sustainability

MDPI

Article

From Polarization to Convergence: How Data Agglomeration Shapes Within-City Income Disparities via Technological Agglomeration and Diffusion

Yanqing An [1], Jiaxi Kang [1], Junxia Zhang [2], Juan Liu [1] and Zijian Cao [1,*]

[1] School of Economics, Lanzhou University, Lanzhou 730000, China
[2] School of Economics and Management, Lanzhou University of Technology, Lanzhou 730050, China
* Correspondence: caocao_d320@163.com

Abstract

This paper develops a mathematical model grounded in traditional agglomeration theory to analyze the impact of data factor agglomeration on income disparity. Using panel data from 290 prefecture-level cities over the period 2013–2023 and employing the National Big Data Comprehensive Pilot Zone as a proxy for data factor agglomeration, we conduct empirical tests. The results show that: (1) Data factor agglomeration narrows within-city income disparities on average and generates significant income gains for cities within 300 km, while distant cities beyond 600 km experience widening disparities. However, due to the limited number of clusters at the province level, the statistical significance of this effect is sensitive to the clustering level. (2) Data factor agglomeration widens income disparities in the eastern region, while it helps narrow disparities in the central and western regions, in cities with smaller economic and population scales, and in areas with lower levels of digital information infrastructure. Meanwhile, at the 10th percentile, data factor agglomeration narrows income disparities, whereas at the 90th percentile, it exacerbates income disparities. (3) Mechanism analysis shows that data factor agglomeration significantly promotes both technological agglomeration and technological diffusion, with the former associated with widening and the latter with narrowing income disparities.

Keywords: data factor agglomeration; income disparity; technological agglomeration; technological diffusion

1. Introduction

With the deepening of the new technological revolution and industrial transformation, data has become a key production factor driving economic development and productivity growth. It has permeated production, distribution, circulation, consumption, and services, reshaping labor, capital, and production processes, and has generated large-scale accumulation and spillover effects in China's data factor market [1]. According to the Digital China Development Report (2024), data production reached 41.06 zettabytes in 2024, up 25% year on year, and total computing power reached 280 EFLOPS, with over 60% of new capacity concentrated in eight major hub regions. Despite these spillover benefits, structural challenges remain. On the supply side, large volumes of data are underutilized, market-based allocation mechanisms remain underdeveloped, and data silos and circulation bottlenecks limit the conversion of data resources into productive capital. On the demand side, firms face constraints in data adoption, technical capacity, and value exploitation, restricting the contribution of data to the real economy. This supply–demand mismatch indicates that

Check for updates

Academic Editor: Fabrizio D'Ascenzo

Received: 10 September 2026
Revised: 25 October 2026
Accepted: 27 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

China's data factor market is still at an early stage, with substantial room for improvement. To exploit data agglomeration effects and position data as both a foundational resource and an innovation engine, the National Development and Reform Commission, the Ministry of Industry and Information Technology, and the Cyberspace Administration of China jointly approved the first national comprehensive big data pilot zone in February 2016. This marked a pivotal step in the market-oriented reform of data factors and provided an institutional platform for building a data market system. As data agglomeration reshapes production, economic activities, and the spatial distribution of income within cities, designing a distribution mechanism that balances efficiency and spatial equity has become central to promoting inclusive urban development, narrowing within-city income disparities, and advancing common prosperity.

Academia has developed a relatively systematic understanding of data factors and within-city income disparities, which is primarily reflected in the following aspects. First, the theoretical evolution concerning the essential characteristics of data factors. As the primary factor of production in the digital economy, data has been vividly likened to the "oil of the new era" [2]. From an information processing perspective, the academic community defines data as unprocessed facts and observations, emphasizing that it can be transformed into valuable information through analysis and processing under specific circumstances [3]. From a production perspective, data is regarded as a by-product of economic activities [4,5]. With the in-depth development of the digital economy, the attributes of data are undergoing a fundamental transformation, gradually evolving from passive generation to active collection, specialized production, and systematic accumulation as a strategic resource [6]. Furthermore, data exhibits unique techno-economic characteristics, including low cost, large-scale availability, ease of replication and transmission, as well as non-competitiveness, non-excludability, non-consumability, externalities, and timeliness [7–9]. Second, the inevitability of data factor agglomeration and its value realization mechanism. As a factor of production, data not only possesses the potential for value creation but also requires integration with the means of production; that is, it must enter the production process to achieve value appreciation [10,11]. Single data does not have intrinsic value and is difficult to directly generate economic value. Only by aggregating and integrating scattered data into large-scale data resources, forming raw data [6], and using big data analysis techniques to transform raw data into knowledge or effective information can we uncover the potential economic value hidden behind macro portraits of different regions and groups, and ultimately achieve value creation. Therefore, the aggregation of data elements is not only necessary in reality but also an inevitable path for its value realization. Third, the primary research directions and limitations of existing studies on data factor agglomeration. Current research has primarily focused on the impacts of data factor agglomeration on industrial structure, technological innovation, economic development, total factor productivity, development resilience, and green development [12–16]. However, from the perspective of social welfare, particularly its systematic impact on income distribution effects, research remains relatively scarce. Fourth, the causes of within-city income disparities. Although studies indicate that the income gap among Chinese residents has passed the peak or turning point of the income distribution differentiation curve and is showing a narrowing trend, within-city disparities remain significant [17]. Some scholars attribute these disparities to institutional factors such as land policies and urban planning that cause resource allocation to deviate from spatial efficiency [18], while others emphasize the roles of urban spatial structure, job displacement, and the core-periphery dynamics between downtown centers and suburban outskirts [19,20]. To narrow within-city income disparities, some scholars argue that labor mobility [21], infrastructure construction [22], industrial structure upgrading [23], and regional innovation [24,25] play decisive roles in reducing regional disparities. With the

advancement of the Fourth Industrial Revolution, scholars have gradually begun to pay attention to the impact of the digital economy on within-city income disparities, yet limited attention has been given to the influence of data factor agglomeration on within-city income disparities.

This paper contributes to the sustainability literature by examining whether the agglomeration of data factors, as a key driver of the digital economy, promotes inclusive urban development rather than deepening intra-city spatial inequality. Understanding this relationship is essential for designing digitalization policies that align economic efficiency with social equity, a central concern of sustainable development research. The remainder of this paper is structured as follows. Section 2 presents the theoretical framework and model derivation. Section 3 describes the data and empirical models in detail. Section 4 conducts the baseline empirical analysis. Section 5 performs mechanism testing. Section 6 examines spatial spillover effects. Section 7 concludes with policy implications.

2. Theoretical Analysis

Unlike traditional agglomeration, which is closely tied to geographical distance and typically produces convergence at short distances but divergence over long distances [26], data factor agglomeration concentrates data circulating in virtual space—structured and unstructured—within specific geographic spaces through technology, markets, or policy. Physical space provides the foundation for virtual agglomeration, while virtual space feeds back through value creation. This bidirectional interaction enables the large-scale aggregation, systematic integration, and synergistic interaction of data factors, forming a new resource-allocation paradigm in the digital economy (Figure 1).

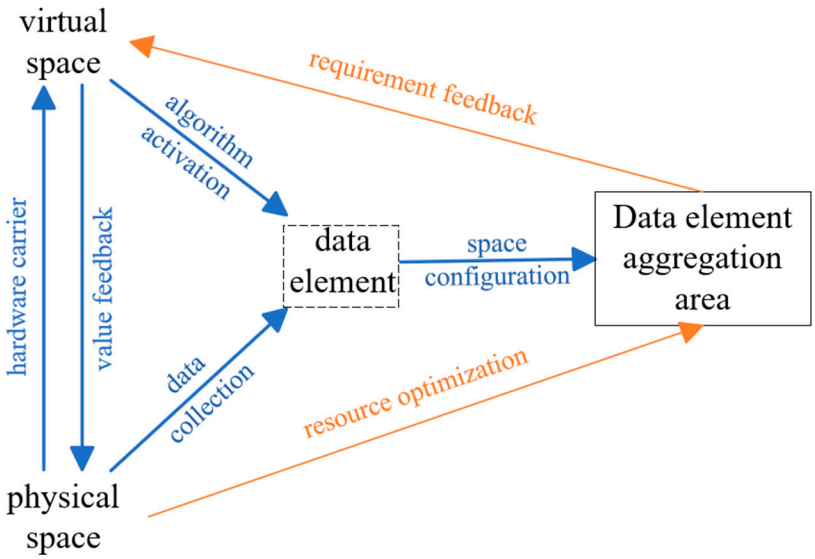


Figure 1. Aggregation of Data Elements. Blue arrows indicate the main operational flow, including data collection, algorithm activation, spatial configuration, and virtual-physical value interaction; orange arrows represent the regulation and feedback loop for resource optimization.

2.1. Theoretical Hypotheses

2.1.1. Data Factor Agglomeration and Within-City Income Disparities

According to traditional agglomeration theory, data factor agglomeration exhibits both Marshallian externalities and Jacobs externalities [14], which can promote the geographical concentration of data enterprises and cross-sectoral technological spillovers, thereby achieving economies of scale. However, the extraction of data value is highly dependent on storage capacity and computing power, displaying distinct characteristics of competitiveness

and path dependence [27]. As a result, data agglomeration entities that first achieve scale advantages—typically concentrated in the downtown core areas—tend to attract more factor resources, inducing a "winner-takes-all" phenomenon and a "Matthew effect". This causes high-skilled labor and digital capital to cluster in the city center, widening the spatial income disparity between the downtown core and the peripheral outskirts within the city. This negative outcome is precisely the "polarization" result predicted by classical agglomeration theory. Nevertheless, in the current stage of China's digital economy development, which is characterized by deepening application, orderly development, and inclusive sharing, the spatial allocation and income distribution of data factors exhibit significant policy malleability, providing a possibility to reverse the within-city polarization trend. First, infrastructure reallocation. Under the strategic frameworks of "Smart City" and "Digital Village" initiatives, digital infrastructure investments are extending from the downtown core to the suburban and rural outskirts. This alters the intra-city spatial distribution of digital infrastructure, creating a material precondition for peripheral districts to host and utilize data factors. Second, factor synergy. Following intra-city infrastructure improvements, data factors, through synergistic allocation with traditional factors, empower traditional industries in the outskirts. This enhances the marginal product of capital and labor in these peripheral areas, optimizes factor allocation efficiency, and realizes the value creation of data factors, thereby narrowing within-city income disparities. Third, income redistribution. Unlike the intensification of polarization effects caused by increasing returns to scale in classical theory, under China's inclusive digital policy guidance, the excess returns generated by increasing returns to scale in the core are partially channeled back to peripheral districts through mechanisms such as inclusive digital finance, municipal tax transfer payments, and urban-rural integration policies. This enables a broader sharing of the growth dividends from data factor agglomeration, thereby reducing within-city spatial inequalities.

Therefore, although data factor agglomeration under market mechanisms may induce risks of intra-city imbalance, under the influence of China's current policy guidance and inclusive development mechanisms, the spatial allocation of data factors within cities has achieved a path reversal from polarization to convergence, contributing overall to the narrowing of within-city income disparities (Figure 2).

