Industry Agglomeration and Firm Productivity in China: Does Highway Access Matter?

Minjia Guo

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN ECONOMICS

ADAM SMITH BUSINESS SCHOOL

COLLEGE OF SOCIAL SCIENCE

SEPTEMBER 2024

Abstract

Using detailed GIS and firm-level data, this thesis investigates the effect of highway access on within-industry agglomeration, coagglomeration and firm productivity in China during the period 1998-2007. In order to address the potential endogeneity problem and shed light on causality, this study constructs three types of time-variant instruments for the highway access variable, including the historical routes, least cost path network and straight line network.

This study finds that highway access positively affects within-industry agglomeration, and the results are consistent with IV estimations. The improved highway access enhances agglomeration for downstream industries by increasing their flexibility in location choices. Additionally, input-output adjusted highway access promotes within-industry agglomeration by lowering transportation costs for accessing inputs and outputs from other industries.

Regarding the effects of highway access on coagglomeration of industry pairs, this study finds that better highway access increases coagglomeration at the province and city levels, but the effect is insignificant at the county level. The positive effect of highways on coagglomeration is larger for industry pairs with a lower share of state-owned enterprises, for related industries, and for those with input-output linkages.

The results indicate that highway access has a positive effect on firm productivity, with a 1% change in highway access yielding a 0.018% increase in firm TFP. Four channels are investigated, including within-industry agglomeration, coagglomeration, export, and innovation. This study finds that highway access affects firm productivity through the channels of within-industry agglomeration, coagglomeration, and exports, while the innovation channel is less significant.

Contents

List of Tables

List of Figures

Abbreviations

ACASIAN	Australian Consortium for the Asian Spatial Information and Analysis Network
ACF correction	Ackerberg–Caves–Frazer correction
ASIF	Annual Survey of Industrial Firms
DO index	Duranton-Overman index
EG index	Ellison-Glaeser index
FDI	Foreign Direct Investment
GIS	Geographic Information System
IO table	Input-Output table
IOHA	Input-Output adjusted Highway Access
LCP	Least Cost Path
LCP-MST	Least Cost Path Minimum Spanning Tree
LP method	Levinsohn-Petrin method
MAR	Marshall-Arrow-Romer
MSA	Metropolitan Statistical Areas
NBER	National Bureau of Economic Research
NBS	National Bureau of Statistics of China
NEN	National Expressway Network
NTHS	National Trunk Highway System
OP method	Olley-Pakes Method
SIPO	State Intellectual Property Office
SOEs	State-Owned Enterprises
TFP	Total Factor Productivity
USPTO	U.S. Patent and Trademark Office

Acknowledgements

I extend my sincere gratitude to Prof. Sai Ding and Dr. Hisayuki Yoshimoto, my supervisors, for all the guidance and inspiration they have given me in the course of this academic journey. Their positive influence and valuable insights, and the opportunities they provided me with have greatly shaped both my research and personal development.

Prof. Ding's encouragement and optimistic outlook have not only enlightened my academic path but also enriched my attitude towards life. I am grateful for the opportunity to have been a research assistant under Prof. Ding's mentorship, which has broadened my research horizons. Dr. Yoshimoto's guidance has taught me to prioritize key issues, tackle tasks gradually, and think deeply. I am thankful for his valuable advice, which has influenced my approach to understanding and addressing different situations.

I would also like to express my deepest gratitude to my family for supporting me throughout this journey. I am grateful to my parents for their unwavering support, which has provided me with the chance to focus on my studies. I would like to thank my wonderful husband whose willingness to partake in insightful discussions across various topics, coupled with his intellectual thinking, has significantly contributed to the depth of my understanding.

I am grateful for my dear friends who provided companionship, trust, and joy throughout this journey. Their presence has been a source of strength and happiness, creating cherished memories together. Last but not least, I extend my appreciation to the supportive staff of the Adam Smith Business School, who have created a fantastic learning environment, which I feel fortunate to have been a part of.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Printed Name: Minjia Guo

Signature: Minjia Guo

Chapter 1

Introduction

China has been undergoing increasing agglomeration since the economic reform in 1978, characterized by the clustering of industries in specific regions. Evidence in the literature indicates that Chinese industry experienced a rise in spatial concentration between 1980 and 1995 (Wen 2004) and also a consistent upward trend from 1998 to 2005 (Lu & Tao 2009). China's industry agglomeration is particularly pronounced in the eastern coastal areas, such as the Pearl River Delta (e.g., Shenzhen, Guangzhou) and Yangtze River Delta (e.g., Shanghai, Jiangsu), and the Bohai Bay region (e.g., Beijing, Tianjin), which have developed into major economic hubs. Furthermore, China's western and central regions are progressively becoming new concentrated areas, benefiting from government initiatives to promote balanced regional development.

