

Digital economy and entrepreneurial vitality—Causal inference based on double machine learning

Niyuan Yin

Business School, Soochow University, Jiangsu, China

ynyuan67@163.com

Abstract. While digital economy provides entrepreneurs with convenient data elements, it also narrates their cognition due to its redundant information and data deluge. Based on the contradiction between the dilemma of “information cocoon” and the goal of “wisdom sharing” in the context of digital economy, this paper selects 248 prefecture-level cities in China from 2010 to 2019 as research samples and takes the policy of National Big Data Comprehensive Pilot Zone as the treatment variable, empirically analyzing the impact of digital economy on entrepreneurial vitality by the double machine learning model. The results indicate that: (1) Digital economy has a significant positive impact on entrepreneurial vitality; (2) Compared with the western region, the impact of digital economy on entrepreneurial vitality is more significant in the eastern and central regions, and there is heterogeneity among cities at different levels; (3) Optimizing the business credit environment and coordinating the digital infrastructure are two mechanisms between them.

Keywords: Digital economy, Entrepreneurial Vitality, National Big Data Comprehensive Pilot Zone, Double Machine Learning.

1. Introduction

First proposed by Tapscott in 1996, the connotation of the term “digital economy” has been further expanded with the deeper insights of scholars in recent years [1]. The G20 Digital Economy Development and Cooperation Initiative defines digital economy as a series of economic activities in which digital information and knowledge become production factors, modern information networks become an important space for economic activities, and information technology becomes an important driving force for optimizing economic structure and promoting economic growth. At present, digital economy is a powerful engine driving global development. According to the Global Digital Economy released by the China Academy of Information and Communications Technology at the 2023 Global Digital Economy Conference, digital economy in the world’s five major countries continues to accelerate, including the United States, China, Germany, Japan, and South Korea, accounting for 58% of GDP. Compared with 2016, it has increased by about 11 percentage points [2], indicating a worldwide trend of data value, digital industrialization, industrial digitalization, and digital governance. In 2018, China released the Outline of Digital Economy Development Strategy, the first national digital economy overall strategy, and issued the 14th Five-Year Digital Economy Development Plan in 2022, which proposes to optimize the digital infrastructure and give full play to data elements, aiming to seize the opportunities of digital development and expand new space for economic development.

As an important part of China's digital economy policy system, the construction of the National Big Data Pilot Zone aims to explore new mechanisms for the development of big data, including the open sharing of public data, big data industry aggregation, and other aspects, promoting the efficient use of data resources and industrial innovation to realize the deep integration of big data and economy. In August 2015, the State Council of China issued the Outline of Action to promote the development of big data, proposing to carry out regional pilot projects for state-level big data pilot zones. Guizhou became the first pilot region and officially launched the construction of the first pilot zone in September of the same year. In 2016, the National Development and Reform Commission, the Ministry of Industry and Information Technology, and the Cyberspace Administration of the CPC Central Committee issued a letter to approve the construction list of the second batch of the National Big Data Comprehensive Pilot zone, including Beijing, Tianjin, Hebei, Inner Mongolia, (Liaoning) Shenyang, Henan, Shanghai, Chongqing, and Guangdong. As an exogenous policy, the national big data comprehensive pilot zone is widely used in the research of digital economy policy [3][4], which can significantly promote the activity of urban entrepreneurship [5][6]. Therefore, it is reasonable to take the policy as the treatment variable of digital economy from both theoretical and practical levels.

The rise of digital infrastructure has deeply integrated digitalization and entrepreneurship [7], and the subsequent transformation of business models based on digital technology has triggered a new round of digital entrepreneurship ecology [8]. The value of data and information has become increasingly prominent, being a key resource for enterprises to obtain competitive advantages in decision-making and operation. However, an influx of too much data can also backfire. In the current data deluge, individuals and enterprises in the digital era are surrounded by a large number of redundant and gradually homogenized data, as if they are in an "information cocoon" situation, falling into a passive state of information reception, which leads to narrow cognition and solidification of thinking, far from the previously envisaged "wisdom sharing". Enterprises rely too much on analytical tools but ignore their own innovative thinking and judgment ability, and focus too much on the results of data analysis but miss the fleeting market opportunities, thus stifling the burst of entrepreneurial vitality. Under the phenomenon of "data dictatorship", data is no longer an auxiliary tool to serve people but gradually evolves into the bondage to enslave human beings [9]. The two sides of the value and risk of big data also make the research on the impact of digital economy on entrepreneurial vitality have certain practical significance. Therefore, this paper takes 248 cities in China from 2010 to 2019 as research samples and takes the National Big Data Comprehensive Pilot Zone as the treatment variable to empirically test the impact of digital economy on entrepreneurial vitality, to facilitate the high-quality development of China's digital economy and the efficient circulation of data elements.

The marginal contributions of this study are: (1) on the theoretical level, the double machine learning model is adopted to make full use of its advantages in dealing with high-dimensional control variables and non-parametric prediction, avoiding the dimensional curse and setting bias of traditional multiple linear regression and difference-in-difference models; (2) on the practical level, based on the background of digital economy, this paper can explore the growth effect of urban entrepreneurial vitality and the potential mechanism from the perspective of optimizing the business credit environment and coordinating digital infrastructure in the digital era, providing certain ideas for facilitating the flow of data elements and enhancing the policy formulation of entrepreneurial vitality.

