Opportunities and Challenges in Sustainability
https://www.acadlore.com/journals/OCS

Data Factor Agglomeration and the Green Development of Traditional Enterprises

Lelai Shi [1,2], Qiuhang Chen [1,2*]

[1] School of Economics, Wuhan Textile University, 430200 Wuhan, China
[2] Hubei Modern Textile Industry Economic Research Center, Wuhan Textile University, 430200 Wuhan, China

[*] Correspondence: Qiuhang Chen (2415143005@wtu.edu.cn)

Received: 05-22-2025 **Revised:** 06-16-2025 **Accepted:** 06-25-2025

Citation: Shi, L. L & Chen, Q. H. (2025). Data factor agglomeration and the green development of traditional enterprises. *Oppor Chall. Sustain.*, *4*(4), 240–257. https://doi.org/10.56578/ocs040401.

Abstract: This study investigates the pathways through which data factor agglomeration (DFA) facilitates the green development of traditional firms in the digital economy. First, we construct a micro-theoretical framework to systematically analyze the mechanisms by which data factor agglomeration influences firms' green and sustainable development. Second, exploiting the establishment of China's National Big Data Comprehensive Pilot Zones as a quasi-natural experiment, we employ a difference-in-differences (DID) approach using a panel of A-share listed traditional manufacturing firms from 2011 to 2022. The empirical results indicate that data factor agglomeration significantly promotes green development in traditional firms by accelerating IT and improving capacity utilization (CU) and energy efficiency. These findings remain robust after a battery of robustness checks, including double machine learning (DML) and instrumental-variable approaches. Heterogeneity analyses reveal that the positive effects of data factor agglomeration are more pronounced for state-owned enterprises, firms led by technologically skilled executives, heavily polluting industries, and firms located in regions with stronger government support and stricter environmental regulation. Further analysis uncovers substantial spatial heterogeneity: while the direct effect of data factor agglomeration on local firms' green development is significantly positive, it generates a "siphoning effect" on geographically adjacent regions, whereas no significant spillover effects are observed among economically similar regions. Overall, this study elucidates the mechanisms and key determinants of green development for traditional firms in the digital era, providing important theoretical and practical implications for global economic growth and sustainable development.

Keywords: Data factor agglomeration; Green development; Capacity utilization; Energy efficiency; Artificial intelligence

1. Introduction

Artificial intelligence (AI) has become a key driving force behind technological progress in the current era. The awarding of the 2024 Nobel Prize in Physics to Hopfield and Hinton for their foundational contributions to artificial neural networks further confirms the central role of AI in scientific and economic development. However, the effectiveness of AI and machine learning applications is constrained not only by algorithms themselves but, more importantly, by the quality and accessibility of data as a critical production factor. As a new type of production factor following land, labor, capital, and technology, data possesses unique characteristics such as virtuality, non-rivalry, replicability, non-exclusivity, low marginal cost, economies of scale, and positive externalities. These attributes endow data with enormous potential value within modern economic systems, making it a new engine driving economic growth and industrial transformation (Jones & Tonetti, 2020).

Against the backdrop of slowing growth in traditional factor inputs and a global economy facing insufficient growth momentum, data factors offer a novel pathway to overcoming development bottlenecks. Taking traditional manufacturing as an example, its relatively mature technological system means that extensive factor inputs often lead to increased resource consumption and environmental pollution, which are incompatible with sustainable development requirements. In this context, the importance of data factors is increasingly evident: on one hand, the accumulation and utilization of data within enterprises can effectively reduce information asymmetry and lower

operational uncertainty, thereby enhancing productivity (Farboodi & Veldkamp, 2021); on the other hand, the deep integration of data with traditional production factors facilitates knowledge creation and technological progress, further promoting long-term economic growth and structural transformation (Agrawal et al., 2019). Overall, data is progressively permeating production, circulation, and consumption processes, profoundly reshaping traditional production methods and economic operation logic, and becoming a critical carrier for the transformation of old and new driving forces.

To promote the agglomeration and efficient allocation of data factors, the Chinese government has established a policy-driven ecosystem conducive to data development. A prime example of this ecosystem in action is the establishment of National Big Data Comprehensive Pilot Zones as early as 2016, which pioneered the exploration of data factor agglomeration (DFA) development pathways. In recent years, policy support has been continuously strengthened, with the issuance of key documents such as the "Opinions on Building a Data Infrastructure System to Better Utilize Data Factors" and the "Data Factor × Three-Year Action Plan (2024–2026)," which explicitly call for "fully leveraging the amplification, superposition, and multiplication effects of data factors to build a data-centric digital economy," emphasizing the multiplier effect of data in empowering economic and social development.

Within this context, data, as a critical factor in the emerging digital industry, has experienced exponential growth and formed agglomerations in specific regions, where homogeneous or heterogeneous big data resources are concentrated and aggregated within certain geographic scopes. Whether DFA can not only promote the development of the big data industry but also empower traditional industries remains a crucial question. How can digital factors be effectively transformed into drivers of green and sustainable development for traditional industries? What are the underlying mechanisms? Addressing these questions is essential for understanding how traditional industries can achieve "new" development in the digital era and holds significant theoretical and practical implications for global economic growth and green development.

The literature relevant to this study mainly covers three aspects: the characteristics of data factors, the socio-economic effects of data factor agglomeration, and its ecological and environmental impacts.

1.1 Characteristics of Data Factors

From the perspective of information economics, data is regarded as a by-product of economic activities (Bergemann & Bonatti, 2019), characterized by non-rivalry, replicability, and increasing returns to scale (Jones & Tonetti, 2020). Due to its non-rival nature, data is similar to a public good, and the extensive use of big data can generate substantial benefits. Moreover, its replicability feature implies low usage costs (Goldfarb & Tucker, 2019) and no depreciation or value loss through use, facilitating open sharing, replication, and recombination of data (Lynch, 2008). Varian (2019) provided an economic discussion on machine learning and data factors, emphasizing data's non-rival nature and describing data as "the new oil," which further confirms its increasing returns to scale characteristic. Meanwhile, Arrieta-Ibarra et al. (2018) highlight the notion of "data as labor." Overall, data is increasingly recognized as a new quality production factor succeeding traditional factors such as land, labor, and capital.

1.2 Socioeconomic Effects of DFA

New economic geography theory identifies agglomeration externalities and economies of scale as key drivers behind the "core-periphery" spatial structure (Krugman, 1991). On one hand, DFA attracts data industries, artificial intelligence, and other high-tech sectors, fostering a specialized division of labor and generating Marshallian externalities, which reduce the operational costs of economic activities (Goldfarb & Tucker, 2019) and enable lower-cost access to and utilization of external resources (Wennberg & Lindqvist, 2010). On the other hand, the clustering of data industries further attracts skilled technical talent, optimizes resource allocation, promotes inter-firm cooperation and knowledge sharing, thereby generating significant knowledge spillovers. Additionally, DFA exhibits Jacobs externalities by facilitating close interactions among different industries, leading to external economies of scale, fostering cross-industry technological integration and innovation, which accelerates technology iteration and drives industrial upgrading (Zhang et al., 2018). In summary, DFA produces substantial economic effects through mechanisms such as knowledge spillovers, sharing economy, and specialized division of labor, which further trigger technological transformations, promote the emergence of new products and production technologies, and support secondary innovations and long-term economic growth (Schaefer et al., 2014).

1.3 Ecological and Environmental Effects of DFA

Data factors play a crucial role in pollution control (Guo et al., 2024). On the one hand, data agglomeration directly stimulates green innovation vitality; on the other hand, it indirectly promotes green innovation and green

development by enhancing government support and raising public environmental awareness (Han et al., 2024). Meanwhile, DFA facilitates the development of urban green finance by driving local industrial upgrading and expanding the scale of green finance (Wang et al., 2024). The development of green finance supports low-carbon urban development through green financing. Moreover, DFA promotes the construction of local digital infrastructure, which further encourages green technological innovation and alleviates financing constraints, thereby fostering green development (Guo et al., 2024). From a broader perspective, existing studies also reveal that the development of the digital economy can lead to industrial green transformation (Yang et al., 2025), improve green total factor productivity at the provincial level (Qian et al., 2024), and enhance the efficiency of urban green innovation (Chen et al., 2025). Furthermore, digitalization may help reduce industrial solid waste emissions by improving resource allocation efficiency and promoting technological progress, thereby presenting an inverted U-shaped relationship between digital development and environmental pressure (Zhang & Wang, 2025). Overall, the agglomeration of data factors reflects the digital economy's profound transformation of traditional development models and contributes to ecological transition through optimized resource allocation, expanded green finance, and enhanced incentives for green innovation.

A review of the related literature shows that although scholars widely recognize data as a new factor of production—following traditional inputs such as labor and capital—and acknowledge its role in promoting economic development through integration with conventional factors, research on the impact of data factors on the green development of traditional firms and the underlying mechanisms remains limited. Moreover, while existing studies from the perspective of economic geography have examined the economic effects of DFA and, at a macro level, its environmental and ecological consequences, relatively little attention has been paid to micro-level environmental outcomes. In particular, the literature has yet to systematically integrate economic geography with firm-level analysis to investigate the spatial effects of DFA on firms' green development.

