension extends to financial agglomeration as well, with Yuan et al. (2020) demonstrating how high-density financial clusters disproportionately boost regional green development through spillover effects.

The relationship evolves not just across space, but through time. Carbon emissions represent the cumulative legacy of past industrial decisions, creating complex temporal dynamics. Lei

et al. (2017) uncovered a striking temporal disconnect - while industrial clustering shows immediate pollution reduction effects, these benefits appear transient with no lasting causal relationship. Du et al. (2018) similarly found China's low-carbon progress fluctuating unpredictably, buffeted by shifting socioeconomic conditions, resource availability, and regional disparities. These patterns align with Jircikova et al. (2013)'s life cycle theory of industrial clusters, which posits that agglomeration effects vary significantly across different evolutionary stages—from initial R&D efficiency gains to mature-phase resource optimization. This temporal variability underscores Wang et al. (2022)'s argument for proactive urban cluster planning to harness scale economies while minimizing carbon footprints.

Despite growing scholarly attention, critical gaps remain in our understanding. While numerous studies have confirmed agglomeration's impact on emissions, surprisingly, none have incorporated this factor into predictive carbon emission models. The literature has also tended to examine specialized and diversified agglomeration in isolation, neglecting their potential interactions within real urban systems where both patterns inevitably coexist. This coexistence may involve either symbiotic reinforcement or competitive exclusion, with significant but unexplored implications for emission outcomes. To address these gaps, we propose three key hypotheses:

- H1: Specialized and diversified agglomeration patterns interact dynamically within urban systems
- H2: Their interaction produces variable impacts on carbon emissions
- H3: These impacts evolve nonlinearly over time

Influencing factors of carbon emissions and carbon emissions forecasting. Carbon emission studies typically fall into two distinct but complementary camps. The first seeks to understand what drives emissions, while the second uses these insights to forecast future emission patterns (Wei et al., 2018). Spatial analysis has proven particularly valuable in this regard. Chuai's (2012) spatial error model work revealed how energy-related emissions closely track with economic output and population density. Digging deeper into China's experience, Cheng et al. (2015) employed a spatial Durbin model to identify four key levers: energy mix, industrial structure, energy efficiency, and urbanization pace. But numbers only tell part of the story - public sentiment matters too. Li et al.'s (2023) analysis of Chinese social media uncovered widespread optimism about carbon neutrality's potential to improve quality of life, tempered by concerns about economic trade-offs that policymakers can't afford to ignore.

Traditional econometric approaches often stumble over collinear variables, but machine learning thrives where others struggle. As Moore et al. (1991) first recognized, these algorithms can cut through tangled datasets to reveal stable predictive patterns. Researchers have since harnessed this power across diverse contexts. In China's Yangtze River Economic Belt, Wang et al. (2021) used random forests to uncover how emission drivers shift regionally—from urban density factors in some areas to industrial mix in others. Wei et al. (2018) pushed further by blending random forests with extreme learning machines, creating hybrid models that outperform conventional approaches. Even our built environment yields insights when viewed through this lens—Lin et al. (2021) demonstrated how building density, height distribution, and spatial arrangement significantly impact urban carbon footprints.

The global picture comes into focus through neural network analyses. Alex et al. (2019) tracked eight economic indicators across five nations, revealing how each country's emission sensitivity varies dramatically. Jena et al. (2021) achieved a

remarkable 96% prediction accuracy across 17 economies using just three key variables. Perhaps most ambitiously, Xu et al. (2019) built a dynamic neural network for China that accounts for emission persistence effects while ranking seven critical factors—with industrialization and energy factors leading the pack.

Despite these advances, a glaring omission persists in emission modeling. While studies routinely include economic output, energy use, and population metrics, they consistently overlook how industrial clustering patterns shape emission outcomes. This oversight matters because, as Lan et al. (2021) and Wu et al. (2024) show, specialized industrial districts often become emission hotspots by concentrating polluting activities. Conversely, diversified clusters can act as environmental balancers (Enrenfeld et al., 2003; Capello et al., 2015), offsetting high-emission sectors with cleaner ones.

What makes agglomeration metrics uniquely valuable is their ability to capture both the spatial concentration and sectoral mix of economic activity - two dimensions that fundamentally influence regional emission profiles but remain absent from current models. By incorporating these measures, we can achieve more nuanced predictions that better reflect real-world economic geography. This integration promises to sharpen our forecasting tools while giving policymakers new levers for sustainable development strategies.

Methodology and data

Utilizing panel data from 17 regions in South Korea spanning the years 2013 to 2019, this research seeks to incorporate manufacturing agglomeration as a significant variable within the carbon forecasting model. Considering that there may be obvious collinearity in the two kinds of agglomeration index, this study adopts the Random Forest (RF) model and Gradient Boosting Decision Tree model (GBDT) to establish prediction models. Based on the ensemble learning method, this study also differentiates and ranks the importance of the factors influencing carbon emissions. Finally, partial dependence plots derived from the Ensemble Learning model are used to analyze the individual and joint impacts of specialized and diversified agglomeration on carbon emissions.

Regression tree. The classification and regression tree (CART) model is the most widely used decision tree learning method. The generation of a decision tree is a process of recursively constructing a binary tree. The features must be divided into multiple spaces when constructing a binary tree. If the value of the feature space is divided into M units ($S_1, S_2, \dots, S_M$), the output value of each unit S_m is C_m and the regression tree model can be expressed as $f(x) = \sum_{m=1}^M C_m I(x \in S_m)$. When dividing the features, the i th variable and its value a are selected as the split variable and point, respectively. Thus, two regions, $S_1 = (x_i \leq a)$ and $S_2 = (x_i > a)$, can be defined. The minimum mean square error principle can be used to determine the optimal feature and optimal split point. The optimal value of C_m for each S_m can be obtained by minimizing the squared error, $C_m^* = \min \frac{1}{n} \sum_{X_i \in S_m} (y_i - C_m)^2$. By taking the derivative of C_m and setting the derivative to 0, the formula for calculating C_m^* can be obtained:

$$\frac{\partial \sum_{X_i \in S_m} (y_i - C_m)^2}{\partial C_m} = N_m C_m - \sum_{X_i \in S_m} y_i = 0, C_m^* = \frac{\sum_{X_i \in S_m} y_i}{N_m} \quad (1)$$

Therefore, the optimal value of C_m for units S_m and C_m^* is the mean value of y_i corresponding to all x_i . After C_m^* is obtained, all features and split points are traversed to find the optimal split

feature x_i and split point a as follows:

$$\min_{i(a)} [\min \sum_{X_i \in S_1} (y_i - \bar{y})^2 + \min \sum_{X_i \in S_2} (y_i - \bar{y})^2] \quad (2)$$

The above split process is repeated for each S_m until the preset conditions are met. This process was repeated to produce a regression tree. Regression trees divided according to the principle of minimum square error are also called least-squares regression trees. Because no assumptions are made in advance about the nature of the relationships between variables, the regression tree allows for the possibility of interaction and nonlinearity between variables (Moore et al., 1991). The classification and regression tree (CART) model inherently mitigate collinearity issues through their tree-based structures, as they split features hierarchically based on importance rather than relying on linear relationships. This reduces the impact of multicollinearity on model performance. Therefore, regression tree analysis has obvious advantages over classical statistical methods, such as regression analysis (Prasad et al., 2006). In addition, regression trees can be used as base learners in ensemble learning (Wei et al., 2018), artificial neural networks (Xu et al., 2019; Wen and Yuan, 2020) and other machine-learning methods.

Bagging and random forest. The basic idea of bagging is to recognize that the output errors of a single regression tree are related to the selection of training datasets (Prasad et al., 2006). Therefore, if several similar datasets are created by resampling (bootstrapping), and the results of multiple regression trees (without pruning) are averaged, the variance of the prediction model can be effectively reduced (Breiman, 1996a; Buhlmann and Yu, 2002). Additionally, because bagging does not prune when building regression trees, it can reduce the bias of each regression tree. Therefore, the bagging algorithm can reduce the variance and bias simultaneously, thus reducing the total mean squared error (MSE). The bagging algorithm first carries out sampling of sample data with put back and obtains M bootstrap samples, each of which has n observation values (n = sample size). The i th bootstrap sample can be expressed as $\{x_i^m, y_i^m\}$, $m = 1, \dots, M, i = 1, \dots, n$. Different regression trees were estimated according to each bootstrap sample without pruning. Thereafter, the results of the M regression trees were averaged: $f(x) = \frac{1}{M} \sum_{m=1}^M f(x)_m$. The probability that an observation is never picked is $(1 - \frac{1}{n})^n$ in the n samples that have been returned. The limit value of this probability is approximately $e^{-1} = 37\%$ (Breiman, 1996a). Therefore, in n bootstrapping samples, the proportion of observations that do not appear in the bootstrap sample is 37%. These observations are called out-of-bag (OOB) observations, which constitute a verification set for calculating out-of-bag errors. The out-of-bag error can be expressed as $MSE_{OOB} = \frac{1}{n} \sum_{i=1}^n [y_i - f(x)_{i, OOB}]^2$.