2. Theoretical Analysis and Hypothesis

Based on the theory of information asymmetry, each person has different information in the transaction. As a shortcoming of the traditional financial system [10], the asymmetric information between banks and enterprises in the supply and demand of funds gives rise to the phenomenon of the "information isolated island", resulting in slow and expensive financing for small and micro enterprises which have made great contributions to tax creation, employment supply, and market vitality [11], often acting as the "long tail" that is easily ignored in the financial market. Internet finance based on digital technologies such as big data and cloud computing has become an effective solution to break the barriers of information asymmetry and reduce transaction and opportunity costs [12]. In terms of inter-enterprise

cooperation, business credit is the guide for entrepreneurs to find partners in complex business networks [13]. When every individual in society has a high credit, opportunistic entrepreneurship relying on entrepreneurial information and material resources becomes the mainstream [14]. Through the traceable credit files and historical transactions on the Internet platform, more accurate credit evaluation can be formed between banks and enterprises and between enterprises with multidimensional analysis such as market behavior and consumer feedback benefiting from big data analysis, artificial intelligence, and other technologies, which can act on the prior screening and punishment of dishonesty in the social credit system [15], alleviating the information asymmetry faced by entrepreneurs [16] to provide diversified financing channels for enterprises. This optimization of the business credit environment can reduce the capital threshold of entrepreneurs, facilitate financing activities during the initial stage of entrepreneurship, and promote the overall entrepreneurial vitality of the city.

Digital infrastructure can be divided into connection-oriented telecommunications infrastructure such as broadband, and new infrastructure that integrates new generation information technologies such as artificial intelligence and the Internet of Things, and provides comprehensive digital services [17]. On one hand, the traditional interconnected telecommunications infrastructure can be instrumental in the high-speed flow of information and data beyond geographical limitations, so that start-ups can significantly reduce the cost of information acquisition, and connect with local, national, and even global markets. On the other hand, through the new digital infrastructure, startups can use digital technologies such as big data and cloud computing to dig deep into customer's pain points and future needs, identify market signals, and seize entrepreneurial opportunities, thus expanding the entrepreneurial market [10] and enhancing the entrepreneurial vitality of cities. Taking Guizhou Province, the first batch of pilots in the Big Data Pilot Zone, as an example, it actively implemented the strategy, which was also embodied in digital infrastructure [18]. In terms of telecommunication infrastructure, through the establishment of Guiyang and Gui'an national Internet direct link points, Guizhou's outbound Internet bandwidth capacity has reached 5690Gbps, and the Internet user information index ranks first in the country. In terms of new infrastructure, Guizhou is committed to building a "Cloud Guizhou" platform to open government data in medical, social credit, and other fields to the public, to achieve data collection, overall planning, and sharing.

Based on the above theoretical analysis, the following hypotheses are made:

H1: Digital economy can boost entrepreneurial vitality in cities.

H2: Digital economy promotes entrepreneurial vitality by optimizing the business credit environment.

H3: Digital economy promotes entrepreneurial vitality by coordinating the digital infrastructure.

3. Research Design

3.1. Model Specification

The double machine learning model (DML) was first proposed by Chernozhukov et al. [19], which has unique advantages in variable selection and model specification in the field of causal inference [20]. Compared with traditional multiple linear regression, it can make up for their limitation. Firstly, to maintain the estimation accuracy under high-dimensional control variables, double machine learning is based on several machine learning and regularization algorithms, which can automatically filter out the set of control variables with high prediction accuracy, thus alleviating the model estimation bias problem caused by the curse of multiple covariances under high-dimensional control variables. Secondly, it can effectively avoid the problem of model misspecification [21]. Due to the normalization of non-linear relationships between variables in economic transformation [22], the linear setting of traditional models is prone to estimation bias, leading to the instability of the estimator results, such as double-in-differences model, breakpoint regression which rely on a strict system of assumptions [23]. Whereas the double machine learning model relies on machine learning algorithms in the treatment of the non-linear problem, and addresses the regularization bias of machine learning models through the instrumental variables method. Therefore, this paper refers to Chernozhukov et al. [19] to construct the following

model to empirically study the policy effect of the Quasi-natural experiment of the National Big Data Comprehensive Pilot Zone.

$$Entre_{it} = \theta_0 Event_{it} + g(X_{it}) + U_{it}, E(U_{it}|X_{it}, Event_{it}) = 0 \quad (1)$$

Where i denotes city and t denotes year; $Entre_{it}$ denotes the explanatory variable entrepreneurial vitality, which indicates the entrepreneurial vitality of city i in year t ; $Event_{it}$ denotes the dummy variable for the Big Data Pilot Zone policy, and θ_0 denotes its estimated coefficient; X_{it} denotes the set of multidimensional control variables; the specific form of the function g is unknown, and it needs to be estimated by the machine learning algorithm; U_{it} denotes the error term, which has a conditional mean value of zero. Considering the small-sample traits of data [25], the following auxiliary regression is constructed to ensure unbiased estimation result:

$$Event_{it} = m(X_{it}) + V_{it}, E(V_{it}|X_{it}) = 0 \quad (2)$$

Where $m(X_{it})$ denotes the regression function of the treatment variable on the control variables, which needs to be estimated by the machine learning algorithm; V_{it} denotes its error term, whose conditional mean is 0.