H1. *Data factor agglomeration can narrow within-city income disparities.*

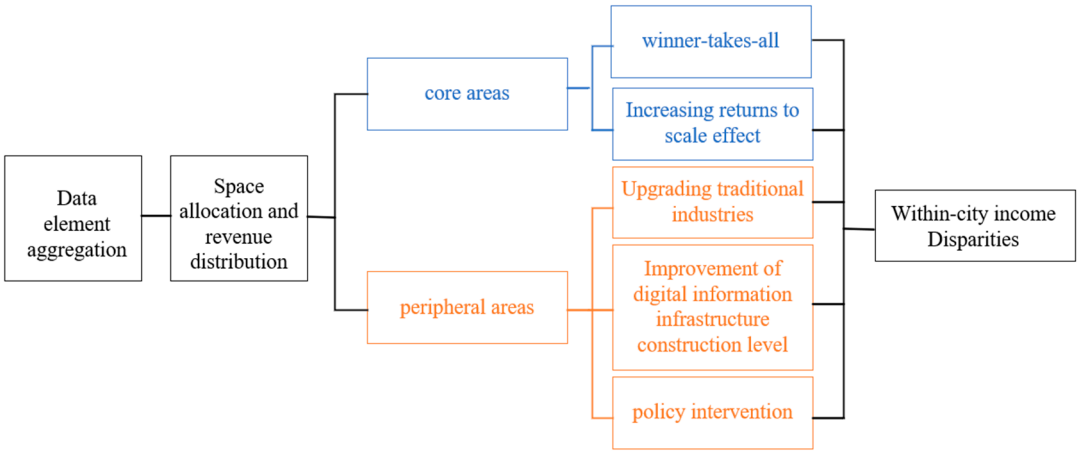


Figure 2. Agglomeration of Data Elements and Within-city Income Disparities. Blue boxes represent the transmission channel in core areas; orange boxes denote the transmission channels in peripheral areas.

2.1.2. Technological Innovation Effect of Data Factor Agglomeration

Data factors serve as the core foundation for technological R&D and upgrading, and their agglomeration inevitably accelerates the localized concentration of technology and economic activities, with the degree of technological agglomeration exceeding that of economic agglomeration [28]. This process amplifies polarization effects through three mechanisms. The first is the data pool formation mechanism. Data factor agglomeration concentrates data from non-agglomeration areas into agglomeration areas, forming massive, diverse, and high-quality data pools that support algorithm training and model optimization, alleviate information asymmetry, reduce R&D uncertainty, attract a large number of innovation entities to cluster, and accelerate the formation of local technological hubs [29]. The second is the digital platform reinforcement mechanism. The development of new technologies such as quantum computing and large-scale model training accelerates the concentration of data factors in computing power hub regions, while also giving rise to comprehensive digital platforms that serve as integrated carriers of data and technology. This further enhances the technological appeal of the urban core, drives enterprises and R&D teams to concentrate in central business districts, creates a spatial concentration of technological activities, and promotes technological progress in the core areas, yet it also widens spatial income disparities within the city. The third is the factor complementarity mechanism. Data factors exhibit complementarity with technological factors such as R&D talent, patented technologies, and innovation equipment: data provide the raw material for technological innovation, while technology endows data with economic value. This complementarity spatially forms a virtuous cycle between data agglomeration and technological agglomeration, making technological agglomeration difficult to reverse once established, continuously enhancing labor productivity in the core urban districts, and further widening within-city income disparities [30].

Technological diffusion facilitates the flow of knowledge and technology across entities and regions within network organizations [31], mitigating the intra-city development imbalances arising from technological agglomeration. Data factors, characterized by low cost, broad coverage, and high adaptability in technology dissemination, significantly lower the barriers to technological diffusion, accelerating the spread of innovation outcomes to non-agglomeration areas. In particular, data factor agglomeration promotes the transformation of tacit knowledge into explicit technology and leverages digital carriers to transcend spatial and temporal constraints. Through continuous data feedback and iterative optimization, it constantly enhances technological adaptability and diffusion depth. From the perspective of spatial evolution, inter-city technological diffusion initially predominantly occurs between the pilot zones and neighboring cities with strong industrial linkages. As the scale of data factors expands and network externalities strengthen, diffusion pathways gradually break through administrative boundaries, achieving leapfrog diffusion to distant cities with robust digital infrastructure and absorptive capacity. Crucially, when neighboring or distant cities absorb these technological spillovers from the pilot zones, digital technologies are typically applied to upgrade local traditional industries. This localized industrial upgrading drives income growth in the peripheral outskirts of the receiving cities, ultimately narrowing their intra-city spatial income disparities [25].

Accordingly, the following hypothesis is proposed (Figure 3):

H2. *Technological agglomeration widens within-city income disparities through the "polarization effect".*

H3. *Technological diffusion narrows within-city income disparities through the "trickle-down effect".*

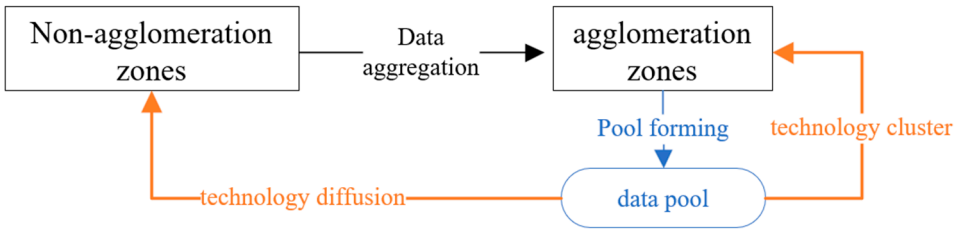


Figure 3. Technological Innovation Effect. Blue lines represent the self-reinforcing loop within agglomeration zones; orange lines denote the cross-region spillover loop.

2.1.3. Spatial Spillover Effects of Data Factor Agglomeration

Data factor agglomeration generates positive externalities by facilitating knowledge spillovers and technological diffusion [32], which serves as a precondition for the formation of spatial spillover effects. Unlike traditional factors, data factors rely on digital infrastructure to achieve low-cost cross-regional flow and replication. Their effects extend to surrounding regions through knowledge spillovers, technological diffusion, and service exports, thereby driving industrial coordination and productivity improvements, and subsequently influencing within-city income distribution patterns. Based on the New Economic Geography theory, the spatial spillover effects of data factor agglomeration exhibit distance-decay characteristics and are shaped by spatial heterogeneity in key factors such as digital infrastructure, human capital endowment, and policy environment. This further accelerates the formation of a "core-periphery" spatial structure, manifesting as radiation effects in close proximity and siphon effects over long distances. Specifically, within the adjacent radiation radius, surrounding cities can more easily access computing power, digital platforms, and industrial spillovers, which attracts the relocation of digital industries. The injection of digital factors empowers traditional industries and creates local employment, thereby driving the development of the peripheral areas of these cities, reducing labor outflows, and consequently narrowing the within-city income gap between urban centers and outskirts. Once the effective radiation radius is exceeded, technology spillover costs rise, and factor return differentials widen. Facing the strong siphon effect of the pilot zones, the core urban areas of distant cities can maintain relative economic resilience due to their more complete industrial foundations. In contrast, their peripheral areas, lacking industrial competitiveness, are more vulnerable to labor and capital outflows, leading to industrial hollowing-out. This asymmetric shock further concentrates resources into the urban centers of distant cities, ultimately widening the within-city income disparities in these cities.

H4. *Data factor agglomeration has spatial spillover effects on within-city income disparities.*

2.2. Model Derivation

2.2.1. Model Setup

Consider two regions, the core region C and the peripheral region P. Assume that the distance between the two regions is d, labor is immobile across regions, and capital is freely mobile with a fixed total amount.

- Consumer Behavior

  The utility function of the representative consumer in each region is:

$$U_i = lnC_i - \frac{\psi}{2}L_i^2, i\epsilon\{C, P\}$$

where C_i is consumption of the final good, L_i is labor supply, and $\psi > 0$ is the disutility coefficient of labor.

The budget constraint is:

$$P_i C_i = w_i L_i + \Pi_i + T_i$$

where the price P_i is normalized to 1, w_i is the wage, Π_i is firm profit, and T_i is the redistribution of data factor returns.

From the utility maximization principle, we obtain:

$$L_i = \frac{w_i}{\psi} \Rightarrow L_i \propto w_i$$

- Final Good Production

The final good producer uses capital K_i, labor L_i, and data factor D_i. The production function is specified as:

$$Y_i = A_i K_i^{\alpha} L_i^{\beta} D_i^{\gamma}, \alpha + \beta + \gamma < 1$$

where A_i is total factor productivity, which is influenced by technological agglomeration. Capital is freely mobile, and the cost of capital r is uniform across regions.

From profit maximization, we obtain:

$$w_i = \beta \frac{Y_i}{L_i}, r = \alpha \frac{Y_i}{K_i}, p_{d,i} = \gamma \frac{Y_i}{D_i}$$

2.2.2. Equilibrium Solution

The data stock D_i is jointly determined by local data generation, the level of data factor agglomeration Z_i, and cross-regional technological diffusion:

$$D_C = \theta_C Y_C + Z_C + \phi e^{-\lambda d} D_P$$
$$D_P = \theta_P Y_P + Z_P + \phi e^{-\lambda d} D_C$$

where θ_i is the efficiency of local data generation, with $\theta_C > \theta_P$; Z_i is the exogenously given level of data factor agglomeration, with $Z_C > Z_P$. Data factor agglomeration not only directly increases the local data stock but also spills over to neighboring regions. ϕ denotes the degree of technological diffusion $(0 < \phi < 1)$, and λ is the distance decay coefficient. Let $\rho = e^{-\lambda d} \in (0, 1)$.

Introducing the technological agglomeration effect into total factor productivity, we set $A_i = A_0 D_i^{\mu}$ and substitute it into the production function

$$Y_i = A_0 K_i^{\alpha} L_i^{\beta} D_i^{\mu+\gamma}$$

where μ is the intensity of technological agglomeration.

- Capital allocation

From the free mobility of capital, $r = \alpha Y_i / K_i$, we obtain the capital stock ratio between the two regions

$$\frac{K_C}{K_P} = \left(\frac{L_C^{\beta} D_C^{\mu+\gamma}}{L_P^{\beta} D_P^{\mu+\gamma}} \right)^{\frac{1}{1-\alpha}}$$

- Labor supply

From $L_i = w_i / \psi$ and $w_i = \beta Y_i / L_i$, we get

$$L_i^2 = \beta \frac{Y_i}{\psi} \Rightarrow L_i \propto Y_i^{1/2}$$

- Output ratio

Substituting the capital ratio and labor relationship into the production function, after algebraic manipulation, we obtain the output ratio between the two regions

$$y = \frac{Y_C}{Y_P} = \left(\frac{D_C}{D_P}\right)^{\Theta}, \Theta = \frac{\mu + \gamma}{1 - \alpha - \beta/2} > 0$$

It is evident that the spatial income disparity is positively correlated with the data stock disparity.

2.2.3. Equilibrium and Comparative Static Analysis

- Determination of the data stock ratio

  By solving the simultaneous data stock equations, we obtain

$$D_C = \frac{\theta_C Y_C + Z_C + \phi\rho(\theta_P Y_P + Z_P)}{1 - (\phi\rho)^2}$$
$$D_C = \frac{\theta_C Y_C + Z_C + \phi\rho(\theta_P Y_P + Z_P)}{1 - (\phi\rho)^2}$$

  Define the data stock ratio $x \equiv D_C/D_P$, the output ratio $y \equiv Y_C/Y_P = x^{\Theta}$, and the data factor agglomeration level ratio $z \equiv Z_C/Z_P > 1$.