Although the bagging method can reduce the variance of the sample mean, if each bootstrapping sample is correlated, then the variance of the sample mean may not be reduced to the ideal level. There are three main reasons for the strong correlation between bootstrapping samples in the bagging algorithm. First, each bootstrapping sample uses the same base learner (e.g., a regression tree). Second, bootstrapping samples are all from the same original data; therefore, there must be a relatively high correlation. Third, each regression tree uses the same features as the training data, and the correlation between the regression trees may be strong. To reduce the correlation between base learners, Breiman (2001) proposed an RF algorithm. Based on the bagging algorithm, the RF model selects only some features as candidate splitting variables for each decision tree (regression tree). In an RF algorithm, each tree is grown with a randomized subset of

predictors, hence the name “random” forests. The number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors. Similar to bagging, RF algorithms allow regression trees to grow to their maximum size and aggregation is achieved by averaging the trees. Using only part of the features at each split node of the regression tree may increase the estimation bias. However, splitting each regression tree with different features can reduce the correlation between regression trees and variance. In addition, in the estimation of multiple regression trees, the unselected characteristic variables can also be selected in other regression trees. Therefore, the RF algorithm achieves a larger decrease in variance by increasing the bias by a small amount, thereby reducing the total MSE. Owing to the growth of many trees in the RF algorithm, there is a limited generalization error, which means that no overfitting is possible. This is an important advantage of the RF algorithm (Prasad et al., 2006). Random forest is one of the most advanced ensemble learning methods and is widely used in various classification and regression analyses (Gislason et al., 2006; Lin et al., 2020).

Boosting tree and GBDT. Another method of ensemble learning is where a decision tree (classification tree or regression tree) as the base learner is called the boosting tree. For regression problems, the boosting tree algorithm constructs a binary regression tree with $x_i \leq a$ or $x_i > a$ as the split point. This form can be regarded as a simple decision tree with a root node connecting the two leaves, also known as a stump. The boosting tree is an additive model, such as the AdaBoost algorithm, and can be calculated using forward stagewise additive modeling. However, instead of adjusting the weight of each regression tree in each iteration, such as AdaBoost, the boosting tree algorithm makes the new regression tree fit the residuals of the previous tree. In the boosting tree algorithm, the weight of each tree is fixed as 1, and the boosting tree model can be expressed as $f_M(x) = \sum_{m=1}^M T(x, \theta_m)$, $m = 1, \dots, M$, where $T(x, \theta_m)$ represents the m th tree and θ_m represents the parameter of the m th tree (feature space division s_m and the value C_m of each feature space).

The m th step of the boosting tree can be expressed as $f_m(x) = f_{m-1}(x) + T(x, \theta_m)$. As $f_{m-1}(x)$ is a fixed constant, the loss function for phase m can be obtained using the following formula:

$$\theta_m = \arg \min \sum_i L[y_i, f_{m-1}(x_i) + T(x_i, \theta_m)], i = 1, \dots, n \quad (3)$$

Because the regression tree adopts MSE as the loss function, the loss function of the boosting tree can be changed to:

$$\sum_i L[y_i, f_{m-1}(x_i) + T(x_i, \theta_m)] = \sum_i [y_i - f_{m-1}(x_i) - T(x_i, \theta_m)]^2 = \sum_i [r_m - T(x_i, \theta_m)]^2 \quad (4)$$

Where, r_m is the residual of the proposed model. Therefore, the boosting tree of the regression analysis only needs to fit the residuals of the current model. Generally, if the loss function is not MSE, an approximation of the residual can be obtained using a gradient boosting machine (Friedman, 2001). The GBDT can be obtained by the following derivation. First, in the boosting algorithm, the increase in the regression tree must reduce the overall loss function, namely, $L[y, f_{m-1}(x_i)] - L[y, f_m(x_i)] \geq 0$. Thereafter, the first-order Taylor expansion of $L[y, f_m(x_i)]$ can be obtained as:

$$L[y, f_m(x_i)] \approx L[y, f_{m-1}(x_i)] + \left\{ \frac{\partial L[y, f_m(x_i)]}{\partial f_m(x_i)} \right\}_{f_m(x_i)=f_{m-1}(x_i)} \times [f_m(x_i) - f_{m-1}(x_i)] \quad (5)$$

$$L[y, f_{m-1}(x_i)] - L[y, f_m(x_i)] \approx \left\{ \frac{-\partial L[y, f_m(x_i)]}{\partial f_m(x_i)} \right\}_{f_m(x_i)=f_{m-1}(x_i)} \times T(x_i, \theta_m) \quad (6)$$

Therefore, when $T(x_i, \theta_m) \approx \left\{ \frac{-\partial L[y, f_m(x_i)]}{\partial f_m(x_i)} \right\}_{f_m(x_i)=f_{m-1}(x_i)}$, the $L[y, f_{m-1}(x_i)] - L[y, f_m(x_i)] \geq 0$ is valid. The contribution of the GBDT is to fit the residual through the negative gradient of the loss function and find a general method to fit the residual.

Variable measurement and data description. Referring to the relevant literature, this study proposed the following variables to establish the model: All data were defined at the metropolitan or provincial levels. It is well known that manufacturing is more energy intensive than any other industry (Boyd et al., 1987). In South Korea, the energy intensity of the manufacturing industry has deteriorated since the mid-1980s because of the expansion of energy-intensive industries, such as steel, cement, and petrochemicals (Korea Energy Economics Institute, 1997). This phenomenon is unique compared with the energy intensity trends in other countries. Because the manufacturing industry accounts for the largest proportion of carbon emissions in South Korea, the carbon emissions of 25 manufacturing sectors in 17 major regions (metropolitan and provincial), published by the Korea Energy Agency, were used as indicators to measure carbon emissions (I). This index is a useful match for the specialized and diversified agglomeration indices of manufacturing.

The specialized agglomeration index considered in this study is the proportion of manufacturing employment in region i divided by the proportion of manufacturing employment at the national level, expressed in the following equation:

$$si_i = \frac{emp_{i,m}/emp_i}{emp_m/emp} \quad (7)$$

where $emp_{i,m}$ and emp_m are manufacturing employment in region i and South Korea, respectively. emp_i and emp are total employment in region i and South Korea, respectively (Combes, 2000). For diversified agglomeration, we adopted the improved method referred to by Henderson et al. (1995):

$$di_i = \sum_s \frac{emp_{i,s}}{emp_i} \left[\frac{1/\sum_{s^*=1, s^* \neq s}^n (emp_{i,s^*}/(emp_i - emp_{i,s}))^2}{1/\sum_{s^*=1, s^* \neq s}^n (emp_{s^*}/(emp - emp_s))^2} \right] \quad (8)$$

where di_i represents the diversified agglomeration level of the manufacturing industry in region i . The larger the value, the higher the diversified agglomeration level. Here, $emp_{i,s}$ represents the employment of manufacturing sector s in region i ; emp_i represents the total quantity of employment in region i ; emp_{i,s^*} represents the quantity of employment of sector s^* excluding sector s in region i ; emp_{s^*} represents the quantity of employment in sector s^* excluding sector s in South Korea; emp_s is the employment in the manufacturing sector s in Korea, where emp represents the total quantity of national employment.

Six control variables were included in this study. First, we used the total regional (metropolitan or provincial) population to measure the population size (po). Second, we used per capita affluence measured regional GDP per capita ($pergdp$). The two variables in question are indicative of the STIRPAT framework posited by Dietz and Rosa (1994). Furthermore, a plethora of studies have established a correlation between carbon emissions and local human capital (Ang, 2009; Han, 2018). In this research, human capital is quantified through the proportion of individuals

Table 1 Definition of the variables.

Variable	Mean	Sdt. dev.	Min	Max
<i>l</i> (carbon emission, 10 ³ ton)	19,272.23	22,510.06	95.6	77,934.1
<i>si</i> (specialized agglomeration, %)	1.049	0.486	0.221	1.865
<i>di</i> (diversified agglomeration, %)	0.3030	0.2865	0.1075	1.349
<i>po</i> (population size in districts, people)	3,037,872	3,175,695	122,153	1.30e+07
<i>pergdp</i> (GDP per capita, won/person)	3.49e+07	1.07e+07	1.90e+07	6.50e+07
<i>edu</i> (the proportion of Bachelor, Master and PhD graduates, %)	0.009	0.004	0.003	0.024
<i>secondrate</i> (the proportion of secondary industry employment, %)	0.398	0.072	0.222	0.588
<i>tecgdp</i> (technological innovation, R&D costs/GDP, %)	0.035	0.041	0.007	0.193
<i>fdi</i> (foreign direct investment, 10 ⁶ dollar)	789.9751	1621.377	0	8339

Table 2 Correlation analysis.