The specific steps for parameter estimation are as follows:

Firstly, divide samples into k subsamples $I_1, \dots, I_k$, whose complementary sets are $I_1^c, \dots, I_k^c$.

Secondly, use I_k^c to estimate $\hat{m}(X_{it})$, the estimator of $m(X_{it})$ in Eq.(2) for each individual in I_k , with the machine learning algorithm to get the residuals $\hat{V}_{it} = Event_{it} - m(X_{it})$. Then, the same machine learning algorithm is used to estimate $\hat{g}(X_{it})$ in Eq.(1).

Thirdly, $\hat{V}_{it}$ is used as the instrumental variable in the regression of $Entre_{it}$ to $Event_{it}$ based on I_k to partialling the effect of X_{it} out from $Event_{it}$, and get the unbiased estimation as follows:

$$\hat{\theta}_0 = \left(\frac{1}{n} \sum_{i \in I, t \in T} \hat{V}_{it} Event_{it} \right)^{-1} \frac{1}{n} \sum_{i \in I, t \in T} \hat{V}_{it} (Entre_{it} - \hat{g}(X_{it})) \quad (3)$$

If plug $Entre_{it}$ into Eq.(3) and multiply by $\sqrt{n}$, the following equation can be obtained:

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = a^* + b^* + c^* \quad (4)$$

$$a^* = [E(V^2)]^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I, t \in T} V_{it} U_{it} \quad (5)$$

$$b^* = [E(V^2)]^{-1} \frac{1}{\sqrt{n}} [m(X_{it}) - \hat{m}(X_{it})][g(X_{it}) - \hat{g}(X_{it})] \quad (6)$$

Where the leading term on the right-hand side will satisfy a normal distribution with a mean of 0, and the second term has a product of two error terms. Let φ_m and φ_g be the convergence rates of $\hat{m}(X_{it})$ to $m(X_{it})$ and $\hat{g}(X_{it})$ to $g(X_{it})$, respectively. Hence, the convergence rate of the whole term is $\sqrt{nn}^{-(\varphi_m + \varphi_g)}$, which can approach 0, and the last term is eliminated in the process of sample splitting. Thus the unbiased estimation coefficient θ_0 in I_k is obtained.

Repeat the steps in other subsamples, and finally, we can take the average of these estimation coefficients to get the estimate $\hat{\theta}_0 = \frac{1}{k}(\hat{\theta}_0^{(1)} + \dots + \hat{\theta}_0^{(k)})$.

3.2. Variable Selection

3.2.1. Explained Variable

Entrepreneurial Vitality (Entre). There are two main indicators to measure it, one of which is the natural logarithm of the number of new enterprises in a city, and the other is the number of innovative enterprises

per 100 people in a city. Referring to Xie et al. [24] and Lin et al. [25], this paper selects the natural logarithm of the number of newly registered enterprises in a city that year as the explained variable for the benchmark regression and adopts the latter for the robustness test.

3.2.2. Treatment Variable

Dummy variable for the National Big Data Comprehensive Pilot Zone policy (Event). The list of policy pilot cities is taken from the official website of the Ministry of Industry and Information Technology of China. If the city was set up as a big data pilot zone and thereafter, it takes the value of 1 for the year, and 0 otherwise. Although Guizhou region was set up as a pilot city in the same year as the other batches, it started the construction of the pilot zone in September 2015. Therefore, the node of policy implementation is set as 2015 for it and 2016 for the other pilot cities.

3.2.3. Mechanism Variable

On one hand, based on the data released in 2010, 2011, 2012, 2015, 2017, and 2019 in China's Urban Business Credit Environment Index (CEI) to construct the mechanism variable Business Credit environment (Credit), the remaining data were filled in by interpolation method. On the other hand, the following digital infrastructure measurement index system is established to comprehensively investigate the digital infrastructure input and output of cities referring to the research of Wang et al. [26], and the entropy weight method is used to calculate the weight of each index to obtain the mechanism variable Digital Infrastructure (Digfra).

Table 1. Digital infrastructure measurement indicator system.

First level indicators	Second level indicators	Indicator calculation	Attribute
Digital infrastructure investment	Optical cable density	Length of long-distance cable line/administrative area	Positive
	Per capita Internet broadband access port	Internet broadband access ports/total population	Positive
	Relevant practitioners	Proportion of employees in information transmission, computer services, and software industries in urban units	Positive
Digital infrastructure output	Telecommunications revenue	Total telecommunications revenue /total population	Positive
	Mobile phone penetration rate	Mobile phone users/total population	Positive
	Internet penetration rate	Internet broadband access users /total population	Positive

Since the selected indicators are all positive, they are first standardized:

$$z_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \quad (7)$$

Secondly, the weight of the sample i under the indicator j is calculated:

$$p_{ij} = z_{ij} / \sum_{i=1}^n z_{ij} \quad (8)$$

Then, the information entropy is calculated for each metric:

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln(p_{ij}) (j = 1, 2, \dots, 6) \quad (9)$$

Thus, the information utility value of each indicator is obtained:

$$d_j = 1 - e_j \quad (10)$$

Finally, the information utility value is normalized to obtain the entropy weight of each indicator:

$$W_j = d_j / \sum_{j=1}^6 d_j (j = 1, 2, \dots, 6) \quad (11)$$

Therefore, we can get the following weight of each indicator:

Table 2. Weight of each indicator based on the entropy weight method.

