Data Aggregation and Enterprise Digital Transformation: A Quasi-Natural Experiment Based on National Big Data Pilot Zones

Wenqian Mu^{a,*}, Jiangyuan Chang^b

School of Economics, Beijing Wuzi University, Beijing, China aWenkymu@163.com, bchangjy_97@163.com *Corresponding author

Abstract: Data as a new type of production factor profoundly influences the development trajectory and competitiveness of enterprises in multiple aspects, including driving innovation, optimizing decision-making, reshaping business models, precision marketing, and improving operational efficiency. This paper uses the establishment of national big data pilot zones as a quasi-natural experiment and employs a difference-in-differences method to study the impact of data aggregation on corporate digital transformation. The findings reveal that data aggregation promotes corporate digital transformation, with more pronounced effects in regions with weaker factor utilization capabilities and in eastern regions. Mechanism analysis indicates that data aggregation enhances corporate willingness to pursue digital transformation by alleviating financing constraints and improving intellectual property protection levels.

Keywords: Data Aggregation; Enterprise Digital Transformation; National Big Data Pilot Zone

1. Introduction

China is currently undergoing a once-in-a-century transformation, with complex and ever-changing domestic and international economic conditions. Domestically, the country has entered a new stage of development, with growth rates adjusting and the transition between old and new growth drivers becoming increasingly evident. Internationally, trade protectionism and unilateralism are on the rise, and anti-globalization trends are gaining momentum, presenting numerous unfavorable factors [1]. The development of the digital economy is the result of the integration of modern information technology with human production and lifestyle. In recent years, developing the digital economy has become a key battleground for countries worldwide to compete for international influence in the information age [2]. Data is a key production factor in the digital economy, and the mining and application of big data can create immense value for the global economy. In fact, after several years of development, the scale of big data elements has grown rapidly, with concentrated effects in specific regions [3]. China issued the "Action Plan for Promoting Big Data Development" in August 2015. To implement the Plan, China began constructing National Big Data Comprehensive Pilot Zones (here in after referred to as pilot zones) starting in 2016, marking a new big data pilot policy introduced by the Chinese government to promote digital economic development. The scope of construction includes the Guizhou Province Big Data Comprehensive Pilot Zone, two cross-regional comprehensive pilot zones, four demonstration comprehensive pilot zones, and one big data infrastructure comprehensive pilot zone. The pilot zone will conduct systematic experiments on seven major tasks: open management and sharing of data resources, integration of data centers, application of data resources, circulation of data elements, aggregation of big data industries, international cooperation on big data, and innovation of big data systems. Through continuous summarization of practical experience that can be referenced, it will ultimately form a ripple effect and demonstration effect.

The Third Plenary Session of the 20th Central Committee of the Communist Party of China pointed out that enterprises should be supported in using digital and intelligent technologies and green technologies to transform traditional industries, thereby enhancing traditional industries. As the main body of social production activities, the digitalization level of enterprises is the core competitiveness of the national digital economy and directly affects the progress of the national digital economy [4]. Existing literature has explored the factors driving enterprise digital transformation from the perspectives of technological development, internal organization, and external environment [5]. For example, Luo et al. found that mixed-ownership reform can significantly promote enterprise digital transformation [6]. Chen

et al. found that digital transformation is influenced by the embeddedness of party organizations [7]. Shu et al. analyzed the physical distance between banks and enterprises and found that enterprises with greater distance from banks exhibit higher levels of digital transformation [4]. Supply chain diversification alleviates the challenges of corporate digital transformation through organizational change effects and technological change effects. However, existing literature has limited research on the impact of data aggregation on corporate digital transformation. Therefore, this paper will build on existing research findings to explore the effects of data aggregation on corporate digital transformation.

The empirical section first uses firm panel data to estimate the impact of data aggregation from the national big data pilot program on corporate digitalization. It then examines the mechanisms through intellectual property protection and financing constraints. The main findings of this paper are as follows: First, the data aggregation triggered by the national big data pilot program significantly accelerates the progress of local corporate digital transformation. Second, the impact of data aggregation on corporate digital transformation exhibits significant heterogeneity: (1) In regions with weaker factor utilization capabilities, the impact of data aggregation on corporate digital transformation is more pronounced than in regions with stronger factor utilization capabilities; (2) Eastern regions are better able to leverage the promotional effects of data aggregation on corporate digital transformation compared to western regions. Third, data aggregation facilitates corporate digital transformation by enhancing intellectual property protection levels and reducing financing constraints.