	<i>si</i>	<i>di</i>	<i>po</i>	<i>pergdp</i>	<i>edu</i>	<i>secondrate</i>	<i>tecgdp</i>	<i>fdi</i>
<i>si</i>	1							
<i>di</i>	-0.786	1						
<i>po</i>	-0.022	0.076	1					
<i>pergdp</i>	0.552	-0.202	-0.023	1				
<i>edu</i>	-0.210	0.002	-0.188	0.034	1			
<i>secondrate</i>	-0.178	0.125	-0.216	0.454	0.186	1		
<i>tecgdp</i>	-0.130	-0.070	0.202	-0.168	0.331	-0.213	1	
<i>fdi</i>	-0.257	0.423	0.655	0.194	0.029	0.054	-0.008	1

holding bachelor's, master's, and doctoral degrees (*edu*). Additionally, the agglomeration of specialized and diversified manufacturing is influenced by the industrial structure, which exerts varying impacts on environmental pollution across different phases of economic development. Given that environmental pollution predominantly arises from waste generated during industrial production, this study employs the ratio of employment in the secondary sector to total local employment as a metric for industrial structure (*secondrate*) (Poumanyong, 2010; Han, 2018).

Scientific and technological advancements are intrinsically linked to environmental degradation. The infusion of innovation in science and technology plays a pivotal role in optimizing the production methodologies and structures of enterprises, thereby facilitating a reduction in environmental pollution, particularly through advancements in green technologies. This research utilized the ratio of research and development (R&D) expenditures from key institutions, such as universities and research organizations, relative to the GDP of various regions as a metric for assessing the level of scientific and technological innovation (*tecgdp*) (Zhou, 2016). Furthermore, foreign direct investment (FDI) typically provides essential capital and employment opportunities for the local economy. FDI is also recognized as a significant enabler of technology transfer and diffusion, assisting local governments in enhancing technological capabilities to achieve carbon emission reductions (Bwalya, 2006). Consequently, this study also examines the influence of FDI on carbon emission mitigation.

This study examines 17 primary administrative regions in South Korea, comprising one special city, seven metropolitan cities, and nine provinces, over the period from 2013 to 2023. These regions were selected due to their extensive data availability and varied economic development and industrial frameworks, rendering them an optimal sample for investigating the correlation between manufacturing agglomeration patterns and carbon emissions. The primary variable data were obtained from the Korea Statistical Information Service (KOSIS) and the Korea Energy Agency. A comprehensive overview of all variables is presented in Table 1.

Results

Correlation analysis. Before building the model, a correlation coefficient analysis was conducted to investigate whether there was strong collinearity among features (Table 2). The results showed that the correlation coefficient between *si* and *di* was high (-0.787), but the other main features were not highly correlated. A significant inverse relationship was observed between *si* and *di*. This reveals a trade-off between the two manufacturing agglomeration modes in South Korea. Appendix A lists the index of specialized and diversified agglomeration in 17 of the main administrative regions of South Korea in 2019; Seoul and Jeju have a higher diversified agglomeration index; the areas with a higher specialized agglomeration index are concentrated in the industrial complex on the southeast coast of South Korea.

In addition, owing to the high correlation coefficients of *si* and *di*, the inclusion of these two features as independent variables in the model may lead to collinearity, which leads to bias in the estimation results of the traditional panel data model. This study also tried to use the fixed effects model for analysis, and the results showed that the coefficient of specialization agglomeration was not significant, which may be caused by collinearity (Appendix B). The ensemble learning method is not affected by collinearity because it uses a tree model as the base learner. Therefore, the influence of features on the target can be correctly captured. This is also one of the reasons that ensemble learning was used for estimation in this study.

Determining the hyperparameters of the model. Before establishing the ensemble learning model, a hyperparameter optimization (HPO) was conducted to determine the hyperparameters that provide the best prediction effect. For the RF algorithm, increasing the number of regression trees (bootstrap samples) generally does not cause overfitting. This was because the predictions of the multiple regression trees were averaged. Therefore, we selected an appropriate number of regression trees (NT) that did not decrease the root mean square error (RMSE) (Wang, 2021). Although NT can control the overall learning ability of the RF model, the learning ability of each weak learner in the RF model is determined by the depth of the regression tree (DT). The

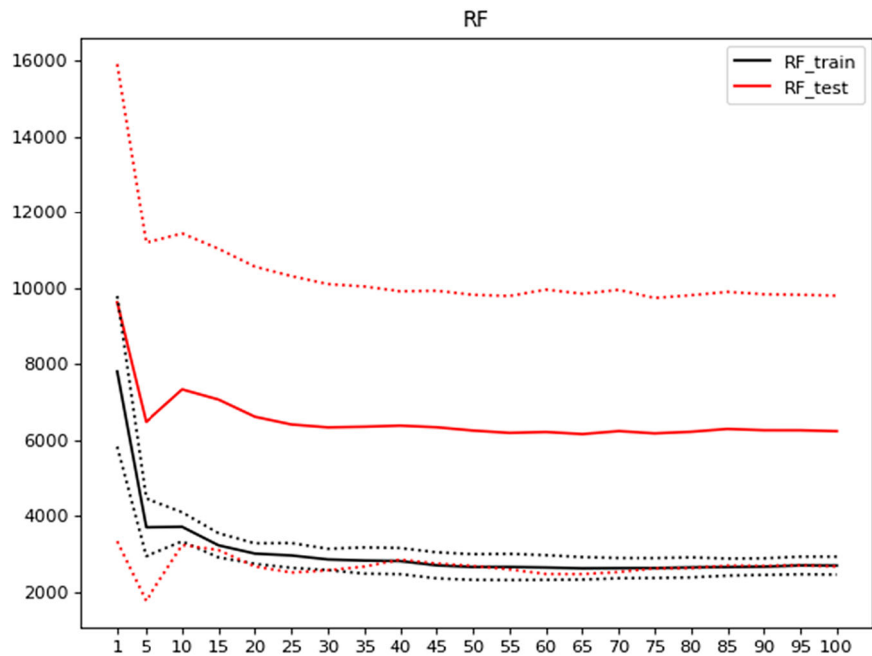


Fig. 1 Curve of the RMSE resulting from RF varying with NT. The RMSE of the training and test sets had the same change trend. Note: Solid and dotted lines represent the mean and variance of RMSE in the 10-fold cross-validation, respectively.

Table 3 Depth range of the trees.							
Count	Mean	Std	Min	25%	50%	75%	Max
300	14.247	1.508	11	13	14	15	19

greater the complexity of a single regression tree, the greater its contribution to the RF model; thus, the number of regression trees can be reduced. To achieve a balance between NT and DT, we also used HPO to determine the optimal DT value. In the RF algorithm, to reduce the correlation between regression trees, only parts of the features (*mtry*) were randomly selected as candidate splitting variables for each regression tree. We selected the number of maximum predictors available for each regression tree (optimal *mtry*) through HPO to achieve a balance between bias and variance. In addition, this study considered the splitting criterion of the regression tree in RF and selected the optimal splitting criterion among MSE, MAE, and Poisson deviation.

Because of the mutual influence of each hyperparameter, the approximate ranges of NT and DT were first estimated in this study. To estimate the approximate range of NT, we first performed a 10-fold cross-validation and then observed how the RMSE of the training set and test set decreased as NT increased. This method divides data into 10 equal portions and uses one portion as test data, and the other nine portions are used as training data to train each regressor independently. The 10-fold cross-validation is especially effective with limited datasets, as it maximizes the use of available data for training and testing while offering more reliable performance estimates by averaging results across all 10 folds, thereby minimizing sensitivity to random data splits. While this study employs ensemble learning methods such as Random Forest and Gradient Boosting Decision Trees for predicting carbon emissions, AutoML has been shown to automate model selection and optimization effectively, as demonstrated by Li et al. (2022) in predicting carpark price indices. However, manual ensemble learning offers greater flexibility and interpretability, especially when examining specific interactions between variables. The RF established in the cross-

validation uses default hyperparameters. As shown in Fig. 1, the RMSE of the training and test sets had the same change trend. The RMSE of the test set no longer decreased when NT was greater than 30 and tended to be stable. The RMSE of the training set no longer changed when NT was greater than 50. Moreover, by analyzing the influence of NT on the RMSE of the test set, it was found that increasing the number of regression trees did not lead to overfitting. Therefore, when conducting the HPO in this study, the search range of NT was set to [10,100].

To estimate the depth range of trees, we first estimated an RF model containing 300 trees with default hyperparameters. The depth of each tree was recorded. As shown in Table 3, in a randomly established RF model with 300 regression trees, the average value of DT was 14.247, the maximum value was 19, and the minimum value was 11. Therefore, to enable the HPO to search for hyperparameters in a controllable range, the depth range of the trees was set to [10,20].