Based on the concept of Ellison & Glaeser (1997), the term 'within-industry agglomeration' refers to the specialized concentration and 'coagglomeration' means the spatial concentration of different industries. Increasing spatial concentration fosters industrial development by attracting foreign direct investment (Barrell & Pain 1999, Guimaraes et al. 2000) and facilitating innovation (Feldman 1999, Antonietti & Cainelli 2011, Zhang 2015, Connell et al. 2014). There is plenty of evidence that both within-industry agglomeration and coagglomeration have significant impacts on region-level and firm-level productivity (Ciccone & Hall 1996, Ciccone 2002, Lin et al. 2011, Hu et al. 2015, Tokunaga & Kageyama 2008, Barrios et al. 2006).

The theories of Krugman (1991) and Weber (1909) indicate that transport cost is a crucial factor that affects agglomeration. The empirical research also stresses the importance of transport cost and infrastructure on agglomeration (Rosenthal & Strange 2001, Holl 2004) and productivity (Holl 2016). Firms can benefit from agglomeration externalities and choose to co-locate. Upgrading transport infrastructure can improve economic efficiency by reducing the requirement to relocate together. China has made significant advancements in expanding the highway system, resulting in a highway length of 177,000 km by the end of 2022 compared with merely 100km in 1988. The enormous network, which exceeded the American Interstate Highway System in 2014, reflects a sustained commitment to infrastructure development in China. The ambitious National Expressway Network in China, initiated in the 1980s and progressing through upgrading construction phases with the 7-5 plan, 7-9-18 plan, and 7-11-18 plan, has connected major cities, ports, and transportation hubs.

This highway expansion, related to transport time reduction, is expected to foster economic development. The relationship between highway expansion, agglomeration, and productivity plays a pivotal role in shaping regional economic landscapes. Understanding how highways contribute to agglomeration effects and, subsequently, affect productivity, is crucial for policy-making and urban planning, especially for big countries like China. China's agglomeration effect, resulting from its extensive scale and varied economic activity, is likely to significantly contribute to the nation's total economic growth.

China's large size can amplify the economic externalities generated by agglomeration, including knowledge spillovers, labour market pooling and input-output sharing (Marshall 1890). In this way, highway expansion may in turn significantly influence firm productivity. Therefore, this thesis is motivated by these questions: To what extent has highway expansion contributed to industry agglomeration? Do the effects on within-industry agglomeration and coagglomeration differ? Does highway expansion increase firm productivity in China? How has the expansion of highways affected firm productivity? Has it occurred through the channels of within-industry agglomeration, coagglomeration, exporting, or innovation?

This thesis contributes to the literature in several ways. First, it investigates the effects of highways on the level of within-industry agglomeration and pairwise coagglomeration. The determinants of within-industry agglomeration and coagglomeration focus mainly on the national advantages and Marshallian externalities (Ellison et al. 2010, Ellison & Glaeser 1999*a*, Rosenthal & Strange 2001, Jofre-Monseny et al. 2011, Faggio et al. 2017, Diodato et al. 2018, Howard et al. 2016, Mukim 2015). Transport costs are found to affect spatial concentration in China (Wen 2004) and coagglomeration at the metropolitan level in the US (Gallagher 2013). This study is the first in the field of rapidly developing highway construction and industry agglomeration in China.

Second, this study examines the impact of highway access on the productivity of Chinese firms and how it affects productivity. The mechanisms encompass intra-industry agglomeration, coagglomeration, export, and innovation. These mechanisms expose the underlying impacts of highways on firm productivity and have not previously been specifically examined. Agglomeration has a crucial role in enhancing company productivity by facilitating resource sharing, knowledge spillovers, and labour market pooling. Enhancements in transport infrastructure can promote industry agglomeration, leading to a rise in firm productivity. Furthermore, it is worthwhile specifically analyzing within-industry agglomeration and coagglomeration as channels. Although the two notions are distinct, the current literature employs vague indicators to measure agglomeration when studying these channels, mostly relying on local density (Wan & Zhang 2018, Holl 2016). Export is also an important channel, as improved highway access results in more exporting activities, thereby enhancing productivity through the learning-by-exporting effect of some Chinese firms.