Against this background, this study adopts a micro-level perspective to examine the impact of DFA on the green development of traditional manufacturing firms and to further explore its spatial effects. The marginal contributions of this paper are fourfold. First, at the theoretical level, we construct a micro-founded analytical model to systematically elucidate the internal mechanisms through which DFA promotes firms' green transformation. Second, at the empirical level, exploiting the establishment of China's National Big Data Comprehensive Pilot Zones as a quasi-natural experiment, we employ a difference-in-differences framework to identify the causal effects of DFA, and further introduce a double machine learning (DML) approach to conduct robustness checks, thereby alleviating potential model misspecification bias and improving the accuracy and reliability of the estimates. Third, in terms of transmission channels, we verify that DFA facilitates green development through three primary pathways—accelerating IT, enhancing capacity utilization (CU), and improving energy efficiency—and conduct heterogeneity analyses along firm characteristics, regional conditions, and industry attributes to uncover differential effects. Fourth, from a spatial perspective, by constructing a spatial Durbin model, we find that DFA generates a "siphoning effect" on geographically adjacent regions, while no significant spatial spillovers are observed across economically similar regions, providing new evidence on the spatial externalities of data factors. Overall, these findings help clarify the boundary conditions under which data factors drive green development, and offer both theoretical insights and empirical evidence to support the formulation of differentiated policy interventions.

2. Theoretical Analysis and Research Hypotheses

This paper first constructs a micro-level mathematical model, drawing on the theoretical framework of Jones & Tonetti (2020), by incorporating data factors and energy inputs into the production function and setting a consumption environment. Through comparative static analysis, it investigates the impact and underlying mechanisms of DFA on the green and sustainable development of traditional firms. The theoretical analysis is further developed to clarify the mechanisms and propose corresponding hypotheses.

2.1 Theoretical Model

2.1.1 Consumer assumptions

Given the heterogeneous preferences of consumers for products, the utility function is assumed to be a Constant Elasticity of Substitution (CES) function: $U \equiv \left(\int_{i\in\Omega} q(i)^{\frac{\sigma-1}{\sigma}} di\right)^{\frac{\sigma}{\sigma-1}}$, where $q(i)$ is the consumption of product i by a representative consumer, and $\sigma > 1$ is the elasticity of substitution between goods. The price of product i is $p(i)$, and the consumer's total income is $R = \int_{i\in\Omega} p(i)\, q(i)\, di = PQ$, where P is the aggregate price index expressed as: $P = \left(\int_{i\in\Omega} p(i)^{1-\sigma} di\right)^{\frac{1}{1-\sigma}}$. Under symmetric equilibrium, the demand function of the representative consumer is:

$$q(i) = \frac{RN^{\sigma-1}}{p(i)} \tag{1}$$

where, N denotes the number of product varieties or firms.

2.1.2 Producer assumptions

Following Jones & Tonetti (2020), information is defined as a collection of non-rival economic goods. In this context, ideas, serving as instructions or blueprints for manufacturing goods, represent a form of information (Romer, 1990). Fundamentally, firm production consists of ideas together with required energy and material inputs. Hence, the firm's production function is assumed as: $y \equiv Idea^a \cdot E^b, a + b = 1$. Where $Idea^a$ represents ideas, E denotes energy and other material inputs, and each firm is associated with a unique idea and produces a unique product. Meanwhile, data, as the carrier and basic unit of information, though non-directional and not instructions for manufacturing, is useful in production by driving knowledge creation and economic growth (Agrawal et al., 2019). The idea function is further specified as: $Idea \equiv K \cdot data$. Where K denotes knowledge level, higher knowledge levels make the same data more valuable, generating new ideas. The final production function becomes:

$$y = (K \cdot data)^a \cdot E^b \tag{2}$$

Data not only assists product manufacturing but also generates new data through consumption and production activities (Bergemann and Bonatti, 2019). Thus, data is modeled as:

$$data = \beta \cdot \tau y + (1 - \beta) \cdot \bar{\tau} B \tag{3}$$

where, $B \equiv (N-1)y$ is the aggregate consumption of other firms' products under symmetric equilibrium. This equation means that a firm's own product consumption y contributes data information proportional to data flow coefficient τ, while the aggregated consumption B of other local firms similarly contributes new data according to data flow coefficient $\bar{\tau}$. Parameter β measures the relative importance of own-product data compared to others. When data is fully enclosed, $\bar{\tau} = 0$, implying firms can only use their own data for production assistance. In regions with DFA, due to agglomeration externalities, data tends to be shared, so $\bar{\tau} > 0$, generating external economies of scale. Integrating Eqs (2) and (3), the final production function is obtained as follows:

$$y = K^{\frac{a}{b}}[\beta\tau + (1-\beta)\bar{\tau}(N-1)]^{\frac{a}{b}}E \tag{4}$$

The initial production function assumes constant returns to scale, but the non-rivalry of knowledge K leads to constant returns to energy and material inputs E, resulting in increasing returns to scale overall. Furthermore, Eq (4) reflects additional scale effects induced by DFA: the more firms in the same region, the more data generated, and with $\bar{\tau} > 0$, output increases. Thus, the non-rival nature of data may bring substantial economic effects (Jones & Tonetti, 2020).

2.1.3 Equilibrium and comparative static analysis

Based on the above assumptions, the firm's profit maximization problem is described as follows:

$$\begin{gathered} max\ \pi_i = p_i y_i - wE_i \\ s.t. \\ P_i = \frac{RN^{\sigma-1}}{q_i} \\ y_i = K_i^{\frac{a}{b}}[\beta\tau + (1-\beta)\bar{\tau}(N-1)]^{\frac{a}{b}}E_i \end{gathered} \tag{5}$$

Solving the above optimization model, the market average time price is obtained: $p^* = \mu w/\psi$; Output: $y^* = \psi RN^{\sigma-1}/\mu w$; Energy and raw material demand: $E^* = RN^{\sigma-1}/\mu w$. Where $\psi = K^{\frac{a}{b}}[\beta\tau + (1-\beta)\bar{\tau}(N-1)]^{\frac{a}{b}}, \mu = \sigma/(\sigma-1)$. For enterprises, their pollution emissions mainly come from energy consumption and raw material processing process, so this paper assumes that the emission function of enterprises is: $y_u^* \equiv \sum_{i=1}^{n} \theta_i E^*, \theta_i > 0$. The coefficient θ_i represents the ratio of various types of emissions generated when a firm consumes factor E. Next, we define the level of green development of enterprises: $GT^* \equiv y_u^*/y^*$. GT^* represents the total amount of emissions per unit of output, and the smaller the value is, the higher the degree of green development of the enterprise is.

In DFA regions, the data flow coefficient $\bar{\tau}$ among firms increases. Comparative static analysis of GT^* with respect to $\bar{\tau}$ yields:

$$\frac{\partial GT^*}{\partial \bar{\tau}} = \frac{\partial GT^*}{\partial y_u^*}\frac{\partial y_u^*}{\partial \bar{\tau}} + \frac{\partial GT^*}{\partial y^*}\frac{\partial y^*}{\partial \bar{\tau}} < 0 \tag{6}$$

This shows that DFA is conducive to the green development of enterprises. Based on the above theoretical model analysis, this paper puts forward a hypothesis.

H1: DFA can promote the green development of traditional manufacturing firms

To further explore the mechanisms, we focus on two key variables: equilibrium output and energy efficiency. Energy efficiency is defined as output per unit energy input: $EEI^* \equiv y^*/E^*$, and the comparative static analyses are:

$$\begin{cases} \dfrac{\partial y^*}{\partial \bar{\tau}} = \dfrac{a\psi(1-\beta)(N-1)RN^{\sigma-1}}{[\beta\tau + (1-\beta)\bar{\tau}(N-1)]b\mu w} > 0 \\ \dfrac{\partial EEI^*}{\partial \bar{\tau}} = \dfrac{a\psi}{\left[\dfrac{b\beta\tau}{(1-\beta)(N-1)} + b\bar{\tau}\right]} > 0 \end{cases} \tag{7}$$

These results indicate that both equilibrium output and energy efficiency improve under conditions of DFA.

2.2 Mechanism Analysis

Based on the derivations of Eqs (6) and (7), DFA not only facilitates firms' green transformation but also significantly enhances production output and energy efficiency. Accordingly, this study summarizes the mechanisms through which DFA affects firms' green development into three main channels.

First, DFA promotes green development by accelerating firms' IT. The core of IT lies in the deep integration of data factors into all stages of firms' production and operational processes. With the accumulation and application of data resources, firms are able to establish data-driven intelligent production modes. For instance, real-time analysis of industrial big data enables production equipment to achieve self-optimization by automatically adjusting operating parameters in response to changing working conditions, thereby significantly improving production precision and reducing material waste (Aghion & Jaravel, 2015). Moreover, predictive maintenance systems based on machine learning algorithms can accurately assess equipment conditions and effectively prevent energy losses caused by unplanned downtime (Smichowski et al., 2023). Such end-to-end IT fundamentally enhances the efficiency of resource allocation and energy utilization, laying a technological foundation for green development. In addition, the accumulation of data factors substantially strengthens firms' capacity for green information processing. By establishing environmental data platforms and carbon accounting systems, firms can realize real-time collection and intelligent analysis of environmental indicators such as energy consumption and emissions, transforming previously fragmented and lagged environmental information into systematic and quantifiable decision-making inputs, thereby markedly improving the precision and timeliness of environmental management.