2. Theoretical Analysis

Within the pilot cities of the National Big Data Pilot Zone, digital resources have shown a significant trend toward concentration. This concentration phenomenon has laid a solid foundation for the digital transformation of local enterprises and provided critical momentum support ^[4]. Following the establishment of the Big Data Pilot Zone, local governments will increase financial and resource investments to advance the construction of digital infrastructure, offering business facilities such as tax incentives and financing guarantees. Additionally, by integrating data resources, a comprehensive database can be established. The big data resources concentrated in this area are deeply embedded in all stages of enterprise production and operations, unleashing immense potential. For example, enterprises can use big data to precisely capture customer needs and enhance marketing effectiveness ^[13]. However, this requires enterprises to possess the capability to unlock the potential and value of big data. Therefore, policy incentives and resource aggregation advantages provide momentum for enterprise digital transformation.

Hypothesis 1: Data element aggregation drives enterprise digital transformation.

The establishment of national big data pilot zones has provided a key platform for the deep integration of finance and big data, attracting financial institutions to cluster in pilot regions [15]. This has created economies of scale, improving the efficiency of capital flows in regional financial markets [12], and effectively alleviating the challenges of corporate financing constraints [11]. On one hand, financial institutions can more accurately assess enterprise risks, innovate and launch tailored products such as "data credit loans"and "supply chain data financing," breaking through the financing bottlenecks caused by information asymmetry under traditional credit models and expanding enterprise financing channels. On the other hand, financial regulation based on big data becomes more intelligent and efficient, enabling timely warnings of financial risks, ensuring the stable operation of financial markets, creating a favorable environment for enterprise financing, and facilitating the smooth flow of funds to enterprises with digital transformation needs. Therefore, the establishment of big data pilot zones alleviates financing constraints for enterprises, thereby promoting their digital transformation.

Hypothesis 2: Data aggregation promotes enterprise digital transformation by alleviating financing constraints.

The establishment of national big data pilot zones has triggered significant data aggregation effects while significantly enhancing intellectual property protection levels [14], thereby opening new pathways for local enterprises' digital transformation. With the development of the digital economy, issues such as difficulties in data rights confirmation, privacy protection, and data transaction mechanisms have emerged [15], leading enterprises to be more cautious when disclosing data and posing challenges to their innovation initiative. National big data pilot zones are actively exploring institutional mechanisms for digital governance, with government, platforms, and other stakeholders collaborating on governance, establishing specialized intellectual property service platforms, and strengthening administrative enforcement efforts to increase penalties for infringement. This protects enterprises' exclusive rights to

intellectual property and the economic benefits of their innovations, creating favorable external conditions for digital element application and open innovation in digitally transforming enterprises. This is conducive to enhancing enterprises' investment in digital R&D funds and R&D human capital. It safeguards the legitimate rights and interests of both parties in technology transactions, forms a developed technology transaction market promoting enterprise digital transformation.

Hypothesis 3: Data aggregation promotes the process of enterprise digital transformation by alleviating corporate financing constraints.

3. Data and empirical model

3.1 Model settings

This paper uses the difference-in-differences (DID) method to examine the impact of data aggregation caused by the establishment of national big data pilot zones on the digital transformation of enterprises. The specific model settings are as follows:

$$DCG_{ijt} = \alpha + \beta bigdata_{j} \times post_{t} + X_{it} + Z_{it} + \rho_{i} + \tau_{t} + \varepsilon_{ijt}$$
 (1)

Where i represents a company, j represents a city, and t represents a year. The dependent variable DCG $_{ijt}$ denotes the digital transformation level of company i located in city j in year t. bigdata $_{j}$ indicates whether city j is a pilot city in the national big data pilot zone; if so, it is 1, otherwise it is 0. Post $_{t}$ is a dummy variable indicating the implementation of big data pilot zone policies, with a value of 0 before 2016 and 1 after 2016. X_{it} and Z_{it} represent control variables at the firm level and city level that may influence a firm's digital transformation. This paper controls for dual roles, equity concentration, debt-to-equity ratio, fixed asset ratio, corporate cash holdings, corporate size, corporate age, net profit margin, economic development level, and financial development level. ρ_{i} represents the firm-level fixed effect, controlling for firm-level characteristics that do not change over time. τ_{t} represents the time-level fixed effect, controlling for national-level characteristics that change over time. ε_{ijt} represents the random disturbance term. β is the core coefficient of interest in this study, reflecting the effect of data aggregation caused by the establishment of big data pilot zones on corporate digital transformation.