Third, this study significantly advances the literature by introducing a novel metric, the Input-Output adjusted Highway Access (IOHA), which better captures the varying use of highways by different industries. Previous measures often overlooked the variations in industry size and transportation volumes, failing to account for how these factors influence highway utilisation. By incorporating input-output tables, the IOHA metric captures both the transportation volume and proximity to highways. This innovative approach addresses a critical gap in existing research and enhances our understanding of how improvements in infrastructure, impact industrial agglomeration and coagglomeration by accounting for the diverse needs of different industries.

Fourth, this study differs from previous literature that relies on aggregate measures of transport infrastructure, such as infrastructure investment or density (Wan & Zhang 2018). Instead, we utilize extensive microeconomic data, which allows for a more thorough investigation of the research subject. For example, the Geographic Information System (GIS) highway data, which accurately represents the locations of highways, together with the location data of firms, are utilized to construct the firm-level highway access variable. Microeconomic firm-level data and customs data are utilized to analyse various mechanisms and heterogeneous characteristics, enhancing the credibility of the findings.

Moreover, in order to address the issue of potential endogeneity and investigate the causal effect, this study employs time-varying instruments related to the highway measure. These instruments include historical ones derived from the courier routes of the Ming and Qing Dynasties, as well as the least cost paths and straight-line routes based on the specific city points in the highway construction plan.

By utilizing GIS data on road networks and a comprehensive dataset of Chinese manufacturing companies from 1998 to 2007, this study reveals that highway access has a beneficial impact on within-industry agglomeration, as indicated by both the baseline model and the instrumental variable estimations. The findings reveal that improved highway access reduces the impact of petroleum on within-industry agglomeration at provincial and city levels but increases it at the county level, highlighting the complex interplay between transportation costs and industry location decisions. Additionally, better highway access enhances agglomeration for downstream industries by improving their flexibility in location optimisation. Overall, the input-output adjusted highway access significantly promotes within-industry agglomeration by reducing firms' transportation costs for inputs and outputs from other industries.

Regarding the effects of highways on coagglomeration, highway access leads to an increase in the pairwise coagglomeration at province and city levels, but the effect is insignificant at the county level. The impact of highways on coagglomeration is more significant for industry pairs characterised by a lower proportion of State-Owned Enterprises (SOEs), for industries that are closely related, and for industry pairings with input-output connections.

With respect to the effects of highway access on firm productivity, this study finds that highways have a beneficial impact on firm-level TFP, and a 1% change in highway access yields a 0.018% increase in TFP measured by the LP-ACF method. The results for four channels, including within-industry agglomeration, coagglomeration, export, and innovation, show that only the innovation channel is less significant. Highways increase within-industry agglomeration and coagglomeration for firms and stimulate exports with IV estimation.

The thesis is organized as follows. Chapter 2 explains background information about highway expansion in China, especially the highway construction plans. Chapters 3, 4, and 5 investigate the effects of highway access on within-industry agglomeration, coagglomeration, and firm productivity, respectively. Chapter 6 concludes this thesis.

Chapter 2

Background on China's Highways

2.1 Background on China's Highway Development

China's transportation infrastructure has undergone a substantial transformation since the 1990s, with the country's highways becoming the longest in the world, increasing from 500 kilometres in 1990 to 177,000 kilometres in 2022. Figure 2.1 gives a comprehensive overview of the expansion of China's highways from 1990 to 2022, reflecting the remarkable progress in the country's transportation infrastructure. In 1988, the highways were in their nascent stage, with a length of merely 100 kilometres. The early 2000s marked a significant acceleration in highway construction, reaching 53,900 kilometres by 2007.

This consistent growth in highway length underscores China's commitment to fostering efficient transportation systems to facilitate economic progress. The highways play a crucial role in China's infrastructure, as more than 70% of goods are transported by roads, enabling more efficient transportation of products across large areas. Figure 2.2 illustrates highways in China from 1998 to 2007 with GIS data from the ACASIAN dataset. More routes are constructed in eastern China, while highway construction is a greater challenge in the western regions where mountainous terrains dominate the landscape.