GBDT also uses a regression tree (CART) as the base learner. However, in GBDT, increasing the number of regression trees may cause overfitting. Therefore, a reasonable NT value must be predetermined. Figure 2 shows the trend of the mean value of RMSE as NT increases in a 10-fold cross-validation based on the default hyperparameter (GBDT). It can be seen from Fig. 2 that the RMSE of the test set did not change drastically when NT approaches 45, and the downward trend stopped completely. The RMSE of the training set almost stopped decreasing when NT exceeded 185. Therefore, the HPO search range for NT was set to [25, 200]. In the GBDT algorithm, the contribution of a single weak learner to the GBDT model was controlled by the learning rate. Therefore, it is necessary to consider the balance between NT and the learning rate in the GBDT. In this study, the search range of the learning rate was set to [0.05, 2.05]. In addition, in GBDT models, we also considered the selection of an appropriate *mtry* to increase the randomness of the model and reduce overfitting. However, if the GBDT randomly selects some features to construct each weak learner, the overall learning ability of the model may be insufficient. Therefore, we also searched for the optimal DT to improve the learning ability of each weak learner in the GBDT, and set the range of DT to [2, 30]. The GBDT

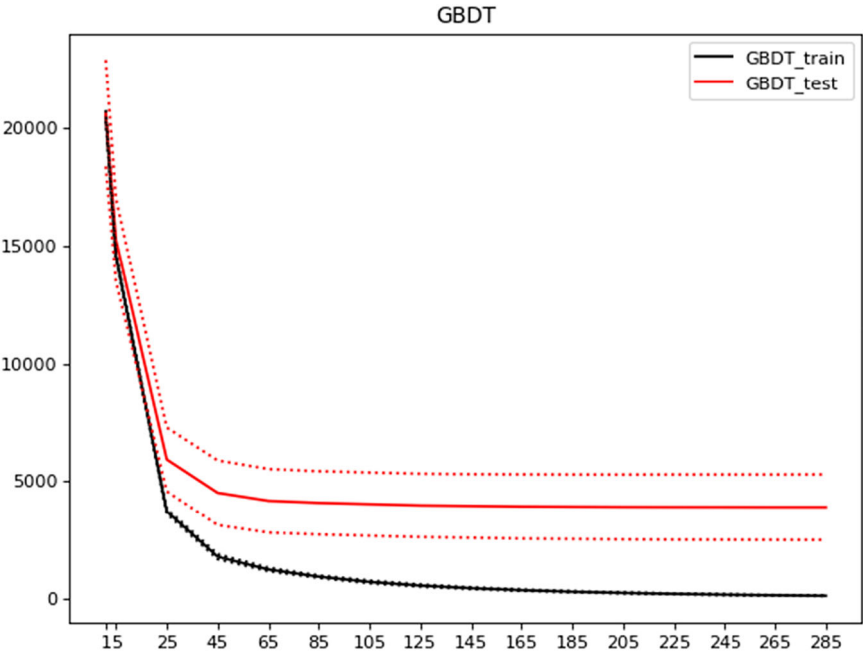


Fig. 2 Curve of the RMSE resulting from GBDT varying with NT. Note: Solid and dotted lines represent the mean and variance of RMSE in the 10-fold cross-validation, respectively. Figure 2 shows the trend of the mean value of RMSE as NT increases in a 10-fold cross-validation based on the default hyperparameter (GBDT).

Table 4 Optimal parameters of each model.							
Model	NT	DT	Mtry	Splitting criterion	Learning rate	Loss function	Initial value
RF	70	10	1	MSE	-	-	-
GBDT	70	9	3	Friedman MSE	0.1	MSE	Zero

builds a weak learner by fitting the current residuals; therefore, the selection of an appropriate loss function was necessary. This study selected the optimal loss function from the MSE, MAE, Huber loss, and quantile loss. We also considered the splitting criterion of the regression tree in the GBDT and selected the optimal splitting criterion from the Friedman MSE, MSE, and MAE. In addition, GBDT outputs the final estimation result by summing up the results of multiple regression trees; therefore, it was necessary to select an appropriate initial value. In this study, zero was considered as the initial value, and the RF model based on the optimal hyperparameter was also considered as an estimator to calculate the initial value of the GBDT.

The optimal hyperparameters were selected by looping through each set of possible values to individually fit the RF and GBDT, thereby determining the models that provided optimized estimations (R^2 and RMSE). For each step of the loop, a 10-fold cross-validation was applied to test the performance of the models. The predictions from the 10 regressors for the corresponding testing data were then combined into a vector of the same length as the data, and R^2 and RMSE were calculated. Python scikit-learn was used for model estimation and cross-validation (Pedregosa et al., 2011). A Hyperopt optimizer based on the tree-structured Parzen estimator (TPE) approach was used for the HPO (Komer et al., 2019; Ozaki et al., 2020).

The optimal hyperparameters of each model, determined by the TPE, and based on 10-fold cross-validation are listed in Table 4. The optimal hyperparameters can make the RF and GBDT exhibit optimal performance on the test set, based on the highest R^2 and

Table 5 Scores of the optimal models on the testing data in the 10-fold cross-validation.				
Model	Default hyperparameters		Optimal hyperparameters	
	R^2	RMSE	R^2	RMSE
RF	0.876	7511.859	0.924	5642.512
GBDT	0.909	6329.890	0.937	5224.746

smallest RMSE. In this study, the RF and GBDT with default hyperparameters were estimated. Models with default hyperparameters were used as benchmarks to compare the results of the HPO.

After obtaining the optimal hyperparameters, the scores (R^2 and RMSE) of the optimal model on the test set in the 10-fold cross-validation were also obtained. The scores on the test set were based on the average of the 10-fold cross-validation. This study compared the scores of the optimal models with those of the models using default hyperparameters. Thus, it was determined whether the prediction accuracy of the model can be improved by using optimal hyperparameters. As can be seen from the Table 5, after HPO, the fitting effect of the RF and GBDT was greatly improved, and the fitting effect of the GBDT was better than that of the RF. In addition, regardless of whether RF or GBDT was used with optimal hyperparameters, R^2 was much larger than the R^2 of the panel regression analysis (0.502) (Appendix B). This showed that the relationship between the features and target was not linear, and the ensemble learning model was a better fit for the data.

Although the prediction accuracy of the RF and GBDT on the test set was improved after HPO, the optimal hyperparameter combination may cause overfitting of the model. To judge the influence of the optimal hyperparameters on model overfitting, we carried out 20-fold cross-validation to compare the RMSE, based on the training and test sets, when the model used default hyperparameters and optimal hyperparameters.

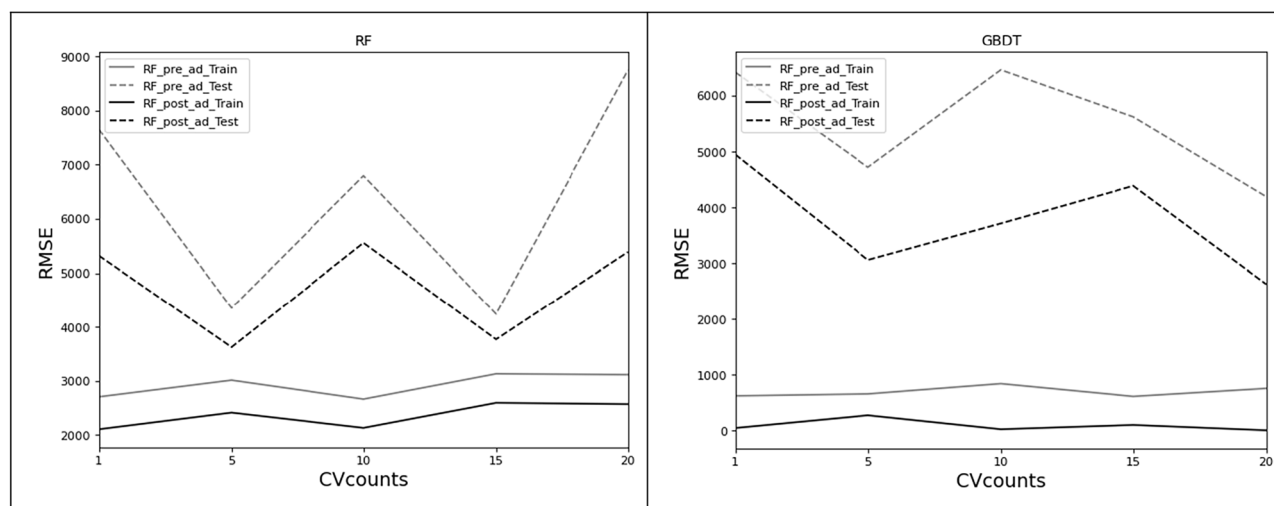


Fig. 3 RMSE when the model uses default hyperparameters and optimal hyperparameters. Figure 3 shows that when the RF used optimal hyperparameters, the RMSE on both the training and test sets decreased, and the distance between the two learning curves also decreased.

When the RF used optimal hyperparameters, the RMSE on both the training and test sets decreased, and the distance between the two learning curves also decreased (Fig. 3). In GBDT, after using the optimal hyperparameter, the distance between the learning curves of the training set and the test set was reduced, indicating that the gap between the fitting results on the training set and the predicted results on the test set was reduced after using the optimal hyperparameters. The results demonstrate that the optimal hyperparameters selected by TPE improve the model's predictive accuracy without leading to overfitting. Therefore, RF and GBDT models with optimal hyperparameters were used for the following analysis.