Second, DFA fosters green development by increasing CU. The distinctive attributes of data factors provide new possibilities for optimizing capacity allocation. Owing to their non-rivalry and replicability, data factors can be shared across different production stages at near-zero marginal cost, effectively easing the scarcity constraints associated with traditional production factors (Karami & Igbokwe, 2025). In practice, data-driven intelligent scheduling systems can coordinate the operating rhythms of different production units in real time, reduce waiting time between processes, and maintain production lines at optimal load levels. Such refined capacity management not only lowers fixed costs per unit of output but, more importantly, reduces resource waste caused by idle capacity or overproduction, thereby realizing more intensive and environmentally friendly production.

Third, DFA contributes to green development by improving energy efficiency. The application of data factors enables firms' energy management practices to shift from extensive to more refined modes. Through the deployment of sensor networks and the construction of energy management platforms, firms can monitor energy use in real time and conduct in-depth analyses. These data-driven insights allow firms to dynamically adjust energy consumption strategies in line with production needs, thereby avoiding unnecessary energy use (Diamantoulakis et al., 2015). Furthermore, improved data infrastructure provides essential support for green innovation. Data infrastructures such as industrial internet platforms not only reduce transaction costs related to the acquisition, processing, and sharing of environmental technologies, but also accelerate the development and diffusion of green technologies by promoting data collaboration among firms, universities, and research institutions. Finally, data-sharing mechanisms help break down information barriers on energy efficiency along the supply chain, encouraging upstream and downstream firms to jointly implement energy-saving upgrades and form a system-wide energy optimization framework. This systematic approach to energy management—from monitoring and regulation to coordination—significantly enhances energy efficiency and effectively reduces carbon intensity per

unit of output.

Based on the above analysis, this paper proposes hypothesis 2.

H2: DFA promotes the green and sustainable development of traditional manufacturing firms by accelerating IT, increasing CU, and improving energy efficiency through three key channels.

3. Research Design

3.1 Model Setting

To examine the impact of DFA on firms' green development, this paper takes the establishment of national big data comprehensive experimental zones as a quasi-natural experiment and constructs the following DID model:

$$GDI_{i,t} = \alpha + \beta DFA_{i,t} + \gamma Controls_{i,t} + \lambda_i + \delta_i + \mu_i + \lambda_t + \varepsilon_{i,t} \tag{8}$$

where, i and t represent the enterprise and year respectively, and $GDI_{i,t}$ is the green development index of the enterprise. $DFA_{i,t}$ represents the aggregation of data factors, which is 1 when the city to which enterprise i belongs is approved to establish the national comprehensive experimental area of big data in year t, otherwise it is 0. ε is the random error term, $\lambda_i, \delta_i, \mu_i$ and λ_t represent industry fixed effect, province fixed effect, firm fixed effect, and time fixed effect respectively, and $Controls_{i,t}$ represent control variables. In this paper, the standard errors of regression coefficients in the model are clustered at the city level.

Since the impact of DFA and control variables on green development may be nonlinear, to avoid model misspecification bias, we adopt a DML approach following Chernozhukov et al. (2018), and specify the following flexible interactive model for robustness:

$$GDI_{i,t} = g_0\left(DFA_{i,t}, \boldsymbol{X}_{i,t}\right) + U_{i,t} \tag{9}$$

$$DFA_{i,t} = m_0\left(\boldsymbol{X}_{i,t}\right) + V_{i,t} \tag{10}$$

where, $\boldsymbol{X}_{i,t} = \left(Controls_{i,t}, \lambda_i, \lambda_t\right)$, the first equation is the main equation, g_0 and m_0 are unknown functions, respectively representing the influence of each variable on the level of green development and the agglomeration of data factors. $U_{i,t}$ and $V_{i,t}$ are random error terms, and the rest of the variables have the same meaning as above. At this time, the average treatment effect ATE of DFA on the level of enterprise green development is defined as:

$$\theta_0^{\mathrm{ATE}} \equiv E\left[g_0\left(1, \boldsymbol{X}_{i,t}\right) - g_0\left(0, \boldsymbol{X}_{i,t}\right)\right] \tag{11}$$

In order to obtain the ATE, this paper refers to the method of Chernozhukov et al. (2018) and uses dual machine learning to solve the problem. Firstly, with the help of K-fold cross validation method, the sample data is divided into K sub-samples $(I_1, I_2, \ldots, I_K)$, and obtain its corresponding complement $(I_1^c, I_2^c, \ldots, I_K^c)$, and then use machine learning model to estimate the data on I_K^c, and use the data on I_K to predict, Get $E\left[GDI_{i,t} \mid DFA_{i,t}, \boldsymbol{X}_{i,t}\right]$ and $E\left[DFA_{i,t} \mid \boldsymbol{X}_{i,t}\right]$ estimate function $\hat{g}_{I_k^c}\left(1, \boldsymbol{X}_{i,t}\right), \hat{g}_{I_k^c}\left(0, \boldsymbol{X}_{i,t}\right)$ and $\hat{m}_{I_k^c}$, finally obtaining the estimate for the average treatment effect:

$$\hat{\theta}_n^{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[\begin{array}{c}\dfrac{DFA_{i,t}\left(GDI_{i,t_i} - \hat{g}_{I_{k_i}^c}\left(1, \boldsymbol{X}_{i,t}\right)\right)}{\hat{m}_{I_{k_i}^c}\left(\boldsymbol{X}_{i,t}\right)} - \dfrac{\left(1 - DFA_{i,t}\right)\left(GDI_{i,t} - \hat{g}_{I_{k_i}^c}\left(0, \boldsymbol{X}_{i,t}\right)\right)}{1 - \hat{m}_{I_{k_i}^c}\left(\boldsymbol{X}_{i,t}\right)} \\ + \hat{g}_{I_{k_i}^c}\left(1, \boldsymbol{X}_{i,t}\right) - \hat{g}_{I_{k_i}^c}\left(0, \boldsymbol{X}_{i,t}\right)\end{array}\right] \tag{12}$$

Among them, $\hat{g}_{I_{k_i}^c}\left(1, \boldsymbol{X}_{i,t}\right), \hat{g}_{I_{k_i}^c}\left(0, \boldsymbol{X}_{i,t}\right)$ and $\hat{m}_{I_{k_i}^c}$ is based on the estimated function $\hat{g}_{I_k^c}\left(1, \boldsymbol{X}_{i,t}\right), \hat{g}_{I_k^c}\left(0, \boldsymbol{X}_{i,t}\right)$ and $\hat{m}_{I_k^c}$ get enterprise i predicted.

3.2 Variable Selection

3.2.1 Dependent variable

Green development index (GDI). Following Zhou et al. (2022), this paper constructs a firm-level GDI based on textual analysis of annual reports. Specifically, a dictionary of 113 keywords related to green development is

compiled. The frequency of these keywords appearing in each firm's annual report is calculated, and the natural logarithm of the word count plus one is taken as a proxy for the firm's green development level.

To further enhance the robustness of the results and alleviate potential measurement errors, we construct an alternative indicator, denoted as GDI_2. Based on GDI, this alternative measure incorporates two additional dimensions: carbon productivity and the ratio of green expenses. The composite index is calculated using the entropy weighting method. Carbon productivity is measured as the firm's main operating revenue per unit of carbon emissions, while the green expense ratio is defined as the proportion of pollution discharge fees, environmental protection fees, and other related expenditures relative to operating revenue. In terms of index properties, carbon productivity is positively associated with a firm's level of green development, whereas the green expense ratio is negatively associated with green development.

3.2.2 Key independent variable

Data factor agglomeration (DFA). The core explanatory variable is a dummy variable indicating whether the city in which a firm is located has been designated as a National Big Data Comprehensive Pilot Zone. Based on the lists released by the State Council in 2015 and 2016, cities approved as pilot zones are assigned a value of 1 from the year of approval onward, and 0 otherwise.

3.2.3 Mechanism variables

Intelligent transformation (IT). A key feature of firms' IT is large-scale investment in artificial intelligence (AI)–related assets, through which production and management processes are restructured in a data-driven manner. From the perspective of firms' asset structure, this study constructs an indicator to measure IT. Specifically, based on the notes to listed firms' financial statements, we identify AI-related intangible assets (e.g., "intelligent systems" and "information platforms") and fixed assets (e.g., "servers" and "computing power equipment") through a keyword-filtering approach. These assets are aggregated to obtain total AI-related investment, which is then normalized by total assets to construct the IT index (IT), defined as the ratio of total AI investment to total assets. This indicator effectively captures the intensity of firms' resource allocation toward AI applications and reflects the depth of their transition toward an intelligent operating model.