Indicator	Optical cable density	Per capita Internet broadband access port	Relevant practitioners	Telecommunications revenue	Mobile phone penetration rate	Internet penetration rate
Weight	0.287	0.265	0.062	0.191	0.076	0.119

3.2.4. Control Variables

To take into the potential impact of urban characteristics on entrepreneurial vitality full consideration, with reference to the practices of Zhao et al. [27] and Zhi et al. [5], the control variables selected in this paper are as follows: level of economic development (lnpgdp), level of financial development (lnfin), degree of fiscal decentralization (Finadp), industrial structure (Struc), human capital (Hum), level of foreign investment (FDI). To improve the fitting accuracy of the model, quadratic terms of control variables are added, and the fixed effects of city and time are introduced in the form of dummy variables. The symbols and definitions of all variables are shown in Table 3.

Table 3. Definition of the variables.

Types	Variables	Symbols	Definition
Explained variable	Entrepreneurial vitality	Entre	Natural logarithm of the number of newly registered enterprises that year
Treatment variable	National Big Data Comprehensive Pilot Zone	Event	1 for the year since the city is set as the National Big Data Comprehensive Pilot Zone, 0 otherwise
Mechanism Variables	Business credit environment	Credit	Data released by China's Urban Business Credit Environment Index after linear interpolation
	Digital infrastructure	Digfra	An index system based on relative research with entropy weight method as the measure for index weight

Table 3. (continued).

Control Variables	Economic development level	lnpgdp	Natural logarithm of per capita GDP
	Financial development level	lnfin	Natural logarithm of outstanding loans of financial institutions to GDP
	Fiscal decentralization degree	Finadp	Budget revenue/budget expenditure
	Industrial structure	Struc	Value added of tertiary industry to value added of secondary industry
	Human capital	Hum	Education expenditure/GDP
	Foreign capital	FDI	Actual use of foreign capital/GDP

3.3. Data processing

In this paper, 248 prefecture-level cities from 2011 to 2019 are selected as the research object, and all continuous variables were winsorized at 1% and 99%. The number of new enterprises in the city in the current year is derived from the registered data of Chinese industrial and commercial enterprises, and the control variables such as the level of economic development and financial development are derived from the China City Statistical Yearbook of the corresponding year. The missing data were filled by linear interpolation method and the data processing software was Stata17.

4. Results and Discussion

4.1. Analysis of Main Effects

Based on the double machine learning model, this paper conducts the regression of main effects, estimating the policy effect of the National Big Data Comprehensive Pilot Zone on entrepreneurial vitality by setting the sample segmentation ratio as 1:4 with the random forest algorithm as the estimation method to predict and solve the main regression and auxiliary regression. The regression results are shown in Table 4. Among them, column (1) only controls the primary term of the control variable, column (2) further controls the city and time fixed effect based on column (1), column (3) adds the quadratic term of the control variable based on column (1), and column (4) also adds the quadratic term of the control variable with the fixed effect of city and time to control the influence of individual differences in the panel data. The coefficients before the policy variable are all significantly positive at the level of 1%, indicating that the establishment of the National Big Data Comprehensive Pilot Zone can significantly promote urban entrepreneurial vitality. Therefore, H1 is verified.

Table 4. Results of main effects.

	(1)	(2)	(3)	(4)
	Entre	Entre	Entre	Entre
Event	0.346*** (0.043)	0.257*** (0.035)	0.362*** (0.044)	0.253*** (0.035)
Linear terms of control variables	Yes	Yes	Yes	Yes
Quadratic terms of control variables	No	No	Yes	Yes
City	No	Yes	No	Yes

Table 4. (continued).

Year	No	Yes	No	Yes
N	2480	2480	2480	2480

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

4.2. Robustness Tests

4.2.1. Change of the explained variable

Referring to the practice of Bai et al. [28], the number of innovative enterprises per 100 people (Entre') was used to replace the original explained variable. The regression results are shown in column (1) of Table 5. Although the regression coefficient before the treatment variable decreases, the symbol and level of significance remain unchanged, indicating the robustness of the regression results.

4.2.2. Exclusion of samples of municipalities

As municipalities have an important position in China's politics and economy and generally have higher economic strength, resource allocation ability, and policy support, they can attract and concentrate more innovative enterprises and talents, leading to a certain gap between them and other cities in terms of entrepreneurial vitality. Therefore, after excluding the samples of four municipalities of Beijing, Tianjin, Shanghai, and Chongqing, this paper conducts regression analysis again to test the robustness of the promotion effect of the establishment of big data pilot zones on entrepreneurial vitality. The regression results are shown in column (2) of Table 5. The regression coefficient before the treatment variable is still significantly positive at the level of 1%. Therefore, it can be concluded that the policies of the big data pilot zones can still significantly boost the growth of entrepreneurial vitality in the city.