  Using the relationships $\frac{Z_C}{Y_P} = \frac{Z_C}{Y_C} \cdot \frac{Y_C}{Y_P} = z_C x^{\Theta}$ and $\frac{Z_P}{Y_P} = z_P$, after substitution and rearrangement, we obtain

$$x = \frac{(\theta_C + z_C)x^{\Theta} + \phi\rho(\theta_P + z_P)}{(\theta_P + z_P) + \phi\rho(\theta_C + z_C)x^{\Theta}}$$

  Let $a = \theta_C + z_C$ and $b = \theta_P + z_P$. Then, we obtain the concise self-consistent equation:

$$x = \frac{ax^{\Theta} + \phi\rho b}{b + \phi\rho a x^{\Theta}}$$

- Comparative Static Analysis

  Rewrite it as an implicit function:

$$F(x, b) = x - \frac{ax^{\Theta} + \phi\rho b}{b + \phi\rho a x^{\Theta}} = 0$$

  Let $M = b + \phi\rho a x^{\Theta} > 0$. By the implicit function theorem, compute $\partial F/\partial b$ and $\partial F/\partial x$:

$$\frac{\partial F}{\partial b} = -\frac{ax^{\Theta}(\phi^2\rho^2 - 1)}{M^2} > 0$$
$$\frac{\partial F}{\partial x} = 1 - \frac{ab\Theta x^{\Theta-1}(1 - \phi^2\rho^2)}{M^2}$$

  In the neighborhood of the equilibrium, it is typically true that $\partial F/\partial x > 0$. Then $\partial x/\partial b < 0$, indicating that an increase in the comprehensive efficiency b of the peripheral region reduces the data stock ratio x, thereby narrowing the income disparity.

  Similarly, treating $\phi\rho$ as a parameter t and taking its partial derivative yields

$$\frac{\partial x}{\partial t} < 0$$

  That is, strengthening technological diffusion also reduces the data stock ratio and thus narrows the income disparity.

  According to classical agglomeration theory, data factor agglomeration reinforces the core-periphery structure and widens income disparities. In our model, if we ignore policy support for the peripheral region, keep b unchanged and rely solely on market forces, then $a > b$ and the diffusion intensity t is small. In equilibrium, $x > 1$. In this case, further increasing the agglomeration level in the core region would enlarge the disparity.

However, our model introduces an exogenous data factor agglomeration policy whose core lies in simultaneously enhancing the peripheral region's comprehensive efficiency b and the technological diffusion intensity t. These measures can effectively reduce x, thereby reversing the polarization trend. From a mathematical perspective, the critical condition for this reversal is:

$$\frac{\partial x}{\partial b} * \Delta b > \frac{\partial x}{\partial a} * (-\Delta a)$$

Since the initial comprehensive efficiency b_0 of the peripheral region is significantly lower than a_0 of the core region, the same policy input generates a larger marginal increase Δb in the periphery than the marginal increase Δa in the core, satisfying $\Delta b > (-\Delta a)$ and consequently reducing x. In policy practice, the peripheral region, supported by policy and benefiting from a lower base, achieves higher marginal returns, making Δb significantly larger than Δa. This drives down the data stock ratio x and ultimately leads to the convergence of within-city income disparities.

2.2.4. Spatial Spillover Effects

Extending the model to a continuous space, let the core region be located at the origin and the peripheral region at a distance d. Then $\rho = e^{-\lambda d}$, and the data stock of the peripheral region is given by

$$D_P(d) = \theta_P Y_P(d) + Z_P(d) + \phi e^{-\lambda d} D_C$$

When d is small, $e^{-\lambda d} \approx 1$, the diffusion effect is strong. This raises D_P and narrows the gap with the core region, exhibiting a short-range convergence effect. When d exceeds a certain threshold d^*, $e^{-\lambda d} \approx 0$, and the diffusion effect disappears. In this case, the core region, through technological agglomeration, attracts capital and labor, leading to factor outflows and declining incomes in distant peripheral regions, which manifests as a polarization effect over long distances. Here, the critical distance d^* is determined by the condition that the net benefit equals zero, i.e., there exists d^* such that $\phi e^{-\lambda d} D_C$ exactly offsets the polarization loss.

Synthesizing the above analysis, when data factor agglomeration policies are coupled with the extension of digital infrastructure to urban outskirts and the development of cross-city data circulation networks, the overall net effect can significantly narrow within-city income disparities. This provides a new perspective for understanding intra-city spatial coordination and inclusive urban development in China's digital economy era.

3. Data Sources, Variable Selection, and Model Specification

3.1. Data Sources

Given data availability and administrative adjustments such as the policy of converting counties into cities, after excluding certain city samples, this study uses panel data from 290 prefecture-level and above cities in China from 2013 to 2023. The data are sourced from the National Bureau of Statistics website, various city statistical yearbooks and national economic and social development bulletins, China City Statistical Yearbook, China Urban Construction Statistical Yearbook, and the CSMAR database. On the basis of manually collecting and matching data year by year from official provincial and municipal websites, statistical bureaus, and national economic and social development statistical bulletins, linear interpolation and other methods were employed to fill in missing values. Table 1 reports the descriptive statistics of the main variables used in this study.

Table 1. Variable definitions and descriptive statistics.

Variables	Obs.	Mean	S.D.	Min	Max
Gap	3190	0.7341	0.2150	0.0049	0.9950
Dec	3190	0.1702	0.3758	0	1
Gdp	3190	16.7373	0.9516	14.2435	19.9729
Ind	3190	0.4482	0.0961	0.1644	0.8485
Pop	3190	5.7195	0.9803	1.3714	8.1756
Hc	3190	0.0210	0.0258	0.00006	0.1848
Med	3190	0.4986	0.1786	0.1573	1.3766
Fir	3190	2.7343	1.2578	0.6704	21.3014

3.2. Variable Selection

3.2.1. Dependent Variable

Within-city Income Disparity (Gap). Current academic approaches to measuring income disparity mainly include absolute income disparity [33] and relative income disparity. Since absolute income disparity incorporates growth-induced gaps between regions, relative measures better reflect relative disparities. Generally, relative income disparity is measured by the ratio of economic indicators between two regions, or by using indices such as the Gini coefficient, Theil index, Atkinson index, and coefficient of variation [34,35]. Nighttime light data are primarily derived from satellite observations of the Earth at night. By capturing the intensity and distribution of ground lighting, these data serve as a proxy for regional economic development and income distribution. Mveyange (2015) [36] used nighttime light data to examine regional inequality across 1235 regions in Africa and found a strong correlation between light inequality and regional income inequality. Subsequently, a growing number of scholars, including Lessmann and Seidel (2017) [37], Galimberti et al. (2023) [38], and Niu et al. (2026) [39], have employed nighttime light data to construct measures of regional income inequality. Accordingly, to measure within-city income disparity, this paper utilizes high-resolution grid-level data. Specifically, we divide each city into 0.5 km × 0.5 km grid cells. By matching the nighttime light intensity with the population density of each grid cell, we rank the cells by per capita light intensity and plot the Lorenz curve based on cumulative population shares. The Gini coefficient is then calculated across the grid cells within each city to reflect the spatial fairness of income distribution between the urban center and the outskirts. The Gini coefficient ranges between 0 and 1, with values closer to 1 indicating larger within-city income disparities, and values closer to 0 indicating smaller disparities.

3.2.2. Core Explanatory Variable

Data Factor Agglomeration (Dec). Due to data availability constraints and the difficulty of quantifying the process of data factor agglomeration, a unified indicator system to capture the flow and aggregation state of data factors in economic activities has yet to be established in the academic community. Against this backdrop, drawing on existing literature [40,41], this paper treats the National Big Data Comprehensive Pilot Zone as a quasi-natural experiment to identify data factor agglomeration, constructing the variable by interacting the group dummy variable ($treat_i$) with the time dummy variable ($post_t$). If city i is located in the National Big Data Comprehensive Pilot Zone, $treat_i$ is assigned a value of 1; otherwise, it is 0. If city i becomes a National Big Data Comprehensive Pilot Zone in year t, $post_t$ is assigned a value of 1; otherwise, it is 0. It should be noted that in 2015, the Action Outline for Promoting the Development of Big Data issued by the State Council explicitly proposed the implementation of regional pilot programs and the advancement of big data Comprehensive Pilot Zone construction, emphasizing the need to

coordinate data resources, leverage data flows to guide technology flows, material flows, capital flows, and talent flows, and strengthen the agglomeration of the big data industry. This indicates that the Big Data Comprehensive Pilot Zone can, to a certain extent, reflect data factor agglomeration. The pilot zones cover developed regions such as Beijing-Tianjin-Hebei, Shanghai, and Guangdong, as well as less developed regions such as Inner Mongolia, Guizhou, and Henan. In February 2016, the National Development and Reform Commission, the Ministry of Industry and Information Technology, and the Cyberspace Administration of China formally approved the construction of Guizhou as the country's first National Big Data Comprehensive Pilot Zone (source: National Development and Reform Commission of the People's Republic of China, "Innovation Experiments Carried Out in the National Big Data (Guizhou) Comprehensive Pilot Zone" (https://www.ndrc.gov.cn/xwdt/ztzl/szhzxhbxd/zxal/202007/t20200703_1233042.html (accessed on 3 July 2025)). Therefore, the implementation year for all National Big Data Comprehensive Pilot Zones is 2016.

3.2.3. Control Variables

The control variables in this study include: Regional economic development level (Gdp), measured by the logarithm of gross regional product of each prefecture-level city; Industrial structure (Ind), measured by the ratio of the gross output value of the tertiary industry to the gross regional product; Population density (Pop), measured by the logarithm of the ratio of the registered population to the land area of the administrative region; Human capital (Hc), measured by the ratio of the number of students enrolled in ordinary higher education institutions to the total population at the end of the year; Healthcare level (Med), measured by the number of hospital and health center beds per hundred people; Financial development level (Fir), measured by the ratio of the outstanding balance of deposits and loans of financial institutions at the end of the year to the gross regional product.

3.3. Model Specification

Using the event study approach [42,43], a Difference-in-Differences (DID) model is employed to examine the impact of data factor agglomeration on within-city income disparities:

$$GAP_{it} = \alpha_0 + \alpha_1 treat_i * post_t + \gamma X_{it} + v_i + \mu_i + \varepsilon_{it} \quad (1)$$

where i denotes the city, t denotes the year, GAP_{it} represents the within-city income disparity, $treat_i$ is a dummy variable indicating whether the city belongs to the National Big Data Comprehensive Pilot Zone, $post_t$ is a dummy variable for the period before and after the implementation of the National Big Data Comprehensive Pilot Zone policy, X_{it} represents a set of control variables at the prefecture-level city that may affect within-city income disparities, v_i denotes the year fixed effects, μ_i denotes the city fixed effects, and ε_{it} is the random error term.

4. Empirical Results and Analysis

To examine the effect of data factor agglomeration on within-city income disparities, this paper conducts empirical analyses from multiple dimensions, investigating the baseline impact of the National Big Data Comprehensive Pilot Zone on within-city income disparities and exploring the spatial spillover effects of data factor agglomeration.