Variable importance and partial dependence plots. In a decision tree, because only one feature is used at each node split, it is possible to measure the contribution of that feature; that is, how much it reduces the residual sum of squares (RSS) or the Gini index. RF and GBDT use a decision tree as a base learner; therefore, it is possible to calculate the importance of each feature, as in a decision tree. Specifically, for feature A, the decrease in RSS in each regression tree in the RF and GBDT caused by feature A can be calculated. Subsequently, the average RSS decreases for all regression trees to obtain the variable importance of feature A (Louppe et al., 2013).

As panel data were used in this study, the differences in carbon emissions from manufacturing in different regions could be compared and analyzed. In this study, RF and GBDT were first generated by applying 187 observations drawn from 17 administrative regions between 2013–2023 to determine the variation in carbon emissions change for each region. RF and GBDT used only 17 regions as predictors to model the corresponding carbon emissions (Wang et al., 2021). In practice, each region was converted into a one-hot variable. The determination of relevant optimal hyperparameters in the RF and GBDT was conducted following the steps described in the section “Determining the hyperparameters of the model”. As discussed above, the greater the mean decrease in RSS, the stronger is the association with the response; also, the greater the average reduction in the RSS caused by a feature, the stronger the correlation between the feature and response. Figure 4 shows the importance of the variables for various regions. The ranking was roughly the same across regions in the RF and GBDT. Overall, variable importance for regions

was generally high in Jeonnam, Chungnam, Gyeongbuk, Ulsan, and Gyeonggi.

These results were consistent with the spatial distribution of the South Korean manufacturing industry. It is worth noting that Seoul and Busan, as the first and second largest cities in South Korea, respectively, had a low level of variable importance. A reasonable explanation is that Seoul and Busan mainly rely on the tertiary industry, with manufacturing accounting for a small proportion only, so carbon emissions from manufacturing are relatively low. Thus, variations in economic levels, industrial structures, and other aspects among regions led to significant spatial differences in the ranking of variable importance.

The RF and GBDT algorithms were then used to investigate the temporal predictability of carbon emission variations in the 17 regions and to study the connection between various influencing factors and carbon emissions. The optimal hyperparameters of the RF and GBDT were adopted from the results obtained in the section “Determining the hyperparameters of the model”. Subsequently, the importance of the eight features in the RF and GBDT was calculated, and the importance of each feature was sorted and plotted to obtain the variable importance plot. Figure 5 shows that *si* and *di* were the most important factors influencing carbon emissions in South Korea's manufacturing sector. In RF, the variable importance of *si* and *di* was 0.158 and 0.156, respectively, which was not significantly different from the GBDT results (0.142 and 0.152). Whether in the RF or GBDT, the variable importance of *si* and *di* was ranked first or second, proving that the manufacturing agglomeration indices have the best predictive power and that their values have a significant impact on the outcome values. Moreover, manufacturing agglomeration is even more important than some traditional predictors, such as *fdi* and *population*.

Variable importance simply measures and ranks the importance of the features. To investigate the marginal effects of *si* and *di* on carbon emissions, a partial dependency plot was used to visualize the marginal impact (Kumar, 2017). A partial dependence plot depicts the functional relationship between a small number of features and a target. They show how the target partially depends on the values of specific features by marginalizing the values of all other features. Owing to the limitations of human perception, the size of the target feature is usually one or two. The partial dependency function of feature x_1 on the predictor can be expressed as $f(x_1) = E_{x_2, \dots, x_p} f(x_1, x_2, \dots, x_p)$,

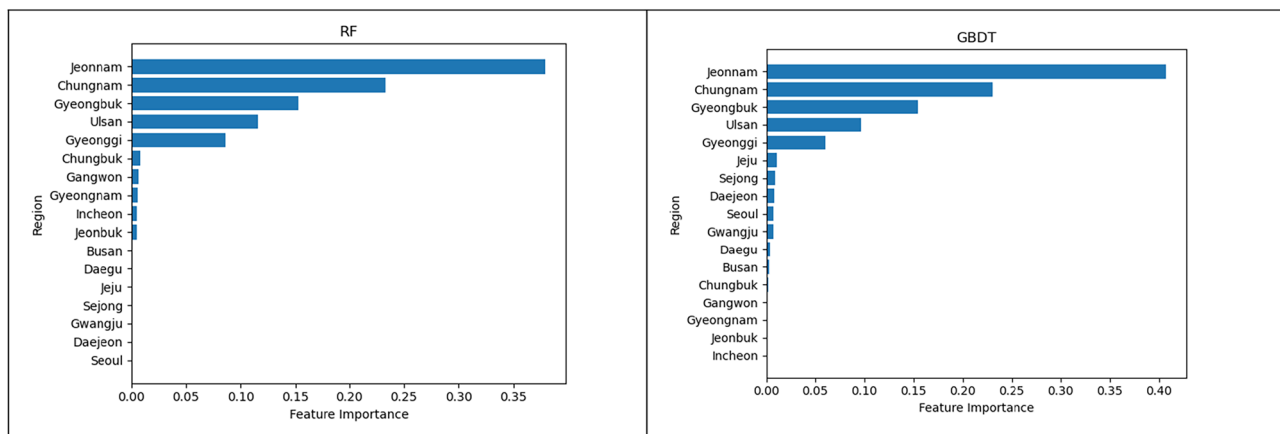


Fig. 4 Mean decrease of RSS of the features (regions) in RF and GBDT. Note: Raw data is normalized to the interval (0,1) to adjust the units in which the data is presented. Figure 4 shows the importance of the variables for various regions. The ranking was roughly the same across regions in the RF and GBDT. Overall, variable importance for regions was generally high in Jeonnam, Chungnam, Gyeongbuk, Ulsan, and Gyeonggi.

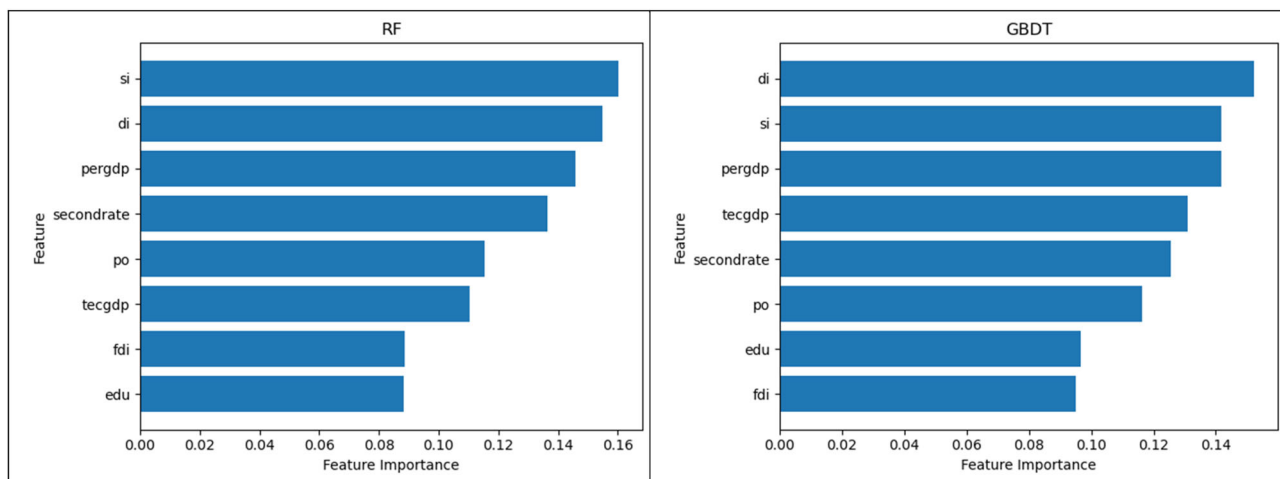


Fig. 5 Feature importance of RF and GBDT. Note: Raw data is normalized to the interval (0, 1) to adjust the units in which the data is presented. Figure 5 shows that *si* and *di* were the most important factors influencing carbon emissions in South Korea's manufacturing sector.

where E is the expectation of the variables $(x_2, \dots, x_p)$. Therefore, if we consider the expectation, $\hat{f}(x_1)$ is a function of x_1 . In actual calculations, the sample mean can be used instead of the population mean, that is, $\hat{f}(x_1) = \frac{1}{n} \sum_{i=1}^n f(x_1, x_{i2}, \dots, x_{ip})$. Given a feature x_i , $\hat{f}(x_i)$ can be calculated and the corresponding partial dependency plot can be drawn. Figure 5 shows how the predicted carbon emissions changed (on the vertical axis) as they moved from small values of *si* and *di* to large values (on the horizontal axis).