4.2.3. Control of province-time interactive fixed effects

Since cities in the same province may have similarities in history, culture, policy, and other aspects [25][32], this paper further controls the province-time interactive fixed effects based on the fixed effects of city and time, to control the differences among different provinces under the change of time. The regression results are shown in column (3) of Table 5. The symbol and significance level of the treatment variable still remain the same, which again verifies the robustness of the regression results.

4.2.4. Treatment of Endogeneity Problems

The main sources of endogenous problems in this paper are as follows. On one hand, in the selection of pilot cities in the big data pilot zones, factors such as the economic development and infrastructure of each city are also taken into consideration, so the selection of pilot cities is non-random [29]. On the other hand, due to the missing data from some cities' statistical yearbooks, the problem of missing variables is unavoidable under this limitation. To solve the above endogenous problem, this paper builds the following partial linear instrumental variable model of double machine learning with reference to the study of Chernozhukov et al. [19]:

$$Entre_{it} = \theta_0 Event_{it} + g(X_{it}) + U_{it} \quad (12)$$

$$IV_{it} = m(X_{it}) + V_{it} \quad (13)$$

Where IV_{it} is the instrumental variable for $Event_{it}$. Based on the practice of Liu et al. and Huang et al., urban relief amplitude is selected as the instrumental variable for the National Big Data Comprehensive Pilot Zone for the following reasons. On one hand, as a natural geographical feature of the city, the relief amplitude of the city is exogenous, which has little impact on the entrepreneurial vitality, thus satisfying the exogenous hypothesis of the instrumental variable. On the other hand, it is related to the cost of infrastructure construction and the operation efficiency of big data platforms. The

smaller the relief amplitude of a city, the lower the cost of big data infrastructure construction, and the higher the operation efficiency of platform equipment, hence the more likely it is to become a pilot city. From this dimension, it meets the correlation hypothesis of instrumental variables. Considering that relief amplitude is cross-sectional, a variable that changes with time is introduced referring to Nunn et al. [30]. Therefore, the interaction term between the time dummy variable and urban relief amplitude is used as the instrumental variable. The regression results are shown in column (4) of Table 5. The former regression coefficient is significantly positive at the 1% level, indicating that after handling endogenous problems, the Big Data Pilot Zone policy still has a significant positive impact on entrepreneurial vitality, which can demonstrate the robustness of the conclusions.

Table 5. Results of robustness tests.

	Change of the explained variable	Control of province- time interactive fixed effects	Exclusion of samples of municipalities	Instrumental Variable
	(1)	(2)	(3)	(4)
Event	Entre' 0.134*** (0.047)	Entre 0.371*** (0.128)	Entre 0.275*** (0.039)	Entre 3.705*** (1.277)
Linear and quadratic terms of control variables	Yes	Yes	Yes	Yes
province-time interactive fixed effects	No	Yes	No	No
City	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes
N	2480	2480	2440	2480

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

4.2.5. Test of Exogenous Impact

Since the growth of entrepreneurial vitality may be affected by the superposition of other policies, this paper controls the corresponding policies from 2010 to 2019 to eliminate the interference of parallel policies and verify the robustness of regression results. In terms of digital economy, there are two main policies “Broadband China” (Broadband) and “Smart City” (Smartcity). The above two policies are important measures to promote the development of digitalization and information technology in China. The former focuses on improving network infrastructure, while the latter mainly focuses on improving urban management and service levels through information technology, which also plays a certain driving role in the disclosure and circulation of information and data in modern society. Therefore, this paper constructs policy dummy variables of “Broadband China” and “Smart city”, and adds them to the regression analysis with reference to the practice of Sun et al. [31]. The regression results are shown in Table 6. After excluding the exogenous impact of parallel policies, the policy dummy variable of the National Big Data Comprehensive Pilot Zone is still significant at the level of 1%, indicating that the establishment of the pilot zones still has a steady influence on the entrepreneurial vitality of cities.

Table 6. Results of excluding the exogenous impact of parallel policies.

	(1)	(2)	(3)
Event	Entre 0.259*** (0.035)	Entre 0.250*** (0.035)	Entre 0.257*** (0.034)
Broadband	Yes	No	Yes
Smartcity	No	Yes	Yes
Linear and quadratic terms of control variables	Yes	Yes	Yes
City	Yes	Yes	Yes
Year	Yes	Yes	Yes
N	2480	2480	2480

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

4.2.6. Reset of Double Machine Learning Models

To correct the setting bias of the double machine learning model, this paper adopts two ways to reset the model. One is to change the sample segmentation ratio from 1:4 to 1:2 and 1:7, and the other is to reset the machine learning algorithm by replacing the random forest algorithm with lasso, gradboost, and ridge algorithms. The regression results are shown in Table 7. The coefficients before the policy dummy variable are all significantly positive at the level of 1%, indicating that the establishment of the National Big Data Comprehensive Pilot Zone can significantly promote the entrepreneurial vitality of cities, which confirms the robustness of the result.

Table 7. Results of reset of double machine learning models.