4.1. Baseline Regression Results

Table 2 reports the estimation results of the impact of the National Big Data Comprehensive Pilot Zone on within-city income disparities. Column (1) presents the regression results using only the core explanatory variable, without including control variables, and employs fixed effects with city and year controls along with cluster-robust standard errors.

Column (2) adds a series of control variables at the city level that may affect data factor agglomeration and within-city income disparities, based on Column (1). The results in Table 2 show that the coefficient of the core explanatory variable is significantly negative at the 5% significance level. After adding control variables, the policy effect becomes more pronounced, with the Gini coefficient reduction expanding to 0.0156, indicating that the role of data factor agglomeration in narrowing income disparities is robust. It is evident that, regardless of whether control variables are included, data factor agglomeration can reduce within-city income disparities compared to non-pilot areas, providing preliminary support for Hypothesis 1.

Table 2. Baseline regression: Data Factor Agglomeration and Within-city Income Disparities.

Variables	(1) Gap	(2) Gap
Dec	−0.0123 **	−0.0156 ***
	(0.0054)	(0.0054)
Gdp		−0.0300 **
		(0.0118)
Ind		−0.0435
		(0.0357)
Pop		0.0768 **
		(0.0336)
Hc		0.5346 ***
		(0.1867)
Med		−0.0210
		(0.0241)
Fir		−0.0030
		(0.0025)
Constant	0.7362 ***	0.8266 ***
	(0.0009)	(0.2740)
City FE	Yes	Yes
Year FE	Yes	Yes
R^2	0.940	0.940
Observations	3190	3190

Note: *** and ** represent the significance of 1% and 5% levels, respectively. Robust standard errors clustered at the city level are reported in parentheses.

4.2. Endogeneity Issues

4.2.1. Instrumental Variable Method

This paper employs the instrumental variable method to mitigate endogeneity concerns such as reverse causality. Drawing on the approach of Huang et al. (2023) [44], the interaction term between the number of post offices per million people in 1984 and a time trend is used as the instrumental variable. From the perspective of relevance, post offices served as core hubs for information transmission and public services during the planned economy period, with their spatial distribution influenced by factors such as population density, administrative hierarchy, and economic activity intensity, exhibiting strong path dependence. In the digital economy era, they are more easily transformed and upgraded into digital infrastructure, thereby establishing an indirect but stable correlation with current data factor agglomeration, satisfying the relevance assumption of the instrumental variable. On the other hand, the number of post offices in 1984 is historical statistical information, temporally distant from contemporary data factor agglomeration, making it less susceptible to reverse influence from current economic variables, thus satisfying the exogeneity requirement of the instrumental variable. To address the concern that the instrumental variable

may affect contemporary income disparities through historical path dependence rather than solely through data factor agglomeration, and considering that Ming Dynasty post stations were the earliest and most systematic nationwide information transmission network in Chinese history, with their locations reflecting the political, economic, and geographical conditions of that time, this paper adds the interaction between the number of Ming Dynasty post stations and a time trend as a control variable, thereby indirectly supporting the exclusion restriction of the instrumental variable. Table 3 reports the estimation results of the instrumental variable method. According to the first-stage regression results, the coefficient of the instrumental variable is significantly positive at the 1% significance level, indicating that the number of post offices in 1984 increases the degree of data factor agglomeration, which is consistent with expectations. The second-stage regression results show that the coefficient of data factor agglomeration remains significantly negative at the 1% significance level. To ensure the validity of the instrumental variable, the weak instrument test yields a KP-Wald F statistic of 47.45, which is substantially above the critical value, indicating no weak instrument problem. The KP-LM statistic for under-identification has a *p*-value of 0, indicating no under-identification issue. In addition, the IV-DID estimation loses more than 700 observations relative to the baseline DID. Of these, 99 observations are lost due to missing values in the instrumental variable and the newly added control variables. The remaining loss arises from the automatic removal of singleton groups and units with no within-city variation in the endogenous treatment variable, which is required for identification in the fixed-effect IV estimator. To ensure comparability between OLS and IV estimates, we re-estimate the baseline DID specification on the exact subsample used in the IV regression; the OLS estimate on this subsample is very close to the baseline estimate.

Table 3. Instrumental variable (IV) estimation results.

Variables	(1) First Stage Dec	(2) Second Stage Gap	(3) New Sample Gap
Iv	0. 0131 ***		
	(0.0019)		
Dec		−0.1081 ***	−0.0167 ***
		(0.0381)	(0.0059)
Ming	0.0441 **	0.0106 ***	
	(0.0119)	(0.0045)	
Control variables	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
KP-Wald F-value		47.45	
KP-LM *p*-value		0.0000	
Observations	2464	2464	2464

Note: *** and ** represent the significance of 1% and 5% levels, respectively.

While the same-sample comparison facilitates a fair evaluation, we acknowledge the remaining differences between the IV and OLS estimates. Specifically, the IV estimate is larger in magnitude than the OLS estimate. This divergence can be attributed to two main factors. First, the IV identifies a Local Average Treatment Effect (LATE) for the compliers—cities whose data agglomeration status is most responsive to historical post office density—which may experience a stronger treatment effect. Second, classical measurement error in the proxy variable for data agglomeration may attenuate the OLS estimate toward zero, whereas the IV approach corrects for this attenuation. Furthermore, although controlling for Ming Dynasty post stations mitigates historical path dependence, it cannot perfectly rule out all unobserved channels. Therefore, we treat the IV estimates not as conclusive validation

but rather as suggestive, consistent evidence that supports the causal interpretation of our baseline findings.

4.2.2. Propensity Score Matching-Difference-in-Differences (PSM-DID)

Since the selection of demonstration cities for the National Big Data Comprehensive Pilot Zone is not random but rather the result of the state selecting pilot cities after comprehensively considering factors such as urban development positioning and resource endowments, there exists sample selection bias. To mitigate the impact of sample differences between the treatment group and the control group on the regression results as much as possible, this paper uses cities that established the National Big Data Comprehensive Pilot Zone during the sample period as the treatment group and employs various PSM methods such as 1:1 nearest neighbor matching, 1:2 nearest neighbor matching, and kernel density matching to perform year-by-year matching for the treatment group. After re-matching, the *p*-values for the differences in sample means between the treatment group and the control group for different variables are all insignificant at the 10% significance level, indicating that the samples of the treatment group and the control group are balanced. On this basis, the difference-in-differences method is reapplied to examine the impact of data factor agglomeration on within-city income disparities. According to Table 4, regardless of the method used, the coefficient of the core explanatory variable remains significant at the 1% significance level, indicating that after correcting for sample selection bias using the PSM-DID method, the baseline regression results of this paper remain stable.

Table 4. PSM-DID Estimation Results.

Variables	(1) 1:1 Gap	(2) 1:2 Gap	(3) Kernel Density Gap
Dec	−0.0155 ***	−0.0163 ***	−0.0151 ***
	(0.0055)	(0.0056)	(0.0055)
Control variables	Yes	Yes	Yes
Constant	0.7260 **	0.6311 **	0.9235 ***
	(0.2923)	(0.2868)	(0.2777)
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
R^2	0.940	0.941	0.941
Observations	3056	2981	3104

Note: *** and ** represent the significance of 1% and 5% levels, respectively.

4.3. Robustness Tests

4.3.1. Parallel Trends Test

This paper employs the difference-in-differences method for estimation. To ensure the validity of the regression results, a key prerequisite is that the parallel trends assumption holds—that is, there is no significant difference in income disparities between cities that established the National Big Data Comprehensive Pilot Zone and those that did not prior to policy implementation, or they share a common trend. Drawing on the method of Jacobson, Lalonde, and Sullivan (1992) [45], this paper adopts an event study approach and constructs the following model:

$$GAP_{it} = \alpha_0 + \alpha_1 pre_{-3} + \cdots + \alpha_1 current + \cdots + \alpha_8 post_5 + \theta Z_{it} + v_i + \mu_i + \varepsilon_{it} \quad (2)$$

Among them, pre_{-3} and $post_5$ represent the three years prior to and the five years after the establishment of the National Big Data Comprehensive Pilot Zone, respectively; current denotes the year of the establishment of the National Big Data Comprehensive Pilot Zone; the

meanings of other variables are consistent with those in Model (1). Based on this, this paper employs Model (2) to test the parallel trend, and the results are presented in Figure 4. The parallel trend test indicates that before policy implementation, the coefficients on the policy dummies are statistically insignificant, suggesting no pre-existing differences in within-city income disparities between treatment and control groups. After implementation, significant differences emerge, satisfying the parallel trends assumption. The dynamic effects show that the pilot zone policy exhibits both a time lag and persistence. The effect is insignificant in the adoption year, gradually becomes significant, and income disparities narrow continuously, although the effect weakens slightly over time. This suggests that complementary policies should be introduced and refined to sustain the long-term effects and consolidate the initial gains.

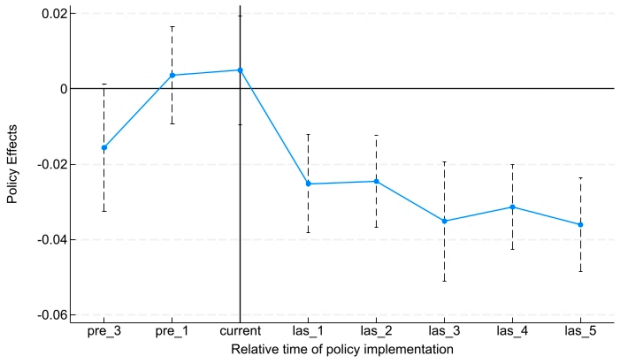


Figure 4. Parallel trend test. The vertical dashed lines represent the 95% confidence intervals.

4.3.2. Sensitivity Analysis

Frontier causal inference theory suggests that conventional parallel trends tests may introduce distortion or bias in the estimator. Therefore, we conduct a sensitivity analysis on the point estimates and confidence intervals by constructing the relative degree of deviation under which the parallel trends test fails and the corresponding confidence intervals of the post-treatment point estimates for different degrees of deviation [46]. If the policy treatment effect remains significant under a relatively large degree of deviation, it indicates that the parallel trends assumption is reasonably robust. Following Biasi and Sarsons (2022) [47], we apply this sensitivity test to the establishment of the National Big Data Comprehensive Pilot Zone. As shown in Figure 5, under both the relative deviation restriction and the smoothness restriction, the parallel trends assumption remains robust, indicating that under a certain degree of shock deviation, data factor agglomeration can narrow within-city income disparities.

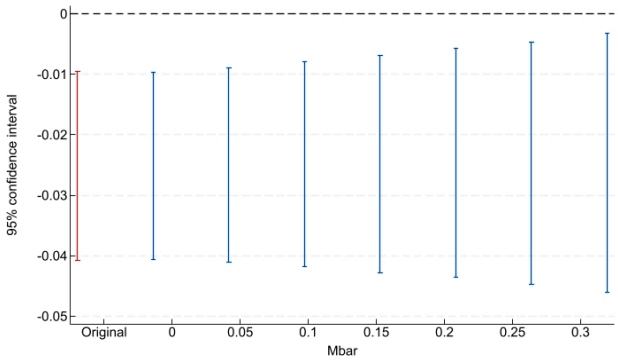


Figure 5. Sensitivity of post-treatment estimates to parallel trend violations. Red vertical line represents the 95% confidence interval of the benchmark estimate; blue vertical lines denote the 95% confidence intervals of placebo estimates under different random assignment parameters.