It can be seen from Fig. 6 that *si* had a positive impact on carbon emissions, which indicates that with the improvement of the specialized agglomeration level in the manufacturing sector, the carbon emissions of the manufacturing sector in relevant regions also increased. In the RF and GBDT models, the partial dependence curves of *si* differed in detail, but the overall trends were similar. This verified the reliability of the results. In addition, the partial dependency curve of *si* presented an evident ladder shape. Both curves showed a maximum increase when *si* exceed 1.6, and then tended to be stable. This indicates that when *si* exceed 1.6, carbon emissions may increase significantly. However, by observing the partial dependence curves of *di* in RF and GBDT, both curves showed the same downward trend. This means that, unlike *si*, *di* had a negative impact on carbon

emissions from the manufacturing sector; that is, carbon emissions decreased with the increase in diversified agglomeration level. Owing to the high value of the variable importance of *di*, it was revealed that the formation and development of diversified agglomeration in the manufacturing sector is very conducive to reducing carbon emissions.

In addition, the values of *si* were relatively averaged on the horizontal axis, but the values of *di* were mostly concentrated between 0 and 0.2. In particular, the partial dependence curves of *di* decreased significantly in the interval (0, 0.2). This result indicates that the carbon emission reduction effect may be realized more quickly in the early stages of diversified agglomeration.

The results of the correlation analysis in the section "Correlation analysis" indicate that *si* and *di* have an opposite relationship, which means that areas with a higher agglomeration level of specialization tend to have a lower agglomeration level of diversification. Therefore, the regions with low *di* values ranging from 0 to 0.2 may include the regions with a high level of specialization agglomeration. The fact that carbon emissions drop sharply in the *di* range of 0 to 0.2 may indicate that diversified agglomeration can bring carbon reduction effects to specialized production areas.

To explore the potential interaction between *si* and *di* and their influence on carbon emissions, we created two-way partial

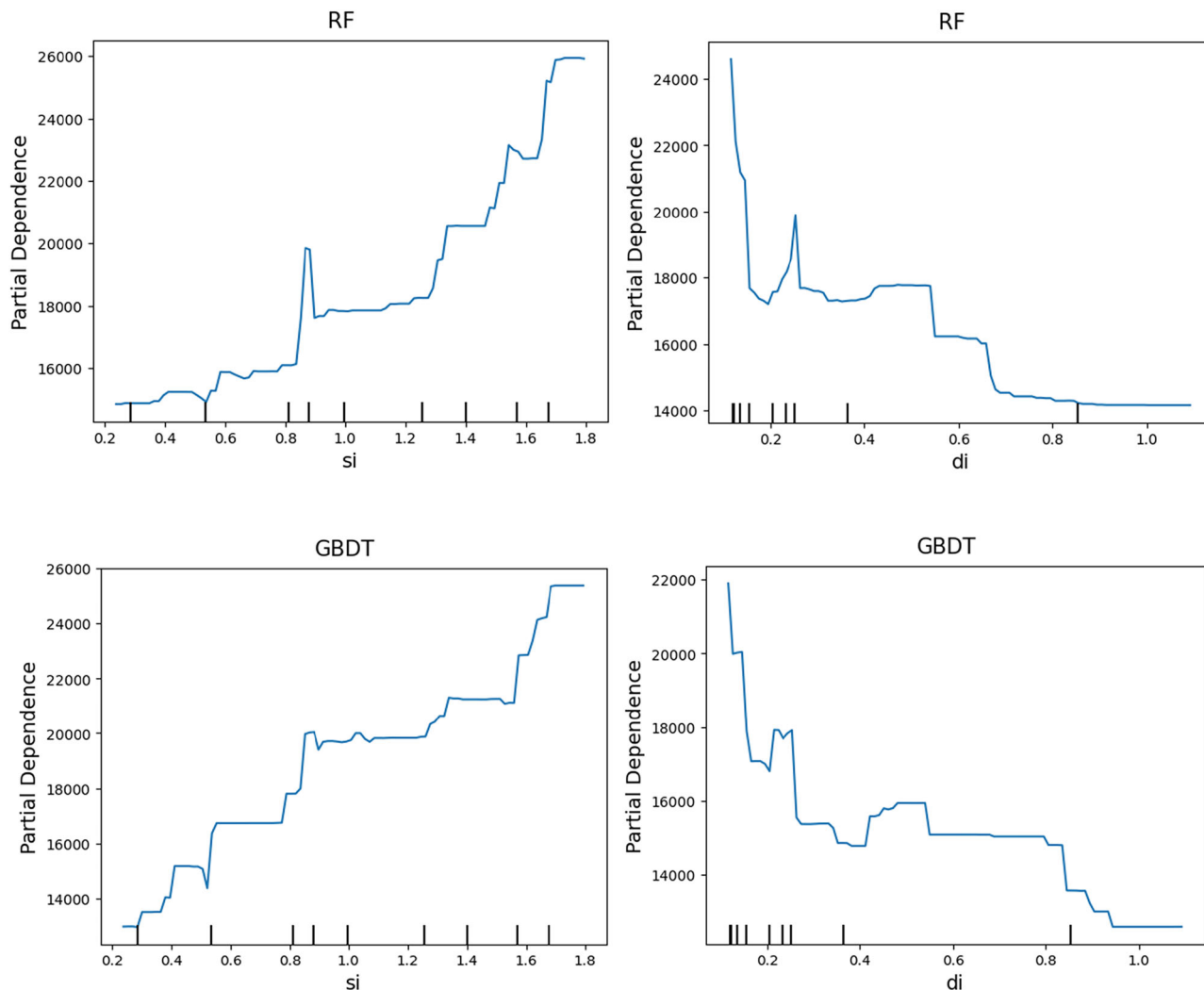


Fig. 6 Partial dependency plot of si and di in RF and GBDT. Figure 6 shows that si had a positive impact on carbon emissions, which indicates that with the improvement of the specialized agglomeration level in the manufacturing sector, the carbon emissions of the manufacturing sector in relevant regions also increased.

dependence plots illustrating their relationship with carbon emissions. Furthermore, to assess the effects of temporal variations, the data were divided into two periods, 2013–2018 and 2018–2023, with separate two-way partial dependence plots generated for each period. Figure 7 shows the two-way partial dependence plots of si and di on the carbon emissions. Partial dependence plots with both si and di enabled us to visualize the interactions among them. The two-way partial dependence plot also shows the dependence of the carbon emissions on the joint values of si and di . First, in the RF and GBDT models, large values of the target are in the regions where si is large and di is small; on the other hand, small values of the target are in the opposite direction, that is, the region where si is small and di is large. This reveals the internal relationship between si and di when they have a joint effect on carbon emissions. Second, the target was more sensitive to the change in si . Regardless of di , the target always reached the highest point when si was at its maximum value. This suggests that si had a greater effect on carbon emissions than di . Third, introducing a specialized agglomeration development pattern in regions already characterized by diversified agglomeration leads to a significant increase in carbon emissions, especially during the early stages of diversification. Conversely, introducing a diversified agglomeration model in regions dominated by

specialization reduces carbon emissions; however, as specialization intensifies, the emission reduction effects of diversification diminish considerably. Last, A comparison of the partial dependence plots across different time periods reveals a clear weakening trend in the impact of the interaction between si and di on carbon emissions in both the RF and GBDT models. This is evidenced by the increasing flatness of the partial dependence function surfaces over time.

Appendix C also presents plots of these partial dependence functions for po , $pergdp$, edu , $secondrate$, $techgdp$, and fdi . First, the impact of population size on carbon emissions seems to have an inverted U-shape; that is, in the early stage of population growth, carbon emissions increased with population growth, but as population growth reached a threshold, it no longer increased carbon emissions. This was verified using both the RF and GBDT models. The growth of per capita GDP had an obvious promoting effect on carbon emissions, especially when the per capita GDP reached about 35 million won, carbon emissions increased the fastest, and the growth trend then tended to decrease more gradually. Human capital had an obvious restraining effect on manufacturing carbon emissions because the partial dependency curve showed an obvious downward trend in both the RF and GBDT models. This may be because the improvement in human capital will promote the