	Kfolds=3 (1)	Kfolds=8 (2)	Lassocv (3)	Gradboost (4)	Rigdecv (5)
Event	Entre 0.261*** (0.037)	Entre 0.257*** (0.037)	Entre 0.081*** (0.018)	Entre 0.297*** (0.038)	Entre 0.065*** (0.017)
Linear and quadratic terms of control variables	Yes	Yes	Yes	Yes	Yes
City	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes
N	2480	2480	2480	2480	2480

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

4.3. Heterogeneity Analysis

4.3.1. Heterogeneity Analysis Based on Different Regions

The regional development focus and policy orientation of cities in different regions vary a lot, therefore, enterprises were divided into East, Middle, and West for heterogeneity analysis. According to the regression results in columns (1), (2), and (3) of Table 8, the influence of the National Big Data Comprehensive Pilot Zone treatment variable on entrepreneurial vitality is significantly positive for eastern and central cities. But for western cities, although the regression coefficient is positive, it is not significant. A possible reason may be the construction of urban agglomerations such as the Beijing-Tianjin-Hebei region and Pearl River Delta in the eastern and central cities, which can achieve effective regional coordinated development and build trans-regional innovation cooperation networks, thus significantly promoting regional innovation performance [32]. Due to the correlation between

geographical location and economic development, the construction of the Big Data Pilot Zone can promote technology exchange, resource sharing, and industrial collaboration in the eastern and central regions, leading to true wisdom sharing, and thus enhancing the entrepreneurial vitality of cities in the region. However, the remote geographical location and harsh natural environment pose hurdles for western cities in the interconnection with the eastern and central cities, and even within the western region. Therefore, it is difficult to build a cross-regional innovation cooperation network in the western region, which affects the circulation of entrepreneurial factors, stifling entrepreneurial vitality.

4.3.2. Heterogeneity Analysis Based on Different City Grades

Cities at different levels have different economic bases and resource endowments, which can reflect the city's development status and residents' living standards. Although a city is set as the pilot zone, the improvement of entrepreneurial vitality is still affected by its factors. Based on the practice of Zhang et al., this paper determines first-tier, second-tier, third-tier, fourth-tier, and fifth-tier cities based on the "2019 City Business Charm Ranking", and classifies the first two as developed cities, the third-tier cities as more developed cities, and the latter two as less developed cities. As can be seen from the regression results of columns (4), (5), and (6) in Table 8, the policy has a significant promotion on all cities at the 1% level, but from the regression coefficient, policies have the most obvious promotion effect on the entrepreneurial vitality of less developed cities, followed by developed cities and more developed cities.

The economic development level of less developed cities is generally low with a single structure, mainly relying on traditional agriculture or labor-intensive industries with backward infrastructure, brain drain, and lack of innovation ability. However, the Big Data Pilot Zone policy plays a greater role in promoting their entrepreneurial vitality. The possible reasons may be as follows: On one hand, based on the late-developing theory, the construction of the Big Data Pilot Zone in less-developed cities brings institutional and technical late-developing benefits to them. In terms of traditional industries and infrastructure, less developed cities generally lag behind developed cities and more developed cities. But the policy is tantamount to an opportunity for leapfrog development. By absorbing the digital development experience of developed cities and shortening the time for digital technology research and development, the less developed can narrow the digital divide and drive economic growth. On the other hand, the Big Data Pilot Zone brings preferential policy and resource support to them, for which policy support such as capital investment, tax incentives and talent introduction is particularly important, especially in the context of accelerating the implementation of innovation-driven development strategy, less developed cities can make use of these advantages to continuously attract innovative enterprises and talents, promote local entrepreneurial activities, and then stimulate high-quality economic development.

Table 8. Results of heterogeneity analysis.

	East (1)	Middle (2)	West (3)	Developed (4)	More developed (5)	Less developed (6)
Event	Entre 0.325*** (0.053)	Entre 0.252*** (0.075)	Entre 0.018 (0.082)	Entre 0.253*** (0.066)	Entre 0.209*** (0.056)	Entre 0.294*** (0.053)
Linear and quadratic terms of control variables	Yes	Yes	Yes	Yes	Yes	Yes
City	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes
N	960	830	690	470	640	1370

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

4.4. Mechanism Analysis

This paper constructs two paths of optimizing the business credit environment and coordinating the digital infrastructure referring to the method of Farbmacher et al. [33] to analyze the causal mediating effect in double machine learning. The mechanism analysis results are shown in Table 9. Under different mechanisms, the total effect of digital economy on entrepreneurial vitality is significantly positive at the level of 1%, and the conclusion of benchmark regression still holds.

4.4.1. Optimize the business credit environment

This paper constructs a mechanism variable business credit environment (Credit) to empirically test the mechanism. It can be seen from the first row of Table 9 that the indirect effects of both the treatment group and the control group are significantly positive. After stripping the path, the direct effects of the treatment group and the control group are also positive and significant at the 1% level. Therefore, H2 is verified that the mechanism of optimizing the business credit environment is established. Personal and corporate credit data are absorbed and integrated by digital tools to form comprehensive credit files, thus breaking data resource barriers, and reducing information asymmetry and transaction costs, which provides entrepreneurs with a more fair and transparent business environment, and access to investment loans. Through credit evaluation, entrepreneurs can more easily obtain the support of new financing channels such as Internet finance, reduce the threshold of entrepreneurship, strengthen the confidence and motivation of entrepreneurs, and finally promote the entrepreneurial vitality of cities.