4.3.3. Placebo Test

In the above test results, the observed relationship between data factor agglomeration and within-city income disparities may potentially reflect a natural response to time trends or be driven by other incidental factors. To rule out this possibility, this paper conducts a placebo test by randomly constructing the data factor agglomeration indicator. Assuming that the timing of the establishment of the National Big Data Comprehensive Pilot Zone remains unchanged and that n cities established the pilot zone in year t, the cities that originally established the pilot zone in the treatment group are designated as the new control group. Meanwhile, n cities are randomly selected from those that had never established the pilot zone in the current year and all previous years to serve as the new treatment group. The estimation is then performed using the new sample, and this process is repeated 500 and 1000 times to conduct simulated regressions. Figure 6 reports the results of the placebo test. It can be observed that the pseudo-estimated coefficients are concentrated around zero and follow a normal distribution, indicating that the probability of the baseline regression results occurring under the placebo condition is extremely small. This largely rules out confounding from unobserved time-varying factors, confirming that the negative relationship identified in the main test is not spurious and provides further support for the validity of our baseline findings.

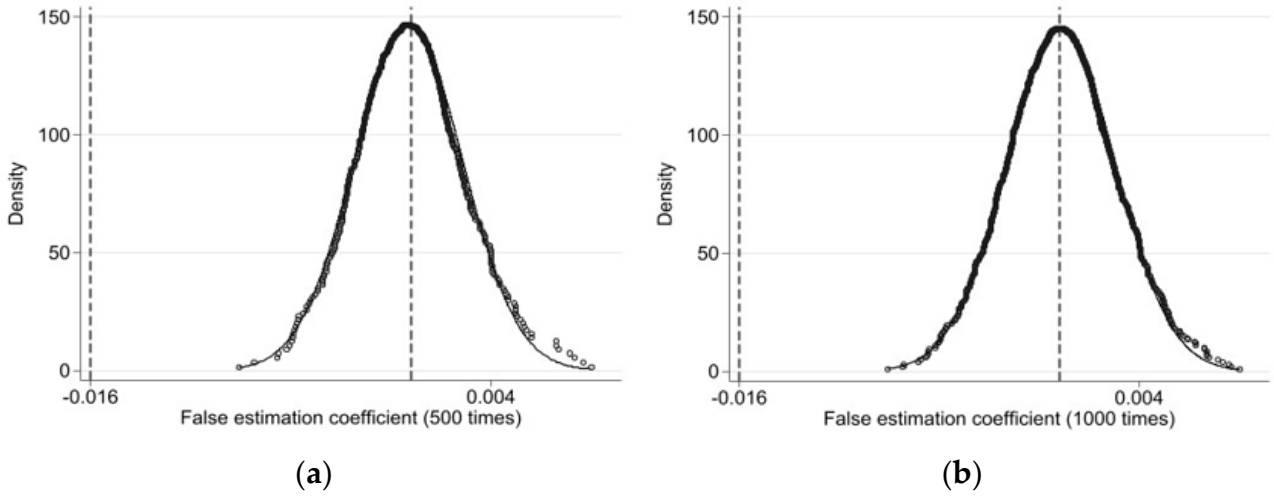


Figure 6. Distribution of placebo estimates. (**a**) Kernel density of estimated coefficients from 500 random assignments; (**b**) Kernel density of estimated coefficients from 1000 random assignments; the vertical dashed line at −0.016 represents the baseline estimate; the vertical dashed line at the distribution peak denotes the mean of placebo coefficients.

4.3.4. Excluding Interference from Other Policies

The cities where the National Big Data Comprehensive Pilot Zone was established exhibit considerable variation in economic development levels. Apart from the exogenous shock of the pilot zone itself, other policies may also influence within-city income disparities, thereby affecting the robustness of the core conclusions of this study. To avoid interference from other policies across regions, this paper collects economic policies related to promoting data factor agglomeration at the prefecture-level city level and focuses on the following three aspects:

- "Smart City" construction. China released three batches of "Smart City" pilot projects from 2012 to 2014. This paper controls for the potential impact of this policy on within-city income disparities by introducing a dummy variable indicating whether a city is a "Smart City" pilot area.

- "Broadband China" construction. China released three batches of lists of "Broadband China" demonstration cities from 2014 to 2016, and this paper controls for the potential impact of this policy by introducing a dummy variable for "Broadband China" demonstration cities.
- "Supply Chain" demonstration cities. China released three batches of demonstration cities and enterprises for supply chain innovation and application from 2021 to 2024. This paper introduces a dummy variable for "Supply Chain" demonstration cities to control for the potential impact of this policy.

Table 5 reports the estimation results after excluding interference from other policies. According to the regression results, after adding the three policy dummy variables, the coefficient of the core explanatory variable remains significantly negative, indicating that the baseline regression results are less affected by shocks from other policies, and data factor agglomeration significantly narrows within-city income disparities. The baseline regression results remain stable.

Table 5. Regression results controlling for concurrent policies.

Variables	(1) Gap	(2) Gap	(3) Gap
Dec	−0.0153 ***	−0.0161 ***	−0.0156 ***
	(0.0054)	(0.0055)	(0.0054)
Sc	0.0245 **		
	(0.0120)		
Bc		−0.0149 ***	
		(0.0053)	
Scda			0.0424 ***
			(0.0097)
Control variables	Yes	Yes	Yes
Constant	0.8253 ***	0.8222 ***	0.9909 ***
	(0.2750)	(0.2753)	(0.2832)
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
R^2	0.940	0.941	0.941
Observations	3190	3168	3168

Note: *** and ** represent the significance of 1% and 5% levels, respectively.

4.3.5. Other Robustness Tests

To further rule out interference from other factors affecting the baseline regression results and to enhance the robustness and reliability of the findings, this paper also conducts the following robustness tests:

Excluding samples. Considering that the sample in this study includes Beijing, Shanghai, Tianjin, and Chongqing, which differ significantly from other cities in terms of economic development level, infrastructure construction, and public service capacity, their special municipal status may introduce certain biases into the regression results. Therefore, this paper excludes the samples of the four municipalities and re-estimates the model. According to Column (1) of Table 6, the coefficient of the core explanatory variable remains significantly negative at the 1% significance level, showing no substantial change compared with the baseline regression results. This indicates that the role of data factor agglomeration in narrowing within-city income disparities is relatively pronounced, and the conclusions of this study are robust.

Table 6. Other robustness testing methods.

Variables	(1) Gap	(2) Gap	(3) Gap	(4) Gap
Dec	−0.0162 ***	−0.0162 ***	−0.0156 **	−0.0156
	(0.0056)	(0.0057)	(0.0061)	(0.0113)
Dig		−2.0217 *		
		(1.1027)		
Control variables	Yes	Yes	Yes	Yes
Constant	0.8389 ***	0.8270 ***	0.8266 ***	0.7975
	(0.2731)	(0.2854)	(0.2701)	(0.5058)
City FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
R^2	0.940	0.941	0.940	0.934
Observations	3146	3075	3190	3190

Note: ***, **, and * represent the significance of 1%, 5%, and 10% levels, respectively.

Controlling for digital infrastructure level. The proportion of keywords related to new digital infrastructure is used as a proxy variable for the level of digital information infrastructure construction, which is then incorporated into the control variables of the baseline regression model for re-estimation. According to the regression results in Column (2) of Table 6, the coefficient of the core explanatory variable is −0.0162, remaining significant at the 1% significance level, indicating that the baseline regression results are consistent and reliable.

The bootstrap method. To assess the sensitivity of the baseline DID estimates to sampling variation, we apply a standard bootstrap procedure that resamples individual observations from the original sample with replacement. The results are reported in Column (3) of Table 6. The results show that the estimated coefficient of the core explanatory variable is significantly negative at the 5% level. This indicates that the narrowing effect of data factor agglomeration on within-city income disparities holds under the bootstrap sampling framework.

Clustering adjustment and wild cluster bootstrap. Because the National Big Data Comprehensive Pilot Zone policy is assigned at the province level, we adjust the clustering of standard errors to the province level. Column (4) of Table 6 reports the province-clustered estimates. The coefficient on the core explanatory variable remains negative but is statistically insignificant at conventional levels. With only 30 clusters, conventional cluster-robust inference may be unreliable. We therefore implement a wild cluster bootstrap at the province level; the wild bootstrap p-value is 0.096, indicating significance at the 10% level. These results indicate that the core coefficient is negative, although its statistical significance is sensitive to the clustering level.

4.4. Heterogeneity Analysis

The preceding sections primarily examine the impact of data factor agglomeration on within-city income disparities based on the full sample. However, these findings may mask differentiated effects across regions due to heterogeneous characteristics. Accordingly, this paper conducts heterogeneity analysis from the following four perspectives.

4.4.1. Heterogeneity in Geographical Location

We examine regional heterogeneity by dividing the sample into eastern, central, and western regions. The results are reported in Table 7. According to the regression results, in the eastern region, the coefficient of the core explanatory variable is significantly positive

at the 10% significance level, whereas in the central and western regions, the coefficient is significantly negative at the 1% and 5% levels, respectively. This indicates that data factor agglomeration widens within-city income disparities in the eastern region while narrowing them in the central and western regions. This heterogeneity may reflect the combined effects of initial regional conditions. In the eastern region, strong industrial bases and abundant high-skilled labor accelerate the concentration of digital industries in the downtown core, where the polarization effect outweighs the diffusion effect, thereby widening the within-city income disparities. In contrast, in the central and western regions, weaker initial infrastructure means that data factors first help bridge the within-city digital divide. Because their traditional industries are largely located in the suburban outskirts, the inclusive application of data disproportionately benefits peripheral areas, thereby narrowing the within-city income disparities.

Table 7. Heterogeneity of Geographic Location.

Variables	(1) Gap	(2) Gap	(3) Gap
Dec	0.0132 *	−0.0370 ***	−0.0153 **
	(0.0075)	(0.0116)	(0.0073)
Control variables	Yes	Yes	Yes
Constant	−0.2297	1.5561 ***	0.9185 **
	(0.3146)	(0.5881)	(0.3954)
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
R^2	0.940	0.926	0.910
Observations	1111	1100	979

Note: ***, **, and * represent the significance of 1%, 5%, and 10% levels, respectively.

4.4.2. Urban Size

We examine heterogeneity along two dimensions: economic scale, which captures resource endowment and market potential, and population scale, which reflects labor force diversity and social complexity. Cities are divided by GDP into small and large economic scale groups; by urban permanent population, they are classified as megacities/super-large cities (≥5 million), large cities (1–5 million), and small/medium-sized cities (<1 million). Results in Table 8 show that data factor agglomeration significantly narrows within-city income disparities in cities with small economic scale, whereas the effect is insignificant in large cities. This likely arises because smaller cities have simpler economic structures, making data empowerment more inclusive across different urban districts; while in larger cities, data and digital industries tend to concentrate in the downtown core and high-value-added sectors, which disproportionately benefits higher-income groups and leaves the outskirts behind. By population size, the effect is significantly negative for large and small/medium-sized cities but insignificant for megacities and super-large cities. Specifically, the small and medium-sized city group contains 13 city samples, the large city group contains 189 city samples, and the megacity and supercity group contains 88 samples. This may be because small/medium cities first bridge the within-city digital divide and allow even spatial diffusion; large cities can effectively spread digital infrastructure to their suburban outskirts via industrial linkages. In megacities and supercities, however, digital infrastructure is already highly developed, so further data factor agglomeration yields limited marginal effects on reducing within-city income disparities, resulting in an insignificant net effect.