Capacity utilization (CU). CU is defined as the ratio of a firm's actual output to its potential maximum output (Kirkley et al., 2002). This paper follows the method of (Fan et al., 2019) and applies a cost function approach to estimate CU. Specifically, it uses indicators including output (Y), capital (K), labor (L), and investment (I), and employs the Olley-Pakes (OP) method to estimate the parameters of the cost function and calculate CU. Output is measured by the main business revenue, capital by the net value of fixed assets, labor by the total number of employees, and investment by the cash paid for the construction of fixed assets, intangible assets, and other long-term assets. To eliminate the influence of price factors, output is deflated using the regional industrial producer price index, while capital is deflated using the regional fixed asset price index.

Energy-environment efficiency (EEI). The existing literature mainly measures energy efficiency from two perspectives: one is single-factor energy efficiency, namely energy intensity, which measures energy consumption per unit of output; the other is total-factor energy efficiency, which integrates various inputs and outputs into a comprehensive indicator. While energy intensity is simple to compute, it does not account for actual pollutant emissions; total-factor energy efficiency is more comprehensive but cannot isolate the impact of energy input alone due to the inclusion of other factors such as labor and capital. Therefore, this paper adopts the energy-environmental efficiency (EEI) as a proxy for energy efficiency, drawing on the work of Zhang et al. (2014). Under the DEA framework, the EEI is measured using the Energy-Environmental Non-Radial Directional Distance Function (ENDDF). In this model, energy input includes both fossil energy and electricity consumption; the desirable output is represented by main business revenue; and the undesirable outputs include air pollutants (SO_2, NO_x, and particulates) and water pollutants (chemical oxygen demand and ammonia nitrogen).

3.2.4 Control variables

Referring to the existing literature, this paper selects the control variables at the enterprise level: Leverage (Lev) is measured by the ratio of total liabilities to total assets at year-end; Profitability (Roa) is calculated as net profit divided by total assets; Growth Capacity (Growth) is derived from the year-on-year change in operating revenue normalized by prior-period revenue; Firm Size (Size) is defined as the natural logarithm of total assets; Ownership Concentration (Top1) reflects the shareholding percentage of the largest shareholder; Ownership Restriction (Balance) quantifies power constraints using the ratio of the second-largest to the largest shareholder's ownership; Control variables at the regional level: Development (Gdp) is proxied by per capita GDP (in 10,000 yuan) of the enterprise's locality; Industrialization Level (Indus) represents industrial added value as a percentage of regional GDP; Trade Openness (Open) is the total import-export volume relative to regional GDP.

3.3 Data Sources

Referring to the methods of Liang &Wang (2022) and Qu et al. (2020), this paper sets the industries not in the catalog as traditional industries according to the matching between the industries in the "Catalogue of High-tech Industries" and the "Industry Classification of National Economy (GB/T4754-2017)". Finally, A-share listed companies in Shanghai and Shenzhen from 2011 to 2022 with industry codes of C13 to C25, C28, and C30 to C32 are selected as the representatives of traditional enterprises. The city where the listed company is registered is used as the regional identification of the enterprise, and a series of data preprocessing is carried out: (1) eliminating ST and *ST enterprises; (2) Eliminating the samples with missing relevant variables; (3) The linear interpolation method is used to supplement the sample data with fewer missing data. Finally, 816 listed companies were selected, and 5931 observations were obtained.

The relevant data at the enterprise level used in this paper are mainly from the CSMAR database, the data at the regional level are from the China Statistical Yearbook, and the pollution emission data are from the China Statistical Yearbook on the Environment.

4. Analysis of Empirical Results

4.1 Descriptive Statistics

First, based on whether DFA occurred in a region, this paper divides the sample into "agglomeration areas" and "non-agglomeration areas". It then calculates the average GDI of enterprises in each area over time and plots the trend of average GDI for agglomeration and non-agglomeration regions, as shown in Figure 1. In the figure, the blue shaded area represents the average GDI of enterprises in non-agglomeration areas, while the red solid line represents the average GDI of enterprises in agglomeration areas. As shown, prior to 2015, the GDI trends of agglomeration and non-agglomeration areas were largely consistent. However, since 2015 and 2016, the GDI growth in agglomeration areas has significantly outpaced that of non-agglomeration areas, and the gap between them has narrowed rapidly. Given that the establishment of National Big Data Comprehensive Pilot Zones was rolled out in batches during 2015 and 2016, the trends illustrated in Figure 1 align closely with the timing of policy implementation, providing preliminary support for the argument that DFA promotes corporate green development.

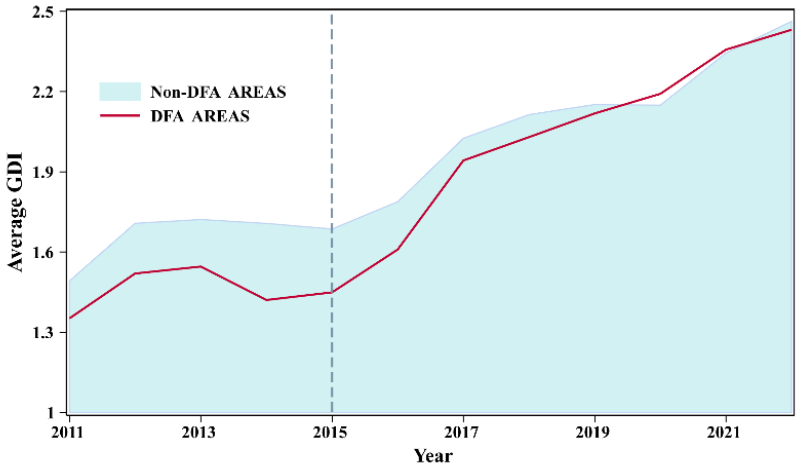

Figure 1. Trend of average GDI

Second, this paper further categorizes the sample into "heavily polluting industries" and "non-heavily polluting industries" based on industry characteristics, and combines this classification with regional categorization to generate group comparison charts of EEI and CU. The industry classification follows the list of heavily polluting industries identified in the "Guidelines for Environmental Information Disclosure of Listed Companies" issued by China's Ministry of Environmental Protection. This includes 16 industries such as textiles, metallurgy, brewing, building materials, pharmaceuticals, petrochemicals, thermal power, steel, tanning, electrolytic aluminum, chemicals, coal, cement, fermentation, papermaking, and mining. The corresponding industry codes for listed companies are B06–B12, C13, C15, C17, C19, C22, C25–C32, D44, and D45. Based on this classification, companies are divided into heavily polluting (HP) and non-heavily polluting (NHP) groups. Meanwhile, regional categorization follows the earlier division into agglomeration and non-agglomeration areas.

As shown in Figure 2, the average EEI and CU in heavily polluting industries are significantly lower than those

in NHOP industries, suggesting that industries with lower pollution levels tend to achieve higher CU and energy efficiency. In contrast, heavily polluting industries are more likely to suffer from outdated capacity and lower energy efficiency, which is consistent with real-world observations. Furthermore, Figure 2 shows that firms located in agglomeration areas exhibit higher average levels of EEI and CU compared to those in non-agglomeration areas, providing additional evidence that DFA contributes to improvements in energy efficiency and CU.

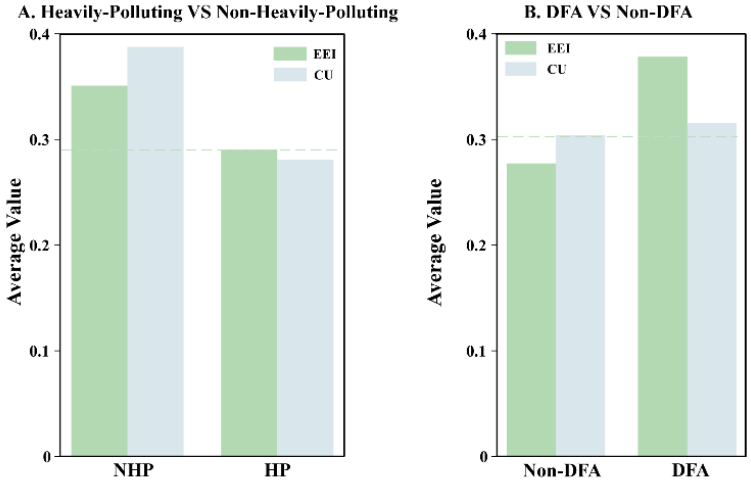

Figure 2. Group comparison of EEI and CU

Third, descriptive statistics of the main variables are presented in Table 1. The results show that the average value of the GDI is 1.965, with a standard deviation of 0.851, indicating substantial variation in green development levels across firms in the sample.

Table 1. Descriptive statistics

Var Name	Obs	Mean	SD	Min	Median	Max
GDI	5931	1.965	0.851	0.000	1.946	4.727
DFA	5931	0.185	0.388	0.000	0.000	1.000
Lev	5931	0.413	0.205	0.007	0.398	1.718
Roa	5931	0.039	0.077	-0.965	0.038	0.482
Growth	5931	0.305	5.407	-0.942	0.084	363.068
Size	5931	22.212	1.269	18.674	22.028	26.710
Top1	5931	0.356	0.150	0.034	0.336	0.900
Gdp	5931	7.486	3.527	1.602	6.761	18.999
Indus	5931	0.350	0.074	0.100	0.365	0.574
Open	5931	0.418	0.307	0.008	0.368	1.464

4.2 Baseline Regression Analysis

Table 2 reports the baseline regression results. Column (1) presents the estimation without control variables and without including any fixed effects. Column (2) extends Column (1) by introducing firm and year fixed effects, while Column (3) further incorporates a set of firm-level and regional control variables. The results indicate that under all three specifications, the estimated coefficient of DFA is positive and statistically significant at least at the 5% level, providing preliminary support for the hypothesis that DFA facilitates green development in traditional firms.