The indicators for measuring the level of digital transformation in enterprises were constructed based on the work of Wu et al ^[10]. Annual reports of listed companies were collected, and text analysis was used to count the frequency of relevant keywords in the annual reports and organize the data. Logarithmic processing was then used to obtain the degree of digital transformation in enterprises.

3.2 Data

The descriptive statistics are reported in Table 1, including the means of the full, treatment, and control samples.

Variable	Sample Size	Mean	Standard Error	Min	Max
DCG	12213	1.2786	1.41	0	5.037
bigdata	12213	0.2318	0.422	0	1
dual	12114	0.312	0.4633	0	1
top10 HHI	12208	0.4246	0.1928	0.105	0.9849
tl	12213	0.3835	0.1961	0.0075	0.9976
tang	12209	0.345	0.1572	0.0041	0.9084
cash	12213	0.2095	0.1505	0.0026	0.9342
size	12213	21.9909	1.2686	19.006	28.6365
lnage	12213	1.7467	0.9053	0	3.4012
roe	12213	0.0412	1.7383	-186.557	2.575
gdp	11826	18.019	1.0146	14.2434	19.7598
finance	12195	18.3572	1.3592	14.2521	20.4198

Table 1 Descriptive Statistics

3.2.1 National Big Data Pilot Zone

This article queries and collects the names of pilot regions for national-level big data pilot zones through the Chinese government website, including Beijing, Tianjin, Shanghai, Chongqing, Shenyang,

Hebei, Guizhou, Guangdong, Henan, and Inner Mongolia.

3.2.2 Enterprise and City Data

The data on listed companies used in this paper is sourced from the CSMAR, while the annual report data on relevant companies is sourced from Inspur Network. Data on urban economic and financial development levels is sourced from the China Urban Statistical Yearbook.

4. Empirical Analysis

4.1 Benchmark regression

Table 2 presents the regression results on the impact of the establishment of pilot cities in big data experimental zones on the level of digital transformation of enterprises. Column (2) includes firm-level and year-level fixed effects, Column (3) includes only firm-level and city-level control variables, and Column (4) includes both firm-level and city-level control variables, as well as firm-level and year-level fixed effects. It can be seen that regardless of whether different levels of control variables or fixed effects are included, the coefficients for pilot cities in big data pilot zones are significant at the 1% level. According to the results in column (4), after the establishment of pilot cities in big data pilot zones, the digital transformation level of enterprises in the pilot regions increased by 0.4688, equivalent to 36.67% of the mean level (1.2786), indicating a significant promotional effect.

	DCG			
bigdata×post	(1) 0.8725***	(2) 0.1277***	(3) 0.4688***	(4) 0.4688***
	(0.0316)	(0.0305)	(0.0323)	(0.0323)
Controls	NO	NO	YES	YES
Firm FE	NO	YES	NO	YES
Year FE	NO	YES	NO	YES
Observations	12,228	11,867	11,734	11,734
R-square	0.0682	0.8217	0.2218	0.2218

Table 2 Benchmark Regression

4.2 Stability Test

4.2.1 Parallel trend test

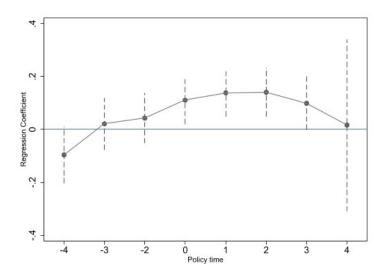


Figure 1 Parallel trend test

The prerequisite for using the double difference method is that the parallel trend assumption must be satisfied, i.e., before the establishment of the national big data pilot zone, the digital transformation of enterprises in the control group and the experimental group should maintain the same trend of change. To test whether the multi-period double difference model set in this paper satisfies the parallel trend test,

following the approach of Beck et al. (2010), the following model is set up:

$$DCG_{it} = \alpha_0 + \sum_{k=-4}^{k=4} \beta_k \, bigdata_j \times post_k + X_{it} + Z_{it} + \rho_i + \tau_t + \varepsilon_{ijt}$$
 (2)