Table 8. Heterogeneity of City Size.

Variables	Economic Scale		Population Size		
	Small Gap	Big Gap	Mega and Mega Cities Gap	Large Cities Gap	Small and Medium-Sized Cities Gap
Dec	−0.0239 ***	−0.0030	−0.0010	−0.0140 *	−0.1033 ***
	(0.0090)	(0.0076)	(0.0086)	(0.0075)	(0.0187)
Control variables	Yes	Yes	Yes	Yes	Yes
Constant	0.7464 **	0.2471	−0.1248	0.7247 ***	1.1425
	(0.3317)	(0.3895)	(0.7123)	(0.3327)	(0.7238)
City FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
R^2	0.911	0.947	0.936	0.940	0.938
Observations	1583	1589	968	2079	143

Note: ***, **, and * represent the significance of 1%, 5%, and 10% levels, respectively.

4.4.3. Heterogeneity in Digital Information Infrastructure Development Level

We split the sample by digital information infrastructure development and estimate the model separately. The results in Table 9 show that data factor agglomeration significantly narrows within-city income disparities in low-infrastructure regions, with the coefficient significant at the 5% level, whereas the effect is insignificant in high-infrastructure regions. This contrast may reflect differences in the marginal reach of digital applications. In low-infrastructure regions, the initial within-city digital divide is wider. The injection of data factors helps bridge this gap, and digital applications are more likely to reach the suburban outskirts, yielding larger marginal gains for peripheral development. In high-infrastructure regions, however, digital infrastructure and industries are already well-developed. Further, data factor agglomeration tends to concentrate in high-value-added sectors within the downtown core, yielding limited marginal effects on reducing within-city income disparities. These findings suggest that the distributional effects of digital infrastructure policy deserve attention, particularly regarding how to balance core development with the inclusive development of urban outskirts.

Table 9. Differences in the level of digital basic information construction.

Variables	(1) Gap	(2) Gap
Dec	−0.0205 **	−0.0054
	(0.0084)	(0.0094)
Control variables	Yes	Yes
Constant	1.1802 ***	−0.0241
	(0.3535)	(0.3945)
City FE	Yes	Yes
Year FE	Yes	Yes
R^2	0.930	0.959
Observations	1523	1522

Note: *** and ** represent the significance of 1% and 5% levels, respectively.

4.4.4. Quantile Regression

To explore the heterogeneity in the impact of data factor agglomeration across cities with different levels of income disparity, this paper conducts quantile regression analyses at the 10th, 25th, 50th, 75th, and 90th percentiles of within-city income disparity. The results are reported in Table 10. According to the regression results, data factor agglomeration exhibits differentiated effects across cities with varying levels of within-city income disparities: for the city group at the 10th percentile (cities with the smallest initial gaps), when the National Big Data Comprehensive Pilot Zone is established, data factors and digital industries are more likely to diffuse evenly across urban space. Rather than excessively concentrating in the downtown core, these digital factors spill over to the suburban outskirts, thereby significantly narrowing

the within-city income disparity. For cities at the 90th percentile (cities with extreme initial gaps), data factor agglomeration significantly widens the within-city income disparity. This is likely because, in highly polarized cities, high-value data factors tend to concentrate further in the downtown core, exacerbating the spatial siphon effect and leaving the peripheral outskirts behind. For the middle-quantile city groups (25th, 50th, and 75th percentiles), the estimated effect of data factor agglomeration is small and statistically insignificant. This may reflect the greater structural heterogeneity of cities in this range or the possibility that the marginal effect of data factor agglomeration on within-city spatial distribution is less pronounced when the initial disparity is at a moderate level.

Table 10. Quantile regression results.

Variables	(1) Q_10	(2) Q_25	(3) Q_50	(4) Q_75	(5) Q_90
Dec	−0.0748 *	−0.0320	−0.0193	−0.0020	0.0165 ***
	(0.0442)	(0.0346)	(0.0172)	(0.0102)	(0.0057)
Control variables	Yes	Yes	Yes	Yes	Yes
Constant	−0.2280	0.4160	0.5726	0.4287	0.4349
	(1.2294)	(1.2137)	(0.5520)	(0.3250)	(0.3206)
City FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	3190	3190	3190	3190	3190

Note: *** and * represent the significance of 1% and 10% levels, respectively.

5. Mechanism Testing

Based on the preceding analysis, data factor agglomeration significantly narrows within-city income disparities. To further investigate the underlying influence pathways, this paper constructs the following model for mechanism analysis, drawing on the theoretical analysis above and following the approaches of Imai et al. (2010) [48]:

$$M_{it} = \gamma_0 + \gamma_1 DEC_{it} + +\theta X_{it} + v_i + \mu_i + \varepsilon_{it} \quad (3)$$

Among them, Model (3) introduces a new mechanism variable based on Model (1). M_{it} are the mechanism variables, and the coefficients γ_1 and δ_1 represent how the mechanism variables influence the effect of data factor agglomeration on within-city income disparities. Other variables remain consistent with those in Model (1). The mechanism test results are reported in Table 11.

Table 11. Mechanism of Action Test.

Variables	(1) Techagg	(2) Techdiff
Dec	0.0007 *	0.1608 ***
	(0.0004)	(0.0460)
Control variables	Yes	Yes
Constant	0.1096 ***	2.4125
	(0.0335)	(1.6993)
City FE	Yes	Yes
Year FE	Yes	Yes
R^2	0.391	0.976
Observations	2820	3190

Note: *** and * represent the significance of 1% and 10% levels, respectively.

5.1. Technological Agglomeration Effect

In selecting a proxy variable for technological agglomeration, considering that data factors provide raw materials for the digital industry and that the core competitiveness of digital industry enterprises such as big data, artificial intelligence, and cloud computing

lies in technology, digital industry agglomeration is thus accompanied by technological agglomeration. Using the level of digital industry agglomeration as a proxy variable for technological agglomeration, we conducted an empirical test. The results are reported in Column (1) of Table 11. Data factor agglomeration significantly promotes technological agglomeration. According to our theoretical framework, technological agglomeration may widen within-city income disparities through a polarization effect. This result is consistent with the first leg of Hypothesis 2.

5.2. Technological Diffusion Effect

Measuring the mechanism variable of technological diffusion is challenging, and a unified indicator has yet to be established. Based on existing literature, most scholars currently use invention patents as the primary indicator for measuring technological diffusion [49]. In line with these existing measurement approaches, this paper uses the logarithm of the number of inter-city technological collaborative innovation patents to capture technological diffusion. This value can, to a certain extent, reflect the diffusion effect of scientific and technological achievements from one city to another, with a higher value indicating a higher level of technological diffusion. Based on this, this paper conducts an empirical analysis. Column (2) of Table 11 reports the mechanism test results for technological diffusion. Data factor agglomeration significantly promotes technological diffusion. According to our theoretical framework, technological diffusion may narrow within-city income disparities through a trickle-down effect. This result is consistent with the first leg of Hypothesis 3.

6. Empirical Evidence on Spatial Spillovers

To further examine the spatial spillover effects of data factor agglomeration on within-city income disparities, this paper employs the Spatial Durbin Model (SDM) for global estimation. The specific model specification is as follows:

$$GAP_{it} = \alpha_0 + \rho WGAP_{it} + \beta_1 treat_i * post_t + \gamma W * treat_i * post_t + \theta X_{it} + v_i + \mu_i + \varepsilon_{it} \quad (4)$$

where W is the spatial weight matrix, $WGAP_{it}$ is the spatial lag term of the dependent variable, $W * treat_i * post_t$ is the spatial lag term of the core policy variable, and ρ captures the degree of spatial dependence of within-city income disparities.

Based on this, this paper simultaneously constructs a geographic distance matrix, an economic distance matrix, and an economic-geographic distance matrix to conduct Spatial Durbin Model regressions. The results, presented in Table 12, show that regardless of the spatial weight matrix used, the direct effect of data factor agglomeration is significantly negative, indicating that data factor agglomeration can significantly narrow local within-city income disparities. The spatial autocorrelation coefficients are all significantly positive at the 1% level, indicating a strong spatial dependence of within-city income disparities across cities. However, the average spatial spillover effect of data factor agglomeration from neighboring cities on a city's internal income disparities is not significant, and the total effects are also insignificant.

Considering that the global spatial spillover effect may be insignificant because it masks the non-linear spatial heterogeneity, this paper further draws on Cao (2020) [50] and adopts a distance-heterogeneity-aware difference-in-differences approach (spillover-robust DID) to examine the nonlinear spatial impact of data factor agglomeration on income disparities in neighboring regions. The model is specified as follows:

$$GAP_{it} = \beta_0 + \beta_1 treat_i * post_t + \sum_{s=100}^{600} \delta_s N_{it}^s + \theta X_{it} + v_i + \mu_i + \varepsilon_{it} \quad (5)$$

Table 12. Spatial Durbin Model Regression Results.

Variables	(1) Geographic Distance Matrix	(2) Economic Distance Matrix	(3) Economic-Geographic Distance Matrix
Dec	−0.0142 **	−0.0125 **	−0.0143 **
	(0.0056)	(0.0059)	(0.0056)
W_ Dec	0.0120	0.0147	0.0208
	(0.0122)	(0.0097)	(0.0131)
Direct effect	−0.0143 **	−0.0116 *	−0.0139 **
	(0.0057)	(0.0065)	(0.0057)
Indirect effect	−0.0386	0.0217	0.1068
	(0.2130)	(0.0450)	(0.1938)
Total effect	−0.0529	0.0101	0.0929
	(0.2141)	(0.0484)	(0.1944)
ρ	0.9421 ***	0.8040 ***	0.9351 ***
	(0.0153)	(0.0207)	(0.0168)
σ^2	0.0031 ***	0.0036 ***	0.0031 ***
	(0.0002)	(0.0002)	(0.0002)
Log-likelihood	4084.2693	3680.5057	4081.1503
Control variables	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
R^2	0.5627	0.4940	0.5475
Observations	3190	3190	3190

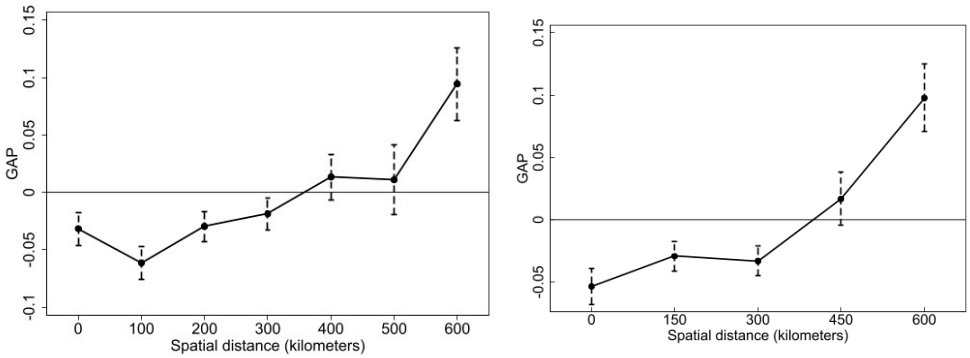
Note: ***, **, and * represent the significance of 1%, 5%, and 10% levels, respectively.