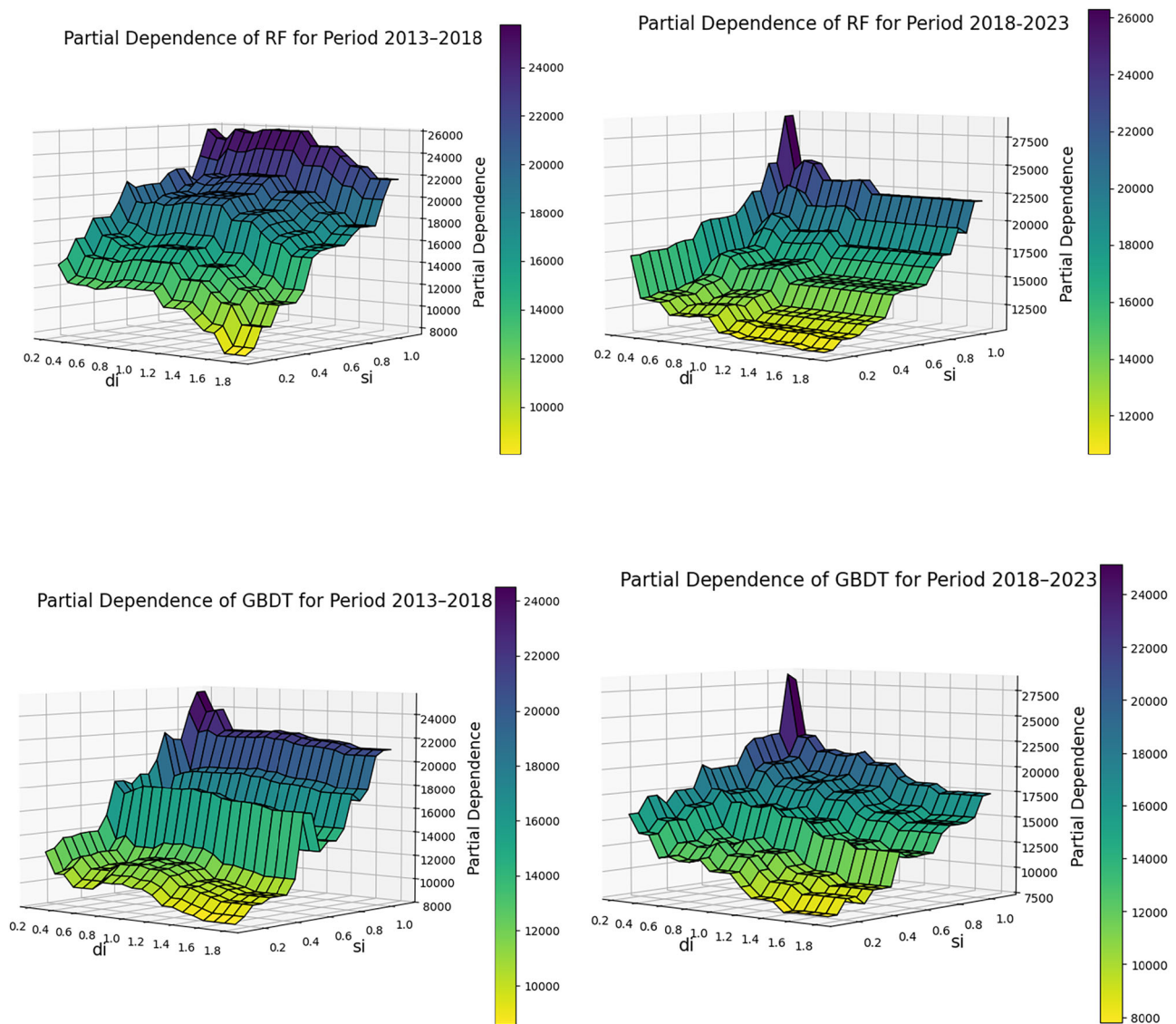


Fig. 7 Partial dependence plot of CARBON emissions on si and di in RF and GBDT. Figure 7 shows the two-way partial dependence plots of si and di on the carbon emissions.

innovation and upgrading of technology within the manufacturing industry and promote carbon emission reduction. The influence of industrial structure on carbon emissions was also obvious. The results showed that an increase in the proportion of secondary industries significantly increased carbon emissions in the manufacturing sector. Therefore, it was concluded that the secondary industry is still the main source of carbon emissions in South Korea's manufacturing sector. Scientific and technological innovation plays a clear role in carbon emission reduction, especially in the initial stage of investment, resulting in a significant reduction in carbon emissions. The impact of foreign investment on manufacturing carbon emissions showed an inverted U-shaped structure, which indicates that foreign investment will greatly increase carbon emissions in the initial stage; with an increase in investment scale, however, carbon emissions will gradually decrease.

Conclusion

This study investigates how specialized and diversified agglomeration patterns influence carbon emissions across regions in South Korea, using ensemble learning models (RF and GBDT). The results demonstrate that manufacturing agglomeration

significantly affects carbon emissions, with specialized agglomeration leading to an increase and diversified agglomeration contributing to emission reduction, particularly in the early stages. Additionally, the trade-off between specialization and diversification suggests that industrial structure plays a crucial role in shaping regional environmental outcomes. Over time, the interaction effects between the two agglomeration patterns weaken, highlighting the need for dynamic policy adjustments.

Discussion

This study makes several significant academic contributions to the fields of environmental economics, regional economics, and sustainable development. First, it provides a detailed understanding of how specialized and diversified industrial agglomeration patterns influence carbon emissions. While previous research has primarily focused on the isolated effects of either specialization or diversification, this study emphasizes the dynamic trade-offs and interactions between the two patterns and their impact on carbon emissions. Second, prior studies have mainly concentrated on predicting carbon emissions using traditional factors such as GDP, energy consumption, and urbanization. By incorporating specialized and diversified

agglomeration indices as key predictive indicators, this study fills a critical gap, offering a novel perspective on the role of spatial and structural economic factors in shaping environmental outcomes. Finally, the study applies ensemble learning models (RF and GBDT) to analyze the impact of industrial agglomeration patterns on carbon emissions. These methods capture the complex relationships between agglomeration patterns and carbon emissions. Moreover, the use of partial dependence plots allows for the identification of dynamic interactions between agglomeration patterns and their effects on carbon emissions. By integrating machine learning into the study of the environmental impacts of industrial agglomeration, this research introduces methodological innovation, providing more precise and robust analyses.

The findings reveal the intricate interactions between specialized and diversified industrial agglomeration patterns and their impact on carbon emissions. These results highlight the need for targeted and adaptive policies to balance economic growth with environmental sustainability. First, carbon emissions in highly specialized regions are more concentrated and challenging to control, necessitating the mandatory adoption of green technologies and stricter emission standards. Governments should establish industry-specific emission caps in these regions, mandate the use of energy-efficient technologies, implement compliance monitoring mechanisms, penalize non-compliant enterprises, and restrict their access to government subsidies. Second, the study shows that specialized agglomeration significantly increases carbon emissions, while diversified agglomeration reduces emissions, particularly in its early stages. To address this, governments should provide tax incentives and subsidies to attract green industries from diverse sectors to highly specialized regions. Policies should also encourage existing industries to adopt green technologies or transition toward low-carbon production methods. Third, diversified agglomeration demonstrates greater potential for adopting green technologies and reducing carbon emissions. Collaboration between specialized and diversified clusters can enable technology sharing and resource complementarity. To facilitate this, governments should establish collaborative R&D funds to support joint projects between the two types of clusters and create regional innovation platforms to promote technology exchange focused on emission reduction. Finally, as the interaction effects between specialization and diversification on carbon emissions weaken over time, dynamic policies are required to address this shift. Governments should design phased support policies tailored to specialization or diversification agglomeration, and regularly assess these policies to align with the evolving dynamics of regional economies and environmental conditions.

Based on the research result, several limitations might be inherent in the study that need to be acknowledged. First, its focus on South Korea may limit the generalizability of the findings to regions with different industrial structures, environmental policies, and economic conditions. Cross-country analyses, particularly comparing developed and developing nations, could provide broader insights into the global applicability of these findings. Additionally, the availability of panel data is somewhat limited, which may affect the comprehensiveness and robustness of the conclusions. Expanding the dataset to cover a longer time span and more regions would enhance the validity of the results. Future research should consider incorporating more extensive data sources to further validate and extend these findings. Second, this research does not explicitly account for spatial spillover effects, where carbon emissions in one region may influence or be influenced by neighboring regions through economic and environmental interdependencies. Future research could address this limitation by applying spatial econometric models, such as the Spatial Durbin Model or Spatial Lag Model, to quantify the spatial spillover effects of industrial agglomeration patterns on carbon emissions. This would help illuminate the regional

interconnections and the broader implications of industrial agglomeration dynamics. Third, while this study focuses on the relationship between agglomeration patterns and carbon emissions, it does not explore other critical environmental outcomes, such as resource depletion, pollution, or biodiversity loss. Future studies should consider expanding the scope of analysis to include additional environmental metrics like air and water quality, resource utilization, and ecosystem health. This broader approach would provide a more comprehensive understanding of the environmental impacts of industrial agglomeration. Fourth, this study observes a temporal weakening of interaction effects between specialization and diversification on carbon emissions but does not deeply analyze the drivers of this change. Future research should investigate potential factors contributing to this weakening, such as shifts in energy policies, technological advancements, or changes in global trade dynamics. Last, this study employs the ensemble learning models RF and GBDT to capture nonlinear relationships, but they provide limited interpretability for understanding causal mechanisms. Future research could address this limitation by combining machine-learning approaches with causal inference methods, such as econometric models. This hybrid approach would enhance understanding of the causal pathways linking agglomeration patterns to carbon emissions.

Data availability

No datasets were generated or analyzed during the current study.