To further strengthen the robustness of the findings, Column (4) augments Column (1) by controlling for year, province, and industry fixed effects, and Column (5) additionally includes the full set of control variables. The estimated coefficient of DFA remains positive and statistically significant at the 5% level, suggesting that the promoting effect of DFA on firms' green development persists after accounting for industry heterogeneity and regional differences. Moreover, we re-measure firms' green development using the alternative indicator GDI2, with the corresponding regression results reported in Column (6). The coefficient of DFA remains positive and significant at the 5% level, thereby further confirming the robustness of the baseline findings and providing initial support for Hypothesis 1.

Table 2. Baseline regression results

	(1)	(2)	(3)	(4)	(5)	(6)
	GDI	GDI	GDI	GDI	GDI	GDI_2
DFA	0.192***	0.147**	0.149**	0.127***	0.106**	0.006**
	(3.292)	(2.422)	(2.475)	(2.878)	(2.072)	(2.379)
_cons	1.930***	1.939***	-2.259**	1.942***	-1.146**	-0.151
	(55.030)	(172.597)	(-2.446)	(93.696)	(-2.493)	(-0.817)
Controls	No	No	Yes	No	Yes	Yes
Year FE	No	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	No	No
Province FE	No	No	No	Yes	Yes	Yes
Industry FE	No	No	No	Yes	Yes	Yes
Observations	5931	5931	5931	5931	5931	5931
R^2	0.008	0.709	0.714	0.324	0.363	0.494

Note：*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively. T-statistics are reported in parentheses and clustered at the city level. The same is below.

4.3 Robustness and Endogeneity Tests

In order to ensure the robustness of the empirical findings, this paper conducts a series of robustness checks, including parallel trend testing, DML, placebo tests, Propensity Score Matching-DID (PSM-DID), and Instrumental Variable (IV) methods.

4.3.1 Robustness tests

(1) Parallel trend test

The parallel trend assumption is a prerequisite for the validity of the DID estimation. As shown in Figure 1, prior to 2015, the average GDI trends for the aggregation and non-aggregation regions are nearly identical, indicating no significant pre-treatment divergence between the treated and control groups. To further validate this assumption, this paper adopts the event study approach following. Figure 3 reports the estimated coefficients and corresponding 95% confidence intervals before and after policy implementation. The year of policy implementation is designated as time 0, with the year prior serving as the baseline, and its coefficient normalized to zero. The results show that in the four years prior to the policy shock, the estimated coefficients are close to zero, and the 95% confidence intervals cross the zero axis, failing to reject the null hypothesis of no difference. This supports the absence of a pre-existing trend in GDI between the treatment and control groups. After the policy implementation, the coefficients exhibit an upward trend, becoming significantly positive at the 10% level in the third year and at the 5% level from the fourth year onwards, indicating that data factor aggregation significantly promotes corporate green development.

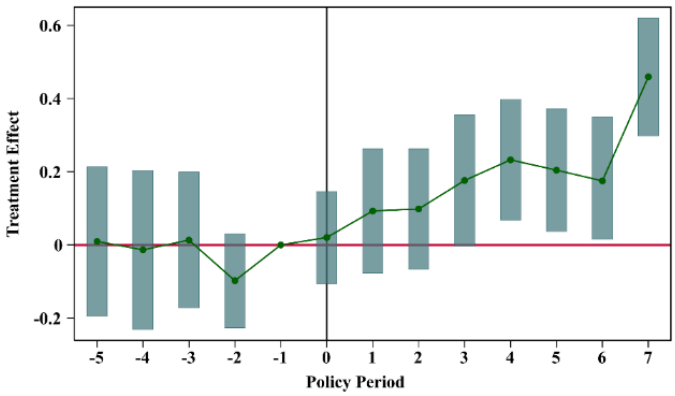

Figure 3. Parallel trend test

(2) Double machine learning

This paper further employs a double/debiased machine learning model for robustness checks. First, variable

selection: the covariates are chosen to be the same control variables as in the baseline regression, with the addition of individual and time fixed effects to account for inherent differences across entities and common time trends. Second, model training method: following the approach of Chernozhukov et al. (2018), a 5-fold cross-validation method is adopted. The sample data is divided into five subsamples ($I_1, I_2, \ldots, I_5$), and their corresponding complements ($I_1^c, I_2^c, \ldots, I_5^c$) are obtained. The machine learning model is then trained on the data in I_i^c to make predictions for the corresponding I_i. Third, model selection: considering that a single machine learning method may be subject to model selection bias, this paper adopts a model stacking approach, integrating four algorithms—Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN), and Gradient Boosting Machine (GBM). Weights are automatically assigned through cross-validation to construct a combined predictive model.

The regression results are presented in Table 3. Column (1) reports the average treatment effect (ATE) estimated by double/debiased machine learning, while column (2) shows the results from the standard DID method. The results indicate that the ATE of DFA is 0.117, significantly positive at the 1% level. The effect of DFA is slightly lower than the standard DID estimate of 0.149, suggesting that, even when avoiding model specification errors, the positive impact of DFA on corporate green development remains robust.

Table 3. Double machine learning interaction model regression

	(1)	(2)
	GDI	GDI
DFA	0.117***	0.149**
	(5.087)	(2.475)
Controls	Yes	Yes
Year FE	Yes	Yes
Firm FE	Yes	Yes
Observations	5931	5931

(3) Placebo test

Further placebo tests are conducted, including time-placebo, spatial-placebo, and combined-placebo tests. First, the time-placebo test results are shown in Figure 4. When the policy shock is artificially lagged by 1 to 4 years, the 95% confidence intervals of all placebo effects include zero, indicating no significant placebo effects. Hence, we fail to reject the null hypothesis of zero placebo effect. Next, Table 4 presents the results of the spatial-placebo test. From 816 listed companies, pseudo-treatment and pseudo-control groups are randomly drawn, and DID estimations are repeated 500 times. The actual estimated treatment effect lies in the far-right tail of the placebo distribution, with a two-tailed p-value of 0.006 and a right-tailed p-value of 0.002, both significant at the 1% level. In the end, in the mixed placebo test, pseudo-treatment groups are randomly drawn along with randomly assigned placebo treatment years. Again, 500 iterations of DID are conducted. The two-tailed and right-tailed p-values are both 0.000, strongly rejecting the null hypothesis of zero treatment effect and further confirming the robustness of the main results.

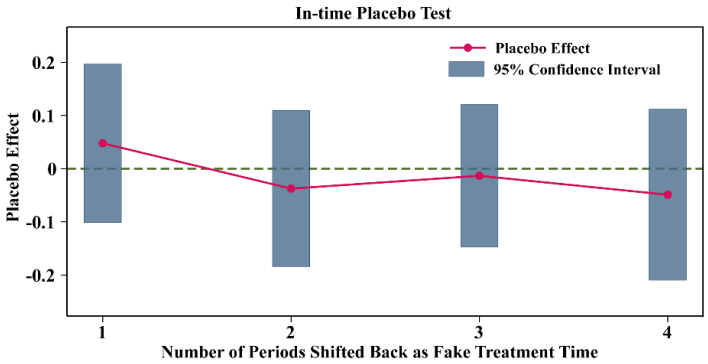

Figure 4. In-time placebo test

Table 4. Spatial and mixed placebo test results

	Treatment Effects	Bilateral P-Values	Left P-Values	Right P-Values
Spatial placebo test	0.149	0.006***	0.998	0.002***
Mixed placebo test	0.149	0.000***	1.000	0.000***

4.3.2 Endogeneity tests

To address potential endogeneity concerns, this paper employs the PSM-DID and IV methods.

First, we construct the propensity score using firm-level control variables and perform 1:1 nearest neighbor matching with replacement. Subsequently, we execute the DID estimation on the matched sample. As shown in columns (1) and (2) of Table 5, the coefficient on DFA remains significantly positive at the 5% level, supporting the main findings after accounting for sample selection bias.

Second, the paper uses regional topographic relief as an IV. On the one hand, data factor aggregation depends on the supporting infrastructure, which is significantly affected by terrain; greater relief increases infrastructure construction and operating costs and destabilizes digital signal transmission. On the other hand, terrain, as a geographic characteristic, is exogenous to socio-economic conditions and less likely to directly affect corporate green development. Thus, topographic relief satisfies both the relevance and exclusion conditions for a valid IV.

Two-stage IV regression results are reported in columns (3) and (4) of Table 5. In the first stage, topographic relief is significantly negatively associated with DFA at the 1% level. The Kleibergen-Paap rk LM statistic is significant at the 1% level, passing the under-identification test. The Kleibergen-Paap rk Wald F-statistic is 17.596, exceeding the 10% Stock-Yogo threshold, indicating the instrument is not weak. In the second stage, the DFA coefficient remains significantly positive at the 5% level, further validating the robustness of the main findings.