Among these, represents the level of digital transformation of export enterprise i in year t, which is a series of dummy variables indicating the kth year since the city where enterprise i is located became a national-level big data pilot zone. Specifically, when k=0, it indicates the year when China-Europe trade was launched; when k=4, it indicates the four years prior to the establishment of the big data pilot zone; and when k=4, it indicates the fourth year after the establishment of the big data pilot zone. If the city where firm i is located becomes a national-level big data pilot zone in year k, the value is 1; otherwise, it is 0. The year prior to the implementation of the policy is used as the baseline group. Figure 1 reports the coefficient estimates for equation (2) and their changes within the 95% confidence interval. When k<0, all coefficient estimates pass through 0, indicating that they are not statistically significant. The impact of data aggregation on corporate digital transformation did not pass the significance level test prior to the establishment of the big data pilot zone, satisfying the parallel trend test assumption.

4.2.2 Placebo test

To further exclude the influence of other non-random factors on the research conclusions, this paper randomly generated pilot cities for the big data experimental zone and conducted 500 placebo tests based on this, obtaining 500 pseudo-regression coefficients. As shown in Figure 2, the majority of pseudo regression coefficients are distributed around 0 and follow a normal distribution, indicating that the vast majority of regression results are not significant. Furthermore, the estimated coefficients of the benchmark regression are independent of the distribution of these pseudo regression coefficients, suggesting that no other non-random factors influence the test results, thereby passing the placebo test.

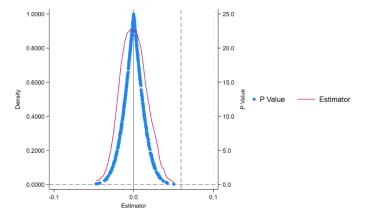


Figure 2 Placebo test

4.2.3 Shorten the sample time span

Samples after 2019 were affected by the COVID-19 pandemic, which caused companies to experience tight cash flow and limited access to resources. These factors reduced companies' willingness to undergo digital transformation and affected the level of digital transformation. Therefore, in order to exclude the impact of the pandemic, this paper excluded samples after 2019. The regression results are shown in column (1) of Table 3 and are significant at the 1% level.

(1)(2) Before 2019 PSM-did bigdata×post 0.1151*** 0.1101*** (0.0325)(0.0314)Controls YES YES Firm FE YES YES Year FE YES YES Observations 11,073 11,356 R-squared 0.8231 0.8252

Table 3 Robustness Test

4.2.4 PSM-DID

To address the endogeneity issues caused by sample selection bias, we adopted a combination of

propensity score matching and difference-in-differences (PSM+DID) to re-regress the model, drawing on the method proposed by. Control variables were treated as covariates, and the Logit model was used to calculate the propensity scores for each sample. A 1:1 bootstrap nearest neighbor matching method was then applied with a bootstrap range of 0.01. The results of the matched data were regressed, with the results shown in Column (2) of Table 3. The coefficient of the core explanatory variable was 0.1101, which was significant at the 1% level, indicating that the baseline regression results are robust.

5. Mechanism Analysis

To explore the mechanism by which data aggregation caused by national big data pilot zones affects the digital transformation of enterprises, this paper draws on the approach focuses on examining the impact of core explanatory variables on mechanism variables. Based on this, it further analyzes the impact of mechanism variables on the explained variables.

$$M_{it} = \alpha + \beta bigdata_i \times post_t + X_{it} + Z_{it} + \rho_i + \tau_t + \varepsilon_{ijt}$$
(3)

Among them, the mechanism variables are replaced by the financing constraint (SA_index) and intellectual property rights protection level (IPR) variables, respectively, while the settings of other variables remain consistent with the benchmark regression model.

5.1 Financing constraints (SA index)

The financing constraints faced by enterprises significantly hinder their digital transformation and upgrading. Digital transformation requires high investment, and high financing costs exacerbate the difficulty of enterprise capital management, thereby dampening enterprises' enthusiasm for digital transformation [8]. This study uses the SA index to measure financing constraints. Since the SA index is negative, a higher SA value indicates that the company faces fewer financing constraints. Table 4 Column (1) presents the regression results between the core explanatory variables and the mechanism variable of financing constraints. The coefficients are positive and significant at the 1% level. This indicates that the establishment of national big data pilot zones promotes digital transformation by reducing financing constraints, thus supporting Hypothesis 1.