Model (5) builds upon the baseline regression model (1) by introducing a new set of control variables, N_{it}^{s}, relaxing the "stable unit treatment value assumption"(SUTVA) and requiring only partial satisfaction of the SUTVA to estimate the policy treatment effect and the "neighbor" treatment effect. The parameter s denotes the geographic distance (in kilometers) between each city and the National Big Data Comprehensive Pilot Zone. If city i is located within 100 km of a National Big Data Comprehensive Pilot Zone, N_{it}^{s} takes a value of 1; otherwise, it takes a value of 0. The coefficient δ_s is used to examine the impact of the establishment of the Big Data Comprehensive Pilot Zone on the within-city income disparities of cities at varying distances. In the specific regression analysis, this study adopts 100 km as the unit and reports the regression results for $s = 100, 200, \ldots, 500, 600$, respectively. By comparing the economic and statistical significance of the coefficient δ_s under different thresholds, this study tests the spatial heterogeneity of data factor agglomeration, using the National Big Data Comprehensive Pilot Zone as a case study.

Based on the estimation using Model (5), Figure 7 presents the trend of how the impact of data factor agglomeration on within-city income disparities varies with the geographic distance between each city and the National Big Data Comprehensive Pilot Zone. According to the estimation results, when the distance to the pilot zone is within 300 km, the establishment of the pilot zone narrows the internal income disparities of surrounding cities, with the narrowing effect first increasing and then decreasing as distance grows. When the distance reaches 400 km, the effect on their internal income disparities becomes insignificant. When the distance extends to 600 km, the effect reverses and becomes significantly positive, indicating a widening of within-city income disparities. To ensure that the spatial thresholds are not driven by a particular bandwidth choice, this paper further re-estimates the model using 150 km, 300 km, 450 km, and 600 km as alternative bandwidths, and the results remain stable.

This distance-dependent pattern reflects the interplay of three mechanisms: transmission losses in data radiation, the boundary constraints of industrial linkages, and friction costs in policy coordination. Within 300 km, the digital infrastructure of the pilot zones radiates data resources and technologies to neighboring cities. Close industrial ties and lower coordination costs enable these cities to integrate into the pilot zone's digital industrial chain, which drives the transformation and upgrading of local traditional industries,

thereby narrowing their within-city income disparities. Beyond 300 km, increased data latency, rising institutional costs, and declining industrial compatibility erode the radiation effect. By 400 km, neighboring cities can no longer effectively absorb the digital dividends from the pilot zones, leaving their internal income disparities largely unaffected. When the distance reaches 600 km, the strong agglomeration capacity of the pilot zones generates a significant siphon effect, drawing talent and capital away from distant cities. This depletes local resources, suppresses resident income growth, and triggers industrial hollowing-out and labor outflows in their peripheral areas, ultimately widening their within-city income disparities. These results support Hypothesis 4, confirming that data factor agglomeration exerts a distance-dependent spatial spillover effect on within-city income disparities.



(**a**) (**b**)

Figure 7. Spatial spillover effect of data element aggregation. (**a**) Baseline distance-band estimates; (**b**) Estimates with alternative bandwidths of 150 km, 300 km, 450 km, and 600 km for bandwidth sensitivity check; vertical dashed lines denote 95% confidence intervals.

7. Conclusions and Discussions

7.1. Main Findings

Based on panel data from 290 prefecture-level cities in China from 2013 to 2023, this paper uses the National Big Data Comprehensive Pilot Zone as a proxy variable for data factor agglomeration and employs a difference-in-differences (DID) model to empirically examine the intrinsic relationship between data factor agglomeration and within-city income disparities. The findings are as follows: First, data factor agglomeration, represented by the National Big Data Comprehensive Pilot Zone, significantly narrows within-city income disparities. However, due to the limited number of clusters at the province level, the statistical significance of this effect is sensitive to the clustering level. While the estimate remains statistically significant under baseline city-level clustering, its significance weakens when standard errors are clustered at the more conservative province level. Second, heterogeneity analysis reveals significant spatial and structural differences in the impact of data factor agglomeration on within-city income disparities. Specifically, in the eastern region, data factor agglomeration further widens these internal disparities, whereas in the central and western regions and areas with lower levels of digital information infrastructure, it helps narrow them. From the perspective of city size, data factor agglomeration alleviates internal income disparities in cities with smaller economic scales, as well as in large, medium, and small cities. Furthermore, quantile regression results indicate that at the 10th percentile, data factor agglomeration narrows within-city income disparities, whereas at the 90th percentile, it exacerbates them. Third, from a spatial perspective, data factor agglomeration, represented by the National Big Data Comprehensive Pilot Zone, exhibits significant distance-dependent spatial spillover effects on within-city income disparities. This is mainly reflected in the pilot zone's ability to significantly narrow the internal income gaps of cities within 300 km,

while exerting a siphon effect that widens the internal gaps of cities beyond 600 km. This indicates that the pilot zone promotes inclusive urban development within its effective range, contributing to intra-city spatial coordination. Fourth, mechanism tests indicate that data factor agglomeration significantly enhances both technological agglomeration and technological diffusion, consistent with our theoretical prediction that technological agglomeration tends to widen within-city income disparities while technological diffusion tends to narrow them.

7.2. Policy Implications

Based on the above findings, this paper proposes the following policy implications:

First, promote the orderly agglomeration of data factors while strengthening spatial guidance. The direction of agglomeration should be clearly defined to avoid homogeneous and inefficient clustering. Through differentiated positioning of the National Big Data Comprehensive Pilot Zones, data resources should be guided not only toward regions with weak digital infrastructure but, crucially, toward the peripheral outskirts of cities that urgently need digital transformation. This approach helps prevent the excessive concentration of high-value-added digital industries exclusively in downtown cores, while also preventing resource waste across regions. Supporting policies, including talent attraction and public service provision, should be improved to create a favorable environment. Meanwhile, antitrust and regulatory measures should be utilized to prevent the monopolization of data factors by a few regions or firms, ensuring that digital dividends are widely shared across both regions and intra-city spaces.

Second, open up channels for factor circulation to strengthen cross-city spatial spillovers. Since data factor agglomeration generates significant spatial spillovers within a specific radius, policies should facilitate cross-regional data flows and value realization. This requires breaking down administrative barriers and data silos by leveraging the national unified data market and establishing cross-regional cooperation mechanisms. Furthermore, it requires reducing technological barriers by optimizing cross-regional computing power dispatch networks and enhancing digital infrastructure connectivity. By extending the technological outcomes of the pilot zones to surrounding cities, these policies can further drive the transformation of their suburban industries, enabling peripheral areas to share the digital dividend and fostering intra-city spatial coordination.

Third, adopt a differentiated digital infrastructure strategy that balances high-end development with spatial inclusiveness. Although technological agglomeration may initially widen within-city spatial disparities, it is a prerequisite for subsequent technological diffusion. Downtown core areas, especially the pilot zones and eastern digital economy hubs, should accelerate the construction of intelligent computing centers and data trading platforms to strengthen high-end carriers for technological breakthroughs. For urban outskirts with weak digital infrastructure, inclusive deployment should be prioritized through fiscal subsidies and inter-city assistance. This includes expanding edge computing nodes and small-scale data centers in suburban areas, which will open up the hardware channels for data factors to reach lagging peripheral areas. This supports the long-term mechanism of "agglomeration in the core first, diffusion to the outskirts later," ultimately enabling data factors to contribute to narrowing within-city income disparities.

7.3. Discussion

This paper investigates the causal effect of data factor agglomeration on within-city income disparities, using the establishment of China's National Big Data Comprehensive Pilot Zones as a quasi-natural experiment. Our findings make several contributions to the existing literature and raise important theoretical and policy considerations.

Compared with existing studies, the innovations of this paper are as follows:

- The reversal of the polarization prediction. Contrary to classical agglomeration theory, which predicts that agglomeration deepens core-periphery disparities, we find that under China's policy-guided data environment, data factor agglomeration significantly narrows within-city income disparities on average and generates spatial income gains within a 300 km radius. This counter-intuitive result arises from policy-induced efficiency gains in urban outskirts and enhanced technological diffusion, which counteract the market-driven polarization tendency.
- The spatial decomposition of technological progress. Instead of treating technological progress as a homogeneous mediator, we spatially decompose it into technological agglomeration and technological diffusion. We show that data factor agglomeration promotes both technological agglomeration and technological diffusion, thereby providing important evidence for the proposition that data factor agglomeration widens within-city income disparities through a polarization effect and narrows them through a trickle-down effect.
- The theoretical threshold for policy-induced convergence. Within an integrated core–periphery framework, we incorporate data factors, technological agglomeration, and technological diffusion into a unified analysis and derive the critical condition under which data factor agglomeration shifts from polarization to convergence. This condition indicates that when the efficiency gains in peripheral outskirts exceed those in the core region and the intensity of technological diffusion surpasses a threshold, the net effect of data factor agglomeration may reverse. This result provides a theoretical reference for understanding the effectiveness of policy interventions and offers testable hypotheses for further empirical examination.

Several caveats merit attention. First, our proxy for data factor agglomeration—the National Big Data Comprehensive Pilot Zone—is a policy designation rather than a direct measure of actual data flows or data stock. While we take care to address endogeneity concerns, future work could construct more granular measures of data transactions, data center capacity, or cross-regional data flows to refine the identification. Second, although we utilize high-resolution grid data to measure within-city spatial disparities, our analysis primarily captures the spatial divide between urban centers and outskirts, rather than income inequalities at the individual or household level. Future micro-level studies using household and firm survey data would complement our findings by revealing how data agglomeration affects specific population groups within the city. Third, our identification of technological diffusion relies on spatial econometric techniques that are subject to their own assumptions; alternative approaches such as structural estimation or network-based diffusion models could provide additional robustness. Fourth, the general equilibrium effects of data agglomeration—including potential impacts on labor markets, housing costs, and regional fiscal capacity—are not fully explored in this paper and warrant dedicated investigation. Finally, the transferability of our conclusions to other national contexts with different data governance regimes remains an open question; comparative studies across countries would be valuable for delineating the boundary conditions of the convergence effect we document.

In summary, this paper contributes to the nascent literature on the economics of data as a factor of production by demonstrating that, under China's policy-guided data environment, the spatial agglomeration of data can serve as a force for spatial convergence within cities rather than divergence. The dual mechanisms of technological agglomeration and diffusion, and the theoretical threshold at which one dominates the other, offer a richer understanding of how data-driven growth can be harnessed to reduce within-city spatial inequalities

and promote inclusive urban development. We hope that these findings stimulate further theoretical and empirical work on the distributional consequences of the data economy.