Overall, after controlling for endogeneity, the positive effect of data factor aggregation on corporate green development remains significant, lending strong support to Hypothesis 1.

Table 5. Endogenous tests

	(1)	(2)	(3)	(4)
	PSM-DID	PSM-DID	First-stage	Second-stage
Degree of relief			-0.190***	
			(-4.195)	
DFA	0.282**	0.313**		1.271**
	(2.284)	(2.434)		(2.207)
Controls	No	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Observations	1721	1721	5864	5864
K-P Wald_F			17.596	
K-P LM			10.415***	

4.4 Mechanism Tests

In order to further explore the mechanism through which the agglomeration of data factors influences the green development of traditional enterprises, this study draws on the methodology proposed by Jiang (2022). On the basis of identifying the causal relationship between the core explanatory variable and the explained variable, it further investigates the effect of the explanatory variable on the mediating variable, and theoretically analyzes the pathway through which the mediating variable affects the explained variable, thereby testing the validity of the mediation effect. The empirical models are specified as follows:

$$GDI_{i,t} = \alpha + \beta DFA_{i,t} + \gamma Controls_{i,t} + \lambda_i + \delta_i + \lambda_t + \varepsilon_{i,t} \quad (13)$$

$$M_{i,t} = \eta + \mu DFA_{i,t} + \varphi Controls_{i,t} + \lambda_i + \delta_i + \lambda_t + \varepsilon_{i,t} \quad (14)$$

where, $M_{i,t}$ represents the mediating variables, including IT, CU, and EEI.

It is noteworthy that, although capacity utilization is widely regarded as an important indicator of resource allocation efficiency, its improvement does not inherently imply better green performance. Essentially, capacity utilization reflects the extent to which a firm uses its existing production capacity, and its environmental impact depends on the energy structure, production technology, and pollution control measures employed during capacity utilization. Therefore, even if a firm's CU increases, green development goals may remain elusive if production continues to rely on high-energy-consumption or high-pollution processes. Based on this, after assessing the impact of DFA on CU and EEI, this paper further analyzes the effect of CU on EEI to test whether capacity utilization constitutes a key mediating mechanism through which DFA affects green development.

Table 6 reports the results of the mechanism tests. Column (2) shows that the coefficient of DFA on IT is significantly positive at the 1% level, indicating that DFA effectively promotes firms' IT. In columns (3) and (4), the coefficients of DFA on EEI and CU are significantly positive at the 1% and 10% levels, respectively. Furthermore, column (5) shows that the coefficient of CU is significantly positive at the 1% level, suggesting that data factors simultaneously enhance energy efficiency and optimize capacity utilization.

Table 6. Analysis of impact mechanisms

	(1)	(2)	(2)	(3)	(4)
	GDI	IT	EEI	CU	EEI
DFA	0.106**	0.002***	3.203***	0.029*	
	(2.072)	(2.748)	(2.635)	(1.764)	
CU					13.895***
					(3.120)
_cons	-1.146**	0.004	7.703	1.167***	-6.117
	(-2.493)	(0.815)	(0.857)	(11.779)	(-0.611)
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Year FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Province FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Industry FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	5931	5931	5931	5928	5928
R^2	0.363	0.057	0.516	0.208	0.544

These results validate the three transmission mechanisms proposed in this paper: First, DFA promotes green development through IT. IT enables firms to optimize production processes based on data analysis and algorithms, improving resource utilization efficiency and reducing environmental burdens directly from production processes. Second, DFA promotes green development by improving CU. Data-driven precise scheduling and resource allocation reduce idle equipment and production bottlenecks, allowing firms to achieve more output with the same energy input, thereby lowering carbon intensity per unit of output. Third, DFA promotes green development by enhancing energy efficiency. Real-time data monitoring and energy usage analysis allow firms to dynamically optimize energy structures, achieving energy conservation and emission reduction.

In practice, taking Guangzhou as an example, as a national comprehensive big data pilot zone, the city has promoted extensive integration of industrial enterprises into industrial internet platforms. A typical firm, Zhijing Technology, has achieved fully digitalized production management through data systems, significantly improving equipment utilization while reducing energy consumption per unit of output. This case confirms the internal logic through which data factors promote green development via multiple pathways.

In summary, both the empirical results and the case study indicate that DFA systematically promotes firms' green and low-carbon development through three mechanisms: driving IT, optimizing CU, and enhancing energy efficiency. Hypothesis 2 is thus confirmed.

4.5 Heterogeneity Analysis

Building on the extended production function incorporating knowledge and data factors, this study posits that firms' capacity to absorb and utilize data factors in the process of green transformation is significantly shaped by both internal and external conditions. Internally, the knowledge structure of the management team largely determines the firm's ability to understand and apply data resources. Executives with technical backgrounds often possess superior capabilities in recognizing and transforming digital technologies, thus playing a more proactive role in promoting the green transition. Additionally, state-owned enterprises (SOEs), due to institutional guarantees and resource allocation advantages, may achieve higher efficiency in both data acquisition and green governance.

Externally, fiscal support serves as a crucial government tool to incentivize enterprise transformation by easing financial burdens. Environmental regulation, on the other hand, may have dual effects: it may stimulate innovation through a "compliance-forcing" mechanism consistent with the Porter Hypothesis, but may also suppress firm vitality if overly stringent or poorly implemented. Moreover, the pollution intensity of the firm's industry constitutes another key driver of green transformation. Heavily polluting industries face greater regulatory pressure and have stronger incentives to build environmental capabilities through green upgrading, thereby gaining a competitive edge.

Accordingly, this paper conducts heterogeneity analysis from five dimensions: executives' technical background, ownership type, regional fiscal support, environmental regulation intensity, and industry pollution attributes. The specific classification criteria are as follows:

(1) Executives' Technical Background: Following Wang et al. (2023), we construct a proxy based on the academic majors of senior managers. Those with majors containing keywords such as "software," "information," "intelligence," "systems," "electronics," "communication," "wireless," "computer," "network," or "automation" are classified as having a technical background.

(2) Ownership Structure: Firms are categorized as state-owned or non-state-owned based on ownership status.

(3) Environmental Regulation Intensity: Drawing on Chen et al. (2018), we analyze local government work reports and measure environmental regulation intensity by calculating the frequency of environment-related keywords as a proportion of the total text. Regions are then divided into high and low regulation groups based on the median.

(4) Fiscal Support: Regional fiscal support is proxied by the ratio of general public budget expenditure to regional GDP. Regions are divided into high- and low-support groups using the sample median.

(5) Industry Pollution Attribute: Firms are categorized as either heavily polluting or non-polluting, following the aforementioned industry classification standard.

Figure 5 presents the regression results of DFA on the GDI across these five heterogeneity groups. From top to bottom, Groups 1 to 5 correspond to executive background, ownership type, environmental regulation intensity, fiscal support, and industry pollution level, respectively. The solid dots represent DFA coefficients for each subgroup, while green and yellow lines indicate the 95% confidence intervals for the "Low-Level" and "High-Level" groups, respectively.

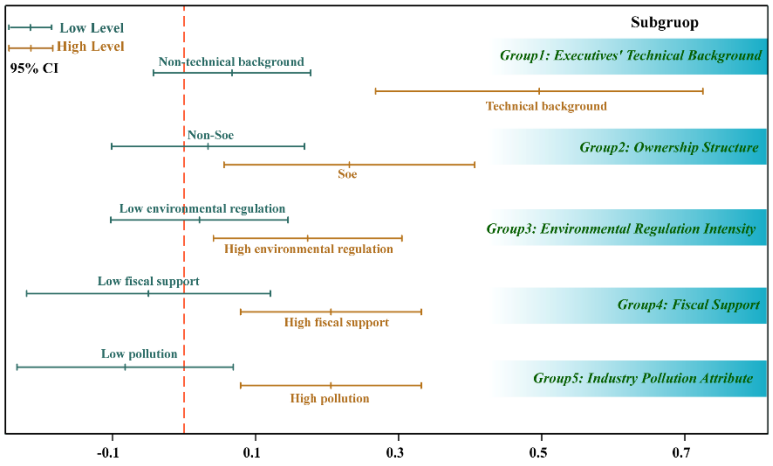

Figure 5. Heterogeneity analysis

The results reveal that in the Low-Level groups, the effects of DFA on GDI are generally insignificant or even negative. In contrast, the coefficients in the High-Level groups are all significantly positive at the 5% level, indicating that the impact of DFA on green development is more pronounced in firms with technical executives, SOEs, regions with strong environmental regulation, strong fiscal support, and heavily polluting industries.