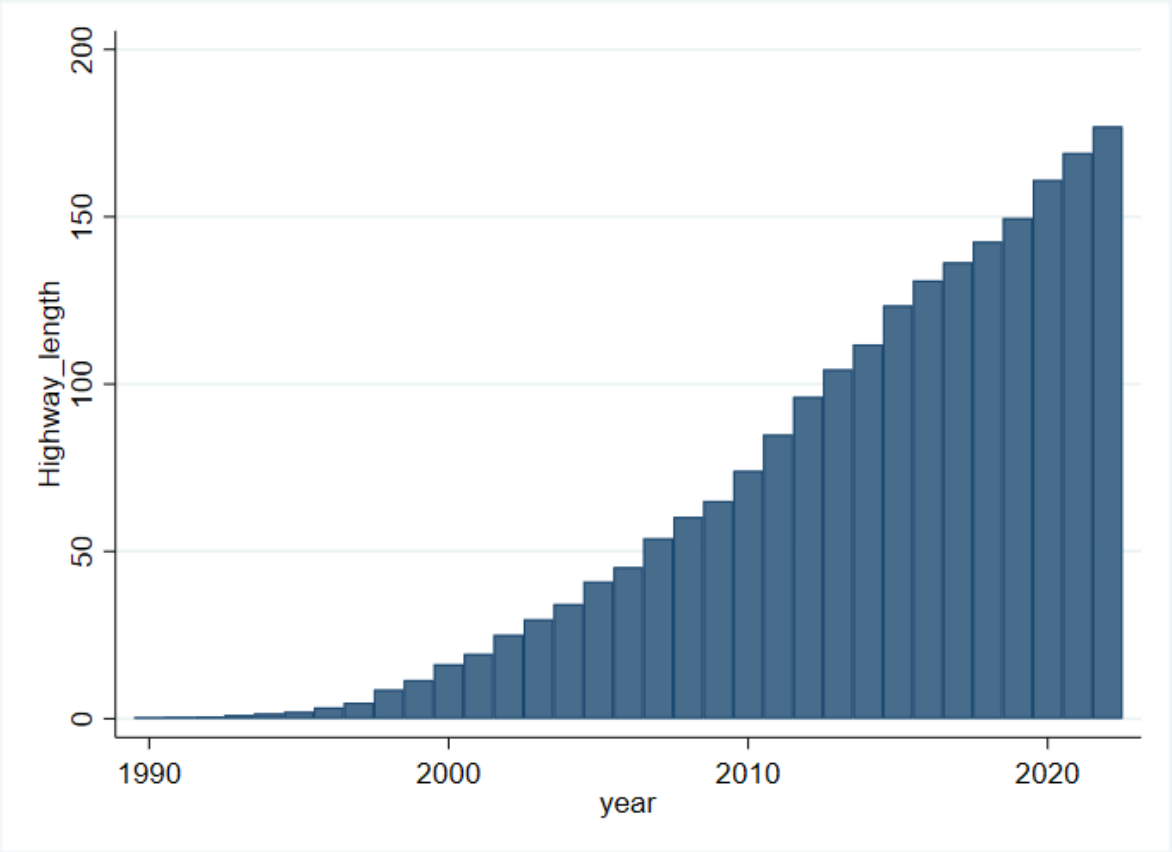
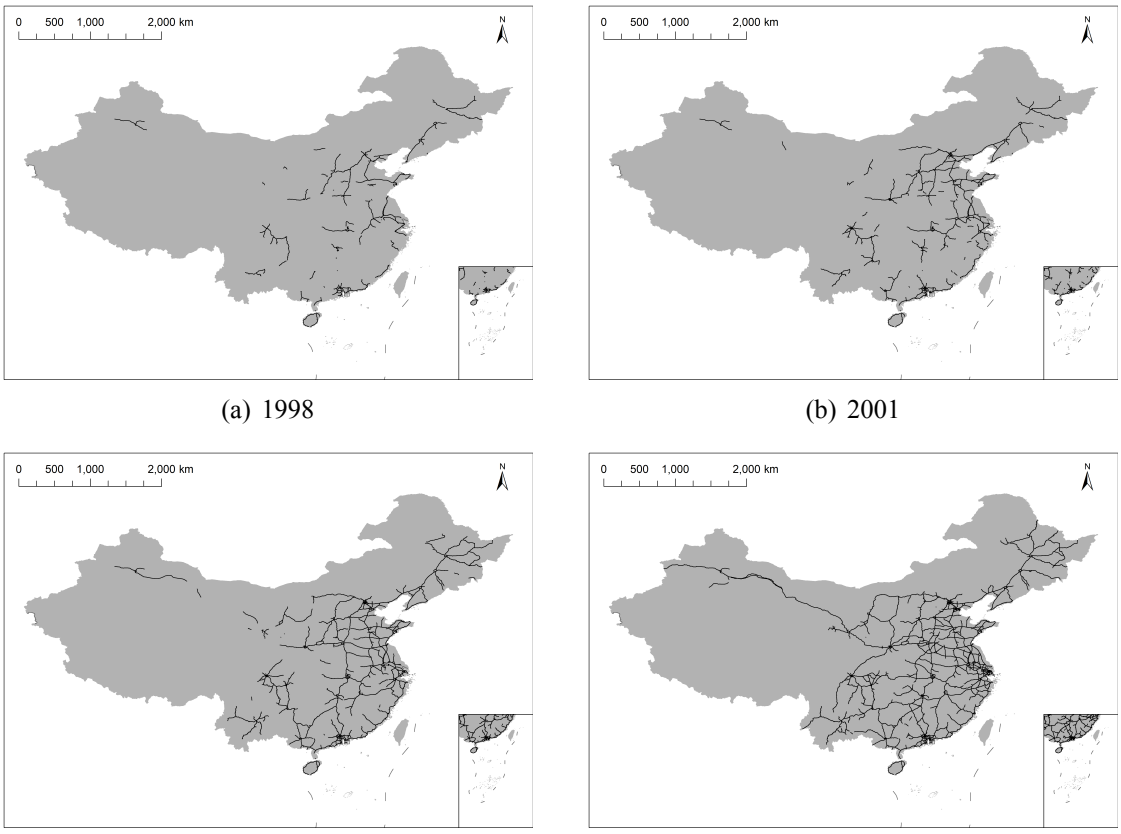