4.4.2. Coordinate the digital infrastructure

This paper constructs the mechanism variable Digital Infrastructure (Digfra) for regression analysis to test the mechanism. It can be seen from the second row of Table 9 that the indirect effects of the disposal group and the control group are significantly positive at different levels, indicating that digital economy can promote entrepreneurial vitality by coordinating infrastructure construction. Therefore, H3 is verified. In the digital era, digital economy provides entrepreneurs with powerful data storage and analysis capabilities through the overall construction of infrastructure such as broadband networks and data centers, which promotes the circulation of data elements, improves the availability of data, and helps entrepreneurs quickly build and expand their businesses, reducing the threshold of entrepreneurship and improving entrepreneurial efficiency. Driven by the infrastructure for such digital technologies, cities can attract startups and venture capital institutions to jointly establish a digital economy mass innovation space, provide entrepreneurs with capital, technology, and market support, and then form an ecosystem of wisdom-sharing within the city or even regional cooperation network, finally promoting industrial digital transformation and high-quality economic development.

Table 9. Results of mechanism analysis.

Mechanism Variables	Total effect	Direct effect of the treatment group	Direct effect of the control group	Indirect effect of the treatment group	Indirect effect of the control group
Credit	0.630***	0.604***	0.575***	0.055***	0.025***
Digfra	0.587***	0.563***	0.563***	0.024*	0.024**

Note:***、**、* represent that is significant at the 1%、5%、10% level respectively.

5. Conclusions and Policy Implications

Based on the contradiction between the dilemma of “information cocoon” and the goal of “wisdom sharing” under the background of digital economy, this paper selects 248 cities from 2010 to 2019 as research samples, and takes the policy of National Big Data Comprehensive Pilot Zone as the treatment variable of digital economy with the double machine learning model, empirically testing the effect and mechanism between digital economy and entrepreneurial vitality. The results indicate that: firstly, digital economy has a significant positive impact on entrepreneurial vitality, which remains robust after the test of endogenous treatment, exogenous impact test, and resetting of double machine learning model;

secondly, according to heterogeneity analysis, the impact of digital economy on entrepreneurial vitality is more obvious in eastern and central regions with developed regional innovation cooperation networks, and the impact on less developed cities from the perspective of different city grades is more significant, which can narrow the digital divide and realize the intelligent sharing of information resources and data elements; thirdly, the mechanism analysis shows that digital economy affects the entrepreneurial vitality through the two paths of optimizing the business credit environment and coordinating the digital infrastructure.

Based on the above findings, this paper makes the following recommendations:

Firstly, foster digital technology application scenarios and encourage the transformation and upgrading of traditional industries. As digital economy effectively drives the entrepreneurial vitality of cities, the government should vigorously promote the innovative application of digital technology. For example, in the Big Data Pilot Zone policy, the exploration of new business models and technology applications can be integrated, which not only includes supporting the innovative and entrepreneurial activities of high-tech enterprises but also involves helping enterprises in traditional industries to realize the upgrading of business models and supply chains through digital transformation.

Secondly, improve the policy environment for digital entrepreneurship and give full play to regional collaborative innovation. As the treatment variable of digital economy, the policy effects on the entrepreneurial vitality of cities cannot be ignored. To create an environment conducive to digital entrepreneurship in the region, the government can continue to introduce a series of supportive policies such as tax incentives, research and development subsidies, and entrepreneurship guidance to improve the urban digital entrepreneurship environment. At the same time, the establishment of regional cooperation platforms, and the sharing of infrastructure such as data centers are important channels for resource sharing and mutually beneficial cooperation among different regions to give full play to regional collaborative innovation. Local governments in the region can coordinate policy formulation, form a consistent development direction and policy support, and promote regional community construction, thus boosting the flow of data elements.

Thirdly, optimize the regional business credit environment and promote the construction of digital infrastructure. According to the mechanism analysis, a good business credit environment is crucial to stimulate entrepreneurial vitality, and the government can improve the transparency and availability of credit data by establishing and improving the credit evaluation system to reduce transaction costs and risks between enterprises. At the same time, the construction of digital infrastructure, including both traditional telecommunications infrastructure and new digital infrastructure, can be taken into practice to lay a solid foundation for digital economy. Reliable digital infrastructure can not only promote the rapid flow of information but also support emerging online business services like e-commerce, telemedicine, etc., to open up a broader market and development space for entrepreneurs.