The heterogeneity analysis indicates that the impact of DFA on green development varies significantly across different firm and regional characteristics. This variation may be attributed to several key factors. Specifically, firms led by executives with technical backgrounds tend to exhibit stronger data literacy and awareness of digital applications. These executives are better positioned to identify the value of data resources and integrate them into strategic management and green transformation efforts; in data-rich environments, they are more inclined to adopt intelligent tools and optimize production processes, thereby improving energy efficiency and resource allocation. Meanwhile, state-owned enterprises benefit from advantages in resource access, policy responsiveness, and strategic orientation. They not only enjoy greater external support in accessing data platforms—particularly within state-led green and digital initiatives—but also face stronger institutional pressures to fulfill environmental responsibilities, with green performance often embedded in their performance evaluation systems. In areas with stricter environmental regulations, firms are subject to greater external pressure to enhance environmental performance. At the same time, data factors offer the technological means for precise environmental monitoring and governance, facilitating a positive synergy between regulatory enforcement and digital transformation. For firms in regions with stronger fiscal support, the marginal effects of data factors are further amplified, as public investment in data infrastructure and industrial digitization reduces the cost and risk of adopting digital technologies, thereby unlocking their green potential. Finally, in heavily polluting industries, the driving effect of data factors on green development is particularly pronounced, due to both higher transformation pressure and wider application scenarios for data technologies in emission monitoring, resource allocation, and energy conservation. In contrast, such effects appear to be more limited in less-polluting sectors.

In summary, the heterogeneity analysis underscores the varying effects of DFA across different firm types and regional contexts. These findings provide important implications for the differentiated and targeted design of green development policies.

5. Further Analysis

Within the theoretical framework of New Economic Geography, the agglomeration of data factors is regarded as a key manifestation of the spatial evolution of the digital economy. When data factors exhibit spatial

concentration, driven by economies of scale and network effects, the marginal value of data increases with its level of agglomeration. This, in turn, attracts more firms, capital, and talent to concentrate in "data core zones," giving rise to a Matthew effect in regional development. In this process, data agglomeration exhibits a strong centripetal force, drawing in resources, factors, demand, and innovation toward the core, thus reinforcing its dominant position and generating a pronounced "siphon effect." However, as data access costs, land costs, and labor costs continue to rise in core areas, some firms may face marginalization or be crowded out due to intensified competition, giving rise to centrifugal forces. These forces may promote the diffusion of data resources, services, and technological capabilities to surrounding areas, thereby generating spillover effects under certain conditions. Especially when core platforms achieve a first-mover advantage in data governance and technological capability, they may extend their influence to peripheral regions through mechanisms such as data openness, platform sharing, or technology outsourcing, thus contributing to the digital transformation of the broader region.

Taken together, the spatial agglomeration of data factors is influenced by both centripetal and centrifugal dynamics, giving rise to complex spatial-economic effects. To further explore the potential economic consequences of these dynamics, this paper adopts a spatial econometric perspective and examines the spatial spillover and diffusion mechanisms associated with DFA. The following Spatial Durbin Model (SDM) is constructed:

$$GDI_{i,t} = \alpha + \rho GDI_{i,t} + \beta DFA_{i,t} + \gamma W \cdot DFA_{i,t} + \varphi Controls_{i,t} + u_i + v_t + \varepsilon_{i,t} \quad (15)$$

where, ρ is the spatial autocorrelation coefficient, capturing the influence of local firms' green development on neighboring regions; W denotes the spatial weight matrix; and $W \cdot DFA_{i,t}$ represents the spatial lag of DFA. The control variables are consistent with those in Model (8). The focus of this study lies in the coefficients ρ and γ, which measure the spatial impact of green development and DFA, respectively.

The regression results are presented in Table 7. Column (1) uses a geographic adjacency weight matrix W_1, while column (2) employs a weight matrix based on economic distance, following the method of Shao et al. (2016), where the weight elements are calculated as the inverse cube of the absolute difference in the annual average per capita GDP between cities. The results indicate that, under different spatial weight settings, the direct effect of DFA remains significantly positive, suggesting that DFA continues to significantly promote the green development of local traditional enterprises even when spatial correlations are taken into account. However, the spatial effects of DFA exhibit clear heterogeneity:

Table 7. Spatial econometric regression results

	(1) Geographical distance	**(3)** Economic distance
DFA	0.230***	0.096**
	(2.981)	(1.985)
W·DFA	-0.176**	0.040
	(-2.324)	(0.705)
ρ	-0.002	-0.041*
	(2.200)	(-1.735)
Direct-DFA	0.230***	0.095*
	(2.978)	(1.951)
Indirect-DFA	-0.163**	0.025
	(-2.019)	(0.386)
Total-DFA	0.067	0.121*
	(1.222)	(1.751)
Controls	Yes	Yes
Year FE	Yes	Yes
Firm FE	Yes	Yes
Observations	1416	1416
R^2	0.041	0.006

From the geographic proximity perspective, the regression results based on geographic distance weights show that the coefficient of $W \cdot DFA$ is significantly negative at the 5% level, and the indirect effect is also significantly negative. This indicates that DFA produces a notable "siphon effect" on geographically adjacent regions—that is, the concentration of data factors in core areas actually suppresses green development in surrounding regions. This phenomenon may stem from the high mobility of data factors, making neighboring regions more susceptible to resource extraction from the core areas, leading to a concentration of high-quality factors such as talent and capital in the core region, and generating negative spatial spillovers for adjacent areas.

From the economic similarity perspective, the results based on economic distance weights show that the coefficient of $W \cdot DFA$ is positive but not significant, and the indirect effect is likewise not significant. This

suggests that DFA does not generate significant spatial spillovers among regions with similar economic characteristics. This result may reflect that the diffusion of data factors relies more on geographic proximity, and similarity in economic structure alone is insufficient to facilitate cross-regional data factor flows. It also implies that the market-based allocation mechanism for data factors in China is not yet fully developed, and cross-regional mobility of data factors still faces institutional and technical barriers.

The estimated spatial autocorrelation coefficient ρ provides additional evidence for understanding regional interaction patterns. Under the geographic distance weight matrix, ρ is not significant, indicating that green development levels of firms do not exhibit significant spatial dependence across geographically neighboring regions—that is, a region's green performance is not directly influenced by its geographic neighbors. This finding reinforces the interpretation of the "siphon effect" as the dominant mechanism: the data core region primarily extracts resources from surrounding areas rather than generating knowledge spillovers, preventing neighboring regions from forming a collaborative pattern of green development. In contrast, under the economic distance weight matrix, ρ is significantly negative at the 10% level, revealing that green development among economically similar regions exhibits a "competitive" or "substitutive" spatial pattern, likely due to interregional competition in similar industrial tracks.

In summary, the spatial effects of DFA exhibit a characteristic of "geographic locality": it generates a significant siphon effect on geographically adjacent regions, while producing no significant spillover for economically similar regions. This finding provides important implications for formulating differentiated regional coordination policies: while promoting the development of DFA, attention should be paid to its negative impacts on neighboring regions, and regional coordination mechanisms should be established to mitigate the siphon effect.

6. Conclusions and Policy Recommendations

In the context of the deep integration of the "dual-carbon" strategy and the digital economy, exploring the impact of DFA on the green development of traditional enterprises holds significant theoretical and practical implications. This paper first constructs a micro-level theoretical model to reveal the intrinsic mechanisms through which DFA promotes firms' green transformation. Subsequently, taking the establishment of the National Big Data Comprehensive Pilot Zones as a quasi-natural experiment and using data from listed traditional manufacturing firms in China's A-share market from 2011 to 2022, a DID model is employed for empirical testing. Robustness checks are conducted using methods such as double/debiased machine learning and IV approaches.

The findings indicate that:

(1) DFA significantly promotes the green development of traditional enterprises.

(2) DFA facilitates firms' green transformation through three pathways: promoting IT, optimizing CU, and enhancing energy efficiency.

(3) The effect of DFA is more pronounced for state-owned enterprises, firms with technology-oriented executives, heavily polluting industries, and regions with strong fiscal support but weaker environmental regulation, demonstrating clear heterogeneity.

(4) While DFA strengthens the green transformation capabilities of local firms, it exhibits complex spatial externalities: it generates a "siphon effect" on geographically adjacent regions but shows no significant spillover to economically similar regions.

In summary, this study not only deepens the theoretical understanding of the relationship between data factors and firms' green transformation but also empirically identifies the causal effect and spatial externality characteristics of DFA. These findings provide strong practical support for promoting the efficient allocation of digital factors and the green upgrading of traditional industries. More broadly, the identified mechanisms through which data factors enable green transformation have general explanatory power for economies at different stages of development. Based on the findings, the following policy recommendations are proposed:

First, improve data infrastructure and promote the rational flow and sharing of data factors across regions. Currently, data factors exhibit complex spatial externalities, generating a "siphon effect" that suppresses the green development potential of neighboring regions. Therefore, it is necessary to accelerate the construction of multi-tier regional data infrastructure, reasonably deploying data centers, computing facilities, and industrial internet platforms in central, western, and less-developed regions, preventing data monopolies and widening digital divides. At the same time, cross-regional data sharing and connectivity should be encouraged through the establishment of standardized and unified data trading platforms to release the spillover effects of data factors and promote coordinated green development across regions. This finding also has implications for developing economies in Asia and Africa with uneven infrastructure development: in promoting digitalization, attention should be paid to the spatial layout and equitable accessibility of data infrastructure to avoid excessive concentration of digital resources in core cities or a few regions, which could hinder overall sustainable development.