Received: 14 October 2024; Accepted: 29 May 2025;

Published online: 23 June 2025

References

- Alex O, Acheampong Emmanuel B, Boateng (2019) Modelling carbon emission intensity: application of artificial neural network. *J Clean Prod* 225:833–856
- Ang JB (2009) CO₂ emissions, research and technology transfer in China. *Ecol Econ* 68:2658–2665
- Boyd G, McDonald JF, Ross M, Hanson DA (1987) Separating the changing composition of U.S. manufacturing production from energy efficiency improvements: a Divisia index approach. *Energy J* 8:77–96
- Breiman L (1996a) Bagging predictors. *Mach Learn* 24:123–140
- Breiman L (2001) Random Forests. *Mach Learn* 45:5–32
- Buhlmann P, Yu B (2002) Analyzing bagging. *Ann Stat* 30:927–961
- Bwalya SM (2006) Foreign direct investment and technology spillovers: evidence from panel data analysis of manufacturing firms in Zambia. *J Dev Econ* 81(Issue 2):514–526
- Capello R (2007) Regional economics. Routledge, Milton Park, Abingdon, Oxon
- Capello R (2015) Regional economics. Routledge
- Cheng YQ, Wang ZY, Ye XY, Wei YD (2015) Spatiotemporal dynamics of carbon intensity from energy consumption in China. *J Geogr Sci* 24(4):631–650
- Chuai XW, Huang XJ, Wang WJ, Wen JQ et al. (2012) Spatial econometric analysis of carbon emissions from energy consumption in China. *J Geogr Sci* 22(4):630–642
- Combes PP (2000) Economic structure and local growth: France, 1984–1993. *J Urban Econ* 47(3):329–355
- Dietz T, Rosa EA (1994) Rethinking the environmental impacts of population, affluence, and technology. *Hum Ecol Rev* 1(2):277–300
- Du H, Chen Z, Mao G, Li RYM, Chai L (2018) A spatio-temporal analysis of low carbon development in China's 30 provinces: A perspective on the maximum flux principle. *Ecol Indic* 90:54–64
- Duc TA (2007) Experimental investigation and modeling approach of the impact of urban wastewater on a tropical river: a case study of the Nhue River, Hanoi, Vietnam. *J Hydrol* 33(4):347–358
- Ehrenfeld J (2003) Putting a spotlight on metaphors and analogies in industrial ecology. *J Ind Ecol* 7(1):1–4
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232
- Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random Forests for land cover classification. *Pattern Recognit Lett* 27(4):294–300
- Han F, Feng P, Yang LG (2014) Spatial agglomeration effects of China's cities and industrial energy efficiency. *China Popul, Resour Environ* 24(5):72e79. (In Chinese)

- Han F, Xie R, Fang J, Liu Y (2018) The effects of urban agglomeration economies on carbon emissions: Evidence from Chinese cities. *J Clean Prod* 172:1096–1110
- Henderson V, Kuncoro A, Turner M (1995) Industrial development in cities. *J Political Econ* 103(5):1067–1090
- Hoover EM (1937) Location theory and the shoe and leather industries. Harvard University Press, Cambridge, MA
- Jena PR, Managi S, Majhi B (2021) Forecasting the CO₂ emissions at the global level: a multilayer artificial neural network modelling. *Energies* 14(no. 19):6336
- Jirčiková E, Pavelková D, Bialic-Davendra M et al. (2013) The age of clusters and its influence on their activity preferences. *Technol Econ Dev Econ* 19(4):621–637
- Jung TY, Park TS (2000) Structural change of the manufacturing sector in Korea: measurement of real energy intensity and CO₂ emissions. *Mitig Adapt Strateg Glob Change* 5(3):221–238
- Kim SW, Lee K, Nam K (2010) The relationship between CO₂ emissions and economic growth: The case of Korea with nonlinear evidence. *Energy Policy* 38(10):5938–5946
- Komer B, Bergstra J, Eliasmith C (2019) Hyperopt-sklearn. Automated Machine Learning: Methods, Systems, Challenges. Springer. pp. 97–111
- Korea Energy Economics Institute (1997). Yearbook of Energy Statistics
- Krugman P (1998) Space: the final frontier. *J Econ Perspect* 2(2):161–174
- Kumar A, Saini P (2017) Effects of partial dependency of features and feature selection procedure over the plant leaf image classification. In: International Conference on Recent Developments in Science, Engineering and Technology. Springer, Singapore. pp. 208–225
- Lan F, Sun L, Pu W (2021) Research on the influence of manufacturing agglomeration modes on regional carbon emission and spatial effect in China. *Econ Model* 96:346–352
- Lazăr AI (2018) Economic efficiency vs. Positive and negative externalities. *Rev Gen Manag* 27(1):112–118
- Lee H (2021) Is carbon neutrality feasible for Korean manufacturing firms? The CO₂ emissions performance of the Metafrontier Malmquist–Luenberger index. *J Environ Manag* 297:113235
- Lei H, Wang H, Zhu MX (2017) Industrial agglomeration, energy consumption and environmental pollution. *J Ind Technol Econ* 36(9):58–64
- Li, R.Y.M.; Song, L.; Li, B.; James, C.; Crabbe, M.; Yue, X.G. (2023) Predicting Carpark Prices Indices in Hong Kong Using AutoML. *CMES Comput. Model. Eng. Sci* 134:2247–2282
- Li RYM, Wang Q, Zeng L, Chen H (2023) A study on public perceptions of carbon neutrality in China: has the idea of ESG been encompassed? *Front Environ Sci* 10:949959
- Li YL (2014) An empirical analysis based on marketization, industrial agglomeration and environmental pollution. *Stat Res* 31(8):39–45. (In Chinese)
- Lin J, Wan H, Cui Y (2020) Analyzing the spatial factors related to the distributions of building heights in urban areas: a comparative case study in Guangzhou and Shenzhen. *Sustain Cities Soc* 52:101854
- Lin J, Lu S, He X, Wang F (2021) Analyzing the impact of three-dimensional building structure on CO₂ emissions based on random forest regression. *Energy* 236(C):121502
- Loupe G, Wehenkel L, Sutura A, Geurts P (2013) Understanding variable importances in forests of randomized trees. *Adv Neural Inf Process Syst*. 26: 431–439
- Marshall A (2009) Principles of economics: unabridged eighth edition. Cosimo, Inc
- Moore DE, Lees BG, Davey SM (1991) A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *J Environ Manag* 15:59–71
- Ozaki Y, Tanigaki Y, Watanabe S, Onishi M (2020) Multiobjective tree-structured Parzen estimator for computationally expensive optimization problems. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference. Association for Computing Machinery, Cancún, Mexico, 533–541
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- Poumanyong P, Kaneko S (2010) Does urbanization lead to less energy use and lower CO₂ emissions? A cross-country analysis. *Ecol Econ* 70:434–444
- Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9:181–199
- Qin B, Wu JF (2015) Does urban concentration mitigate CO₂ emissions? Evidence from China 1998–2008. *China Econ Rev* 35:220–231
- Verhoef ET, Nijkamp P (2002) Externalities in urban sustainability: environmental versus localization-type agglomeration externalities in a general spatial equilibrium model of a single-sector mono-centric industrial city. *Ecol Econ* 40(2):157–179
- Wang J, Dong X, Dong K (2022) How does ICT agglomeration affect carbon emissions? The case of Yangtze River Delta urban agglomeration in China. *Energy Econ* 111:106107
- Wang Z, Zhao Z, Wang C (2021) Random forest analysis of factors affecting urban carbon emissions in cities within the Yangtze River Economic Belt. *PLoS One* 16(6):e0252337
- Wei S, Yuwei W, Chongchong Z (2018) Forecasting CO₂ emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine. *Environ Sci Pollut Res* 25(29):28985–28997
- Wen L, Yuan X (2020) Forecasting CO₂ emissions in China's commercial department, through BP neural network based on random forest and PSO. *Sci Total Environ* 718:137194
- Wu Z, Woo SH, Oh JH, Lai PL (2024) Temporal and spatial effects of manufacturing agglomeration on CO₂ emissions: evidence from South Korea. *Humanit Soc Sci Commun* 11(1):1–14
- Xia J, Li RYM, Zhan X, Song L, Bai W (2022) A study on the impact of fiscal decentralization on carbon emissions with U-shape and regulatory effect. *Front Environ Sci* 10:964327
- Xu G, Schwarz P, Yang H (2019) Determining China's CO₂ emissions peak with a dynamic nonlinear artificial neural network approach and scenario analysis. *Energy Policy* 128:752–762
- Yuan H, Feng Y, Lee J, Liu H, Li R (2020) The spatial threshold effect and its regional boundary of financial agglomeration on green development: a case study in China. *J Clean Prod* 244:118670
- Zhou XH, Fan QQ (2016) Mechanism of carbon intensity reduction and optimization design of its industrial allocation. *J World Econ Res* 7:168e192. (In Chinese)

Acknowledgement

This research work is partially funded by Chiang Mai University.

Author contributions

Zhen Wu contributed to data collection, data analysis, analytical graphing, and manuscript writing. Su-Han Woo contributed to the conceptualization and manuscript revision. Po-Lin Lai contributed to data collection, data analysis, and manuscript revision. Pairach Piboonrungraj contributed to manuscript revision and funding acquisition

Competing interests

The authors declare no competing interests.

Ethical approval

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

Informed consent

This article does not contain studies involving human participants engaged by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-025-05150-x>.

Correspondence and requests for materials should be addressed to Pairach Piboonrungraj or Po-Lin Lai.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons

Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025