Author Contributions: Conceptualization, Y.A.; methodology, Y.A.; software, J.K. and J.L.; validation, Y.A. and J.Z.; formal analysis, Y.A.; investigation, J.K.; resources, J.Z.; data curation, J.Z. and J.L.; writing—original draft preparation, Y.A.; writing—review and editing, Z.C.; visualization, J.Z.; supervision, Z.C.; project administration, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Gansu Provincial Science and Technology Program—Soft Science Special Project, China, grant number 26JRZA130 and Lanzhou University of Technology's Humanities and Social Sciences Cultivation Fund Project for 2025, grant number 062507.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request from the authors.

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chu, Z.; Chen, X.; Yang, J. Impact of data factor and data integration on economic development: Empirical insights from China. *Telecommun. Policy* **2025**, *49*, 103004. [CrossRef]
2. Kraft, C. Big data analytics in management reporting: A systematic literature review. In *Applied Economic Research and Trends*; Tsounis, N., Vlachvei, A., Eds.; Springer: Cham, Switzerland, 2024; pp. 537–558. [CrossRef]
3. Schembera, B. Data, Information, Knowledge, Wisdom, and Explainable Artificial Intelligence. In *Designing the Conceptual Landscape for a XAIR Validation Infrastructure*; Springer: Cham, Switzerland, 2025; pp. 122–132. [CrossRef]
4. Jones, C.I.; Christopher, T. Nonrivalry and the Economics of Data. *Am. Econ. Rev.* **2020**, *110*, 2819–5810. [CrossRef]
5. Farboodi, M.; Veldkamp, L. *A Model of the Data Economy*; NBER Working Paper; NBER: Cambridge, MA, USA, 2021; p. 28427. [CrossRef]
6. Veldkamp, L.; Chung, C. Data and the Aggregate Economy. *J. Econ. Lit.* **2024**, *62*, 458–484. [CrossRef]
7. Qi, P.; Sun, D.; Xu, C.; Li, Q.; Wang, Q. Can Data Elements Promote the High-Quality Development of China's Economy? *Sustainability* **2023**, *15*, 7287. [CrossRef]
8. Goldfarb, A.; Tucker, C.E. Digital Economics. *J. Econ. Lit.* **2019**, *57*, 3–43. [CrossRef]
9. Sen, A.; Chen, H.; Grossman, M.R.; Tsao, S. A Welfare Test for Sharing Health Data. *Rev. Income Wealth* **2024**, *71*, e12717. [CrossRef]
10. Okorie, O.; Charnley, F.; Russell, J.D.; Tiwari, A.; Moreno, M. Circular business models in high value manufacturing: Five industry cases to bridge theory and practice. *Bus. Strategy Environ.* **2021**, *30*, 1780–1802. [CrossRef]
11. Vrana, J.; Singh, R. Cyber-Physical Loops as Drivers of Value Creation in NDE 4.0. *J. Nondestruct. Eval.* **2021**, *40*, 61. [CrossRef] [PubMed]
12. He, Y.; Wen, T. Study on the influence mechanism of digital economy on agricultural carbon emission reduction-based on the perspective of fertilizer saving and efficiency improvement. *Front. Sustain. Food Syst.* **2025**, *9*, 1519109. [CrossRef]
13. Liu, J.; Shi, Z. The Impact of Data Element Agglomeration on Inclusive Green Development: Evidence from Threshold and Spatial Spillover Effects. *Sustainability* **2026**, *18*, 2973. [CrossRef]
14. Wang, H.; Hao, Y.; Fu, Q. Data factor agglomeration and urban green finance: A quasi-natural experiment based on the National Big Data Comprehensive Pilot Zone. *Int. Rev. Financ. Anal.* **2024**, *96*, 103732. [CrossRef]
15. Muñoz, R.; Ríos, S.A.; Graña, M. Clustering Cities over Features Extracted from Multiple Virtual Sensors Measuring Micro-Level Activity Patterns Allows One to Discriminate Large-Scale City Characteristics. *Sensors* **2023**, *23*, 5165. [CrossRef] [PubMed]
16. Shen, N.; Zhou, J.; Zhang, G.; Wu, L.; Zhang, L. How does data factor marketization influence urban carbon emission efficiency? A new method based on double machine learning. *Sustain. Cities Soc.* **2024**, *119*, 106106. [CrossRef]
17. Xing, Z.; He, C.; Lin, J.; Pan, Y. Who you are versus where you are: Revealing the importance of determinants of within-city income inequality in China through an interpretable machine learning approach. *Appl. Geogr.* **2025**, *184*, 103759. [CrossRef]
18. Cui, C.; Pang, X.; Wang, W. The impact of de-industrialization on the urban-rural income gap: Evidence from China. *Habitat Int.* **2026**, *177*, 103970. [CrossRef]

19. Xiao-hui, Z.; Zheng, T. Does polycentric development narrow intra-urban income inequality? An empirical examination based on mediation and threshold effects. *Environ. Plan. B Urban Anal. City Sci.* **2025**, *53*, 1484–1502. [CrossRef]
20. Acemoglu, D.; Restrepo, P. Automation and new tasks: How technology displaces and reinstates labor. *J. Econ. Perspect.* **2019**, *33*, 3–30. [CrossRef]
21. Shao, Y. Does the widespread of mobile internet narrow the urban–rural labor income gap? Evidence from China. *Econ. Anal. Policy* **2026**, *94*, 1800–1816. [CrossRef]
22. Liang, K.; Liu, W. Digital infrastructure and urban-rural income gap: Empirical evidence from China. *Int. Rev. Econ. Financ.* **2026**, *107*, 105003. [CrossRef]
23. Chen, D.; Ma, Y. Effect of industrial structure on urban–rural income inequality in China. *China Agric. Econ. Rev.* **2022**, *14*, 547–566. [CrossRef]
24. Guo, Q. Analysis on the relationship between regional innovation and income inequality in Chinese city regions. *Prof. Geogr.* **2019**, *71*, 472–490. [CrossRef]
25. Chen, Y.; Mao, F.; Wu, Y. Intercity Digital Innovation Network Embeddedness and Within-City Income Inequality. *Int. Rev. Econ. Financ.* **2026**, *112*, 105932. [CrossRef]
26. Yu, Y.; Du, D.; Li, Q. Characteristics and Influencing Factors of Provincial High-End Manufacturing Innovation Clusters in China: A Big Data Analysis of Technology-Based Enterprises. *Appl. Spat. Anal. Policy* **2025**, *18*, 41. [CrossRef]
27. Jiang, S. Big Data Sharing: A Comprehensive Survey. *Data* **2025**, *10*, 182. [CrossRef]
28. Gross, D.P.; Sampat, B.N. America, Jump-Started: World War II R&D and the Takeoff of the US Innovation System. *Am. Econ. Rev.* **2023**, *113*, 3323–3356. [CrossRef]
29. Huang, R.; Mao, S. Enhancing the energy level of regional digital innovation ecosystems: A configuration perspective. *PLoS ONE* **2024**, *19*, e0309229. [CrossRef] [PubMed]
30. Kosfeld, R.; Mitze, T. Research and development intensive clusters and regional competitiveness. *Growth Change* **2023**, *54*, 885–911. [CrossRef]
31. Ayerst, S.; Ibrahim, F.; MacKenzie, G.; Rachapalli, S. Trade and Diffusion of Embodied Technology: An Empirical Analysis. *J. Monet. Econ.* **2023**, *137*, 128–145. [CrossRef]
32. Luo, K.; Lee, C.-C.; Zhuo, C. A pathway to coordinated regional development: Energy utilization efficiency and green development—Evidence from China's major national strategic zones. *Energy Econ.* **2024**, *131*, 107402. [CrossRef]
33. Niño-Zarazúa, M.; Roope, L.; Tarp, F. Global Inequality: Relatively Lower, Absolutely Higher. *Rev. Income Wealth* **2016**, *63*, 661–684. [CrossRef]
34. Wang, L.; Li, T.; Peng, X.; Liu, R. Can the integration of cultural and tourism development narrow the regional income gap? The role of foreign direct investment. *Int. Rev. Econ. Financ.* **2025**, *102*, 104–294. [CrossRef]
35. Teng, C.; Jiao, X.; Jin, Z.; Wang, C. Study on the Temporal and Spatial Evolution of Market Integration and Influencing Factors in the Yellow River Basin. *Sustainability* **2025**, *17*, 6920. [CrossRef]
36. Mveyange, A. *Night Lights and Regional Income Inequality in Africa*; WIDER Working Paper; The United Nations University World Institute for Development Economics Research (UNU-WIDER): Helsinki, Finland, 2015; Volume 85. [CrossRef]
37. Lessmann, C.; Seidel, A. Regional Inequality, Convergence, and Its Determinants—A View from Outer Space. *Eur. Econ. Rev.* **2017**, *92*, 110–132. [CrossRef]
38. Galimberti, J.K.; Pichler, S.; Pleninger, R. Measuring Inequality Using Geospatial Data. *World Bank Econ. Rev.* **2023**, *37*, 549–569. [CrossRef]
39. Niu, C.; Qi, W.; Liu, H. How does artificial intelligence affect the carbon emission efficiency of resource-based cities? A nonlinear effect based on regional income inequality. *Energy* **2026**, *36*, 141312. [CrossRef]
40. Wang, Y.; Liu, J.; Wang, Y. The Impact Mechanism and Effect Evaluation of the National Big Data Comprehensive Pilot Zone on the Resilience of Manufacturing Enterprises. *Sustainability* **2026**, *18*, 1505. [CrossRef]
41. Liu, M.; Li, G. Research on the Impact of the Digital Economy on Carbon Pollution Based on the National Big Data Comprehensive Pilot Zone in China. *Sustainability* **2023**, *15*, 15390. [CrossRef]
42. Beck, T.; Levine, R.; Levkov, A. Big Bad Banks? The Winners and Losers from Bank Deregulation in the United States. *J. Financ.* **2010**, *65*, 1637–1667. [CrossRef]
43. Wang, J. The economic impact of special economic zones: Evidence from Chinese municipalities. *J. Dev. Econ.* **2013**, *101*, 133–147. [CrossRef]
44. Huang, J.; Wang, Y.; Luan, B.; Zou, H.; Wang, J. The energy intensity reduction effect of developing digital economy: Theory and empirical evidence from China. *Energy Econ.* **2023**, *128*, 107–193. [CrossRef]
45. Jacobson, L.S.; Lalonde, R.J.; Sullivan, D. Earnings losses of displaced workers. *Upjohn Work. Pap. J. Artic.* **1992**, *83*, 685–709. [CrossRef]
46. Roth, J.; Sant'Anna, P.H.C. When Is Parallel Trends Sensitive to Functional Form? *Econometrica* **2023**, *91*, 737–747. [CrossRef]

47. Biasi, B.; Sarsons, H. *Flexible Wages, Bargaining, and the Gender Gap*; NBER Working Paper; NBER: Cambridge, MA, USA, 2020; p. 27894. [CrossRef]
48. Imai, K.; Keele, L.; Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **2010**, *15*, 309–334. [CrossRef] [PubMed]
49. He, J.; Tian, X. The dark side of analyst coverage: The case of innovation. *J. Financ. Econ.* **2013**, *109*, 856–878. [CrossRef]
50. Cao, Q.F. The Driving Effect of National-level New Areas on Regional Economic Growth: Empirical Evidence from 70 Large and Medium-sized Cities. *China Ind. Econ.* **2020**, *7*, 43–60. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.