Second, strengthen internal corporate capabilities to enhance the conversion of data factors into green productivity. The study shows that technology-oriented executives are more effective in realizing the green benefits of data factors. Therefore, it is important to enhance executives' data literacy and green management

capabilities. Firms should be encouraged to recruit senior managers with digital technology expertise and improve the existing management team's awareness of data-driven green transformation through training and collaborative initiatives. Additionally, firms should establish internal data governance systems to enhance data collection, analysis, and application capabilities, promoting the deep integration of data into energy saving, process optimization, and green innovation. This micro-level mechanism indicates that data factors do not automatically translate into green performance; their effectiveness depends on the alignment of internal governance and human capital structure—a principle that is equally applicable to firms in developing countries where digital foundations are still evolving.

Third, adopt region- and firm-specific policies to stimulate diverse enterprises' motivation for data-driven green transformation. Given that DFA has stronger green effects in state-owned enterprises, regions with high fiscal support, and heavily polluting industries, policies should be tailored according to enterprise type and regional characteristics. On one hand, incentives for non-state-owned enterprises, non-polluting industries, and regions with weak fiscal support should be strengthened—for example, through subsidies for green data technologies, tax incentives, or capacity-building programs—to lower the threshold for green transformation. On the other hand, environmental regulation and data factor governance should be coordinated: regulatory strength should be reinforced while providing supportive digital tools to avoid diminishing marginal returns of data, ensuring synergistic policy effects. From a global sustainable development perspective, this "data factor–industrial policy–environmental regulation" coordination mechanism helps improve resource allocation and energy efficiency, supporting the achievement of the United Nations Sustainable Development Goals related to industrial upgrading (SDG 9), responsible production (SDG 12), and climate action (SDG 13).

Funding

This work is funded by Philosophy and Social Science Research Project of Hubei Provincial Department of Education (Grant no.: 22Q081) and Special Fund Project of Wuhan Textile University (Grant no.: D0401).

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Aghion, P. & Jaravel, X. (2015). Knowledge spillovers, innovation and growth. *Econ. J.*, *125*, 533–573. https://doi.org/10.1111/ecoj.12199.

Agrawal, A., McHale, J., & Oettl, A. (2019). Finding needles in haystacks: Artificial intelligence and recombinant growth. In *The Economics of Artificial Intelligence* (pp. 149–174). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0005.

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Weyl, E. G. (2018). Should we treat data as labor? Moving beyond "free". *AEA Pap. Proc.*, *108*, 38–42. https://doi.org/10.1257/pandp.20181003

Bergemann, D. & Bonatti, A. (2019). Markets for information: An introduction. *Annu. Rev. Econ.*, *11*, 85–107. https://doi.org/10.1146/annurev-economics-080315-015439.

Chen, X., Hou, Y., & Fan, Z. (2025). Club convergence in urban green innovation efficiency: Identification, club features, and the nexus with the digital economy in China. *J. Environ. Manage.*, *391*, 126456. https://doi.org/10.1016/j.jenvman.2025.126456.

Chen, Z., Kahn, M. E., Liu, Y., & Wang, Z. (2018). The consequences of spatially differentiated water pollution regulation in China. *J. Environ. Econ. Manage.*, *88*, 468–485. https://doi.org/10.1016/j.jeem.2018.01.010.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, *21*, C1–C68. https://doi.org/10.1111/ectj.12097.

Diamantoulakis, P. D., Kapinas, V. M., & Karagiannidis, G. K. (2015). Big data analytics for dynamic energy management in smart grids. *Big Data Res.*, *2*, 94–101. https://doi.org/10.1016/j.bdr.2015.03.003.

Fan, L., Wu, W., Yu, D., & Su, T. (2019). Measurement, comparison and dynamic evolution of industrial capacity utilization in China: An empirical study based on firm-level data. *J. Manag. World*, *35*, 84–96. https://doi.org/10.19744/j.cnki.11-1235/f.2019.0107.

Farboodi, M. & Veldkamp, L. (2021). *A model of the data economy*. *NBER Working Paper No. 28427*. https://doi.org/10.3386/w28427.

Goldfarb, A. & Tucker, C. (2019). Digital economics. *J. Econ. Lit.*, *57*, 3–43. https://doi.org/10.1257/jel.20171452
Guo, B., Hu, P., & Lin, J. (2024). The effect of digital infrastructure development on enterprise green transformation. *Int. Rev. Financ. Anal.*, *92*, 103085. https://doi.org/10.1016/j.irfa.2024.103085.
Han, D., Wu, H., & Lu, K. (2024). The effect of data element agglomeration on green innovation vitality in China. *Humanit. Soc. Sci. Commun.*, *11*, 1305. https://doi.org/10.1057/s41599-024-03844-2.
Jiang, T. (2022). Mediating effects and moderating effects in causal inference. *China Ind. Econ.*, *5*, 100–120. https://doi.org/10.19581/j.cnki.ciejournal.2022.05.005.
Jones, C. I. & Tonetti, C. (2020). Nonrivalry and the economics of data. *Am. Econ. Rev.*, *110*, 2819–2858. https://doi.org/10.1257/aer.20191330.
Karami, A. & Igbokwe, C. (2025). The impact of big data characteristics on credit risk assessment. *Int. J. Data Sci. Anal.*, *20*, 4239–4259. https://doi.org/10.1007/s41060-025-00753-8
Kirkley, J., Morrison Paul, C. J., & Squires, D. (2002). Capacity and capacity utilization in common-pool resource industries. *Environ. Resour. Econ.*, *22*, 71–97. https://doi.org/10.1023/A:1015511232039.
Krugman, P. (1991). Increasing returns and economic geography. *J. Polit. Econ.*, *99*, 483–499.
Liang, Z. & Wang, Y. (2022). From “dual misalignment locking” to “dual unblocking”: Scenarios and policy paths for green transformation of traditional Chinese manufacturing. *Soc. Sci. Res.*, *1*, 68–76.
Lynch, C. (2008). How do your data grow? *Nature*, *455*, 28–29. https://doi.org/10.1038/455028a
Qian, J., Zhou, Y., & Hao, Q. (2024). The effect and mechanism of digital economy on green total factor productivity—Empirical evidence from China. *J. Environ. Manage.*, *372*, 123237. https://doi.org/10.1016/j.jenvman.2024.123237.
Qu, X., Lu, P., Wang, H., & Feng, C. (2020). An empirical analysis of the impact of R&D investment on the transformation and upgrading of China's traditional manufacturing industry. *Stat. Decis.*, *36*, 120–123. https://doi.org/10.13546/j.cnki.tjyjc.2020.05.026.
Romer, P. M. (1990). Endogenous technological change. *J. Polit. Econ.*, *98*, S71–S102.
Schaefer, A., Schiess, D., & Wehrli, R. (2014). Long-term growth driven by a sequence of general purpose technologies. *Econ. Model.*, *37*, 23–31. https://doi.org/10.1016/j.econmod.2013.10.014.
Shao, S., Li, X., Cao, J., & Yang, L. (2016). China’s economic policy choices for governing smog pollution based on spatial spillover effects. *Econ. Res. J.*, *51*, 73–88.
Smichowski, B. C., Duch-Brown, N., Hocuk, S., Kumar, P., Martens, B., Mulder, J., & Prüfer, P. (2023). Economies of scope in data aggregation: Evidence from health data. *SSRN Electron. J.*. https://doi.org/10.2139/ssrn.4338447.
Varian, H. (2019). Artificial intelligence, economics, and industrial organization. In *The Economics of Artificial Intelligence: An Agenda.* (pp. 399–422). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0016
Wang, C., Yu, D., & Long, R. (2023). Economic policy uncertainty and enterprise digitalization: Stepping stone or stumbling block? *Bus. Manage. J.*, *45*, 79–100. https://doi.org/10.19616/j.cnki.bmj.2023.06.005.
Wang, H., Hao, Y., & Fu, Q. (2024). Data factor agglomeration and urban green finance: A quasi-natural experiment based on the National Big Data Comprehensive Pilot Zone. *Int. Rev. Financ. Anal.*, *96*, 103732. https://doi.org/10.1016/j.irfa.2024.103732.
Wennberg, K. & Lindqvist, G. (2010). The effect of clusters on the survival and performance of new firms. *Small Bus. Econ.*, *34*, 221–241. https://doi.org/10.1007/s11187-008-9123-0
Yang, X., Hunjra, A. I., Grebinevych, O., Roubaud, D., & Zhao, S. (2025). Roads to sustainable development: Pioneering industrial green transformation through digital economy policy. *J. Environ. Manage.*, *387*, 125721. https://doi.org/10.1016/j.jenvman.2025.125721.
Zhang, N., Kong, F., Choi, Y., & Zhou, P. (2014). The effect of size-control policy on unified energy and carbon efficiency for Chinese fossil fuel power plants. *Energy Policy*, *70*, 193–200. https://doi.org/10.1016/j.enpol.2014.03.031.
Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). A survey on deep learning for big data. *Inf. Fusion*, *42*, 146–157. https://doi.org/10.1016/j.inffus.2017.10.006.
Zhang, R. & Wang, S. (2025). Can the development of the digital economy reduce industrial solid waste pollution? *J. Environ. Manage.*, *386*, 125775. https://doi.org/10.1016/j.jenvman.2025.125775.
Zhou, K., Wang, R., Tao, Y., & Zheng, Y. (2022). Firm green transformation and stock price crash risk. *J. Manag. Sci.*, *35*, 56–69.