	(1)	(2)
	SA_index	IPR
bigdata×post	0.0089***	1,459.9713***
	(0.0021)	(73.5637)
Controls	YES	YES
Firm FE	YES	YES
Year FE	YES	YES
Observations	11,356	11,356
R-square	0.9810	0.8079

Table 4 Mechanism Analysis

5.2 Intellectual property rights protection level (IPR)

The Big Data Pilot Zone is actively exploring governance mechanisms for the digital economy, strengthening enforcement and penalties, and enhancing the intensity of intellectual property protection. Drawing on the research of Wei and Wu ^[9], the G-P indicator is constructed, and an enforcement intensity adjustment is introduced to the G-P indicator. Enforcement intensity is measured from three aspects: the degree of social legalization, government enforcement efficiency, and the rate of patents not being infringed upon. The degree of social legalization is measured by the lawyer-to-population ratio, with a score calculated as the number of lawyers per 10,000 people in the region divided by 5. Government enforcement efficiency is measured by the case closure rate for patent infringement cases. The patent non-infringement rate is calculated as 1 minus the number of patent infringement cases in the province divided by the total number of patents granted in the province that year ^[10]. As shown in Table 4 (Column 2), the coefficient is significant at the 1% level, indicating that the establishment of national big data pilot zones can enhance the region's intellectual property protection level and increase enterprises' willingness to undertake digital transformation.

6. Heterogeneity Analysis

6.1 Level of Factor Utilization

Considering that a company's digital transformation is influenced by its level of access to and utilization of data as a production factor, we draw on the research of Zhang et al. [3] to construct an indicator of a company's factor utilization level, which is divided into two components: technological acquisition capability and technological application capability. Regarding technological acquisition capability, we employ text mining methods to process corporate business registration information data to capture data acquisition capability. Regarding technological application capability, we use the number of digital economy-related invention patents granted at the county level as a proxy indicator for data application capability. Based on the median, enterprises are categorized into those with higher levels of factors utilization level and those with lower levels of factors utilization level. As shown in Table 5, columns (1) to (2), enterprises with lower levels of factors utilization level are more significantly influenced by the establishment of big data pilot zones, thereby more effectively promoting the digital transformation process of enterprises within the region. A possible explanation is that enterprises with weaker data factor utilization, due to their own limited capabilities, found it difficult to undertake digital transformation on their own prior to the establishment of big data pilot zones.

	(1)	(2)	(3)	(4)
	High level of	Low level of	Hu Line -	Hu Line -
	factors utilization	factors utilization	Eastern Region	Western Region
bigdata×post	0.0603	0.0511*	0.1081***	0.6050*
	(0.0507)	(0.0304)	(0.0318)	(0.3226)
Controls	YES	YES	YES	YES
Firm FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Observations	2,019	9,012	11,110	236
R-square	0.8104	0.6438	0.8265	0.6664

Table 5 Heterogeneity Analysis

6.2 Hu Line

The climate, population, economic development, and topography on either side of the Hu Line differ, which may lead to varying degrees of impact on businesses from the policies of the big data pilot zones. In response, this paper categorizes businesses into those in the eastern region and those in the western region based on the Hu Line, and conducts a regression analysis. The results are shown in columns (3) and (4) of Table 5. Enterprises in the eastern region are more susceptible to the influence of pilot policies on digital transformation compared to those in the western region. Possible reasons include higher levels of economic development, more advanced digital infrastructure, and a greater concentration of digital talent in the eastern region. Following the establishment of the big data pilot zone, enterprises in the eastern region face fewer challenges in their digital transformation efforts.

7. Conclusions and Implications

The establishment of national big data pilot zones is a significant measure to foster and strengthen China's digital economy, promote the digital and intelligent transformation of traditional industries, and enhance social governance. Against this backdrop, this paper examines the impact of the establishment of national big data pilot zones on corporate digital transformation and explores two specific mechanisms through which big data pilot zones promote corporate digital transformation: financing constraints and intellectual property protection. The findings are as follows: The pilot program of the National Big Data Pilot Zone has effectively incentivized digital transformation among enterprises, where data aggregation facilitates the process by alleviating financing constraints and enhancing regional intellectual property protection. Notably, the program demonstrates a more pronounced impact on enterprises with lower initial digitalization levels and those located east of the Hu Line.