References

- [1] Ding Y L 2021 Origin, connotation and measurement of digital economy: a literature review *Tren. Soc. Sci.* 57–63
- [2] Zuo X D 2016 The G20 digital economy initiative and cybersecurity *Microcom. Appli.* 1–2
- [3] Qiu Z X and Zhou Y H 2021 Development of digital economy and regional total factor productivity: an analysis based on national big data comprehensive pilot zone *Fin. Res.* 4–17
- [4] Sun Z Y 2022 How does the development of digital economy affect manufacturing enterprises to “get rid of virtual reality”: evidence from national big data comprehensive test area *Mod. Econ. Res.* 90–100
- [5] Zhi Y P and Lu X X 2023 Establishment of national big data comprehensive pilot zone and urban entrepreneurship activity—based on empirical evidence from 284 cities *China Bus. Mar.* 84–96
- [6] He Y K, Niu G, Lu J and Zhao G C 2024 Digital governance and urban entrepreneurial vitality: evidence from the national pilot policy of information benefiting the people in China *Quan. Tech. Econ.* 47–66

- [7] Yu J, Meng Q S, Zhang Y and Jin J 2018 Digital entrepreneurship: the future directions of entrepreneurship theory and practice in the digital era *Sci. Res.* 1801–08
- [8] Guo J and Zhu Y X 2022 Digital economy, regional innovation efficiency and regional entrepreneurial vitality *J. Harbin Univ. Comm.* 98–111
- [9] Bao Z R 2017 Research on the Phenomenon of “Data Kidnapping” under the Background of Big Data (Zhengzhou: Zhongyuan University of Technology)
- [10] Wang K and Chao X J 2023 Research on the impact of new digital infrastructure on urban entrepreneurial activity *J. Xi 'an Univ. Fin. Econ.* 51–63
- [11] Teng L 2021 Research on Financing Constraints of SMEs from the Perspective of Digital Inclusive Finance (Sichuan: Sichuan University)
- [12] Huang M G and Yang Y 2016 The innovative research on the financing models of small and medium-sized enterprises in poor areas—based on the perspective of new Internet finance formats *Tech. Econ. Man. Res.* 55–59
- [13] Wu R and Zhang Y 2023 Research on the influence of social credit on entrepreneurial activity—empirical evidence from creating model cities based on social credit system *J. Inner Mongolia Agri. Univ.* 37–43
- [14] Zhao J J, Wei J, Liu J D and Liu T J 2020 Does trust help improve entrepreneurial performance—empirical test based on 876 farmer entrepreneurs *China Rur. Obs.* 90–108
- [15] Huang Z 2019 Social credit system in China: ex-ant screening or ex-post punishment—evidence from venture investment *J. Shanxi Univ. Fin. Econ.* 27–41
- [16] Cai Y Z 2016 Opportunity and challenge in the innovation and pioneering work of “Internet plus” action: analysis in the perspective of technological revolution and technical-economical pattern *Seek. Tru.* 43–52
- [17] Chao X J, Lian Y M, Luo L K 2021 Impact of new digital infrastructure on high-quality development of manufacturing *Fin. Tra. Res.* 1–13
- [18] Chen J Y 2017 Research on the development of national big data comprehensive pilot area in Guizhou *Soc. Sci. Guizhou* 149–155
- [19] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W and Robins J 2018 Double/debiased machine learning for treatment and structural parameters *Econ. J. C* 1–68
- [20] Athey S, Tibshirani J and Wager S 2019 Generalized random forests *Ann. Stat.* 1148–78
- [21] Yang J C, Chuang H C and Kuan C M 2020 Double machine learning with gradient boosting and its application to the big n audit quality effect *J. Econ.* 268–283
- [22] Zhang T and Li J C 2023 Network infrastructure, inclusive green growth, and regional inequality: causal inference based on double machine learning *J. Quan. Tech. Econ.* 113–135
- [23] Ni X M, Zheng T T and Zhao H M 2023 Can military-civilian integration reduce the cost of equity capital for high-tech enterprises? An empirical study based on double machine learning *Sys. Eng. Theo. Prac.* 1630–50
- [24] Xie X L, Shen Y and Zhang H X 2018 Can digital finance boost entrepreneurship—evidence from China *Econ. Quat.* 1557–80
- [25] Lin X Y, Shen Z X and Zhuang H M 2024 Impact and mechanism of cross border e-commerce reform on urban entrepreneurial vitality: a case study of the cross-border e-commerce comprehensive pilot zone investigation *J. Hunan Agri. Univ.* 91–102
- [26] Wang Q, Li J, Ding K K and Lei L 2023 Digital infrastructure, factor allocation efficiency and urban-rural income gap *Stat. Deci. Mak.* 29–34
- [27] Zhao T, Zhang Z and Liang S K 2020 Digital economy, entrepreneurship, and high-quality economic development: empirical evidence from urban China *Man. Wor.* 65–76
- [28] Bai J H, Zhang Y X and Bian Y C 2022 Does innovation-driven policy increase entrepreneurial activity in cities—evidence from the national innovative city pilot zone *China Ind. Econ.* 73–85

- [29] Tian K, Huang K and Hang W B 2023 Digital economy, market potential and rural revitalization—causal inference based on double machine learning J. Shanxi Univ. Fin. Econ. 73–85
- [30] Nunn N and Qian N 2014 US food aid and civil conflict Am. Econ. Rev. 1630–66
- [31] Sun T Y, Lu Y and Cheng L H 2020 Implementation effect of resource exhausted cities' supporting policies, long-term mechanism and industrial upgrading China Ind. Econ. 98–116
- [32] Su C 2022 Research on the Influence of Cooperative Innovation Network on Regional Innovation Performance (Shanghai: East China Normal University)
- [33] Farbmacher H, Huber M, Lafférs L, Langen H and Spindler M 2022 Causal mediation analysis with double machine learning Econ. J. 277–300