This study provides a theoretical basis for evaluating the National Big Data Pilot Zone policy and offers insights into how to promote enterprise digital transformation. First, Enhanced governance of data aggregation through state-established incentive funds to stimulate data collection, coupled with rigorous multidimensional quality standards and mandatory self-inspections supplemented by regulatory audits. Concurrently, a national data security framework encompassing cryptographic technologies and access protocols must safeguard data utilization. Second, regionally differentiated policies prioritizing eastern

regions with national R&D centers and innovation hubs to attract advanced resources, while central/western regions receive fiscal transfers to subsidize digital infrastructure costs. Enterprise-specific interventions include foundational data literacy training and subsidized loans for low-capacity firms, contrasted with R&D tax exemptions and approval prioritization for digitally advanced enterprises pursuing AI-big data integration. Third, Institutional mechanism optimization featuring innovative financial instruments and public-private investment funds targeting big data-industrial convergence. Strengthened intellectual property regimes complement nationally unified data market rules and a state-level data exchange platform to facilitate efficient resource allocation and market vitality.

References

- [1] Zhao W, Yang D X. Construction Level of China's Modern Economic System: Statistical Measurement, Spatial Pattern and Correlation Network[J]. Statistics & Decision, 2024, 40(24):95-100.
- [2] Zhou W, Ye L. High-Quality Development of State-Owned Enterprises and New Quality Productive Forces: Reform Directions and Institutional Arrangements[J]. Research on Financial and Economic Issues, 2025, (02):17-28.
- [3] Zhang G S, Yan P, Li X Y. Big Data Factor Agglomeration, Technological Capability Gaps and Regional Disparities in Productivity[J]. China Ind. Econ, 2024, (10):118-136.
- [4] Sun W Z, Mao N, Lan F, et al. Policy empowerment, digital ecosystem and enterprise digital transformation: A quasi natural experiment based on the national big data comprehensive experimental zone[J]. China Ind. Econ, 2023, 9: 117-135.
- [5] Shi B, Guo S Q. Fair Competition Review, New Quality Productive Forces and Digital Transformation of Enterprises[J]. On Economic Problems, 2025, (01):1-14.
- [6] Luo J H, Liu H C, Wu Y L. Does the mixed ownership reform of state owned enterprises promote digital transformation?[J].Finance and Trade Research, 2024, 35(11):79-94.
- [7] Chen D, Liu J Z, Zhao C Y. Party Organization Embedding and Digital Transformation of Private Enterprises[J]. Journal of Shanxi University of Finance and Economics, 2024, 46(11):86-102.
- [8] Zeng X Y, Wei M M, Zhou J. How does the development of the digital economy affect corporate transformation and upgrading? Based on a quasi-natural experiment in a national-level big data comprehensive pilot zone[J]. Fujian Tribune, 2023, (09):74-91.
- [9] Wei H, Wu J. Intellectual Property Rights, Import and Innovation of Leading Innovative Enterprises[J]. Journal of Financial Research, 2018, (09):91-106.
- [10] Wu C P, Di T. Intellectual property rights enforcement, corporate innovation and operating performance: Evidence from China's listed companies[J]. Economic Research Journal, 2016, 51(11):125-139.
- [11] Chen D, Ren X G, Xie X J. Development of Big Data, Innovation Ecology and Quality of Enterprise Technological Innovation-Quasi Natural Experiments Based on National Big Data Comprehensive Experimental Zones[J]. Forum on Science and Technology in China, 2024, (09):79-89.
- [12] Liu J B, Liu Q M. Research on the Interaction and Spatial Spillover Effects of Financial Agglomeration and Economic Growth[J]. Commercial Research, 2021, (06):73-80.
- [13] Shi D, Zheng Y. Empowerment Mechanism of Data Elements and Improvement of Total Factor Productivity of Enterprises: Evidence from the National Big Data Comprehensive Pilot Zone[J]. Reform, 2024, (11):1-16.
- [14] Chen H B, Zhang X X. How can National Comprehensive Big Data Pilot Zone promote green transformation of cities: Based on the dual perspective of the growth rate of carbon emissions and carbon productivity[J]. China Soft Science, 2024, (11):172-188.
- [15] Li S X, Huang Z. Digital Economy and High quality Development: Mechanisms and Evidence[J]. China Economic Quarterly,2022,22(05):1699-1716.