Figure 2.1: Highway length from 1990 to 2022

Note: The year is from 1990 to 2022. The data are presented in 1,000km. The highway length in 2022 is 177,000km. The data are obtained from China Statistical Yearbooks.

China's extensive highway development relies on a diverse range of funding. The government, both at the national and local levels, is a primary contributor and Public-Private Partnerships (PPPs) have also emerged as a dynamic tool for financing highway projects. PPPs attract private investments and leverage the efficiency of private expertise in construction and operation. Additionally, financial markets play a role through bond issuance, backed by government guarantees. State-owned banks provide loans, while toll collection on highways provides crucial revenue for sustaining projects. Toll financing ensures self-sustainability, covering construction costs, maintenance, and debt repayment. The diversified funding approach also includes investments from State-Owned Enterprises, international financial institutions, and vehicle-related taxes and fees.



(a) 1998 (b) 2001

(c) 2004 (d) 2007

Figure 2.2: Highways in China from 1998 to 2007

2.2 Roads Classification

The Ministry of Transport and its related agencies in China are responsible for establishing and enforcing highway engineering standards. In China's highway engineering standards, roads are classified into different ranks or classes, which consider characteristics such as traffic volume, design standards, and the intended use of the route. Highways in China refer to expressways and four classes of roads. Expressways are specifically designed to accommodate high-speed traffic with limited access points and are partitioned into lanes that accommodate traffic moving in opposite directions, typically from four to eight lanes in total.

First class roads in China occupy a position below expressways in the road hierarchy, ranging from four to six lanes. They are major routes designed to facilitate regional connectivity, and often have more flexible access points, slightly lower design standards, and may accommodate lower traffic volumes. Then moving down the hierarchy to second, third, and fourth class roads, the focus shifts to regional connectivity, addressing local transportation needs, with corresponding adjustments in traffic volume, design standards, and functionality. This research uses data on expressways with the characteristics of top speed and limited access.

2.3 Highway Construction Plans

Recognizing the crucial role of a well-connected transportation network, the Chinese government has introduced a long-term plan for a National Expressway Network since the 1980s. The network was organized into three phases: 5-7 plan, 7-9-18 plan, and the 7-9-18 network being the focus most recently. The construction of expressways is financed by both the central and local governments.

The National Trunk Highway System (NTHS) plan approved by the State Council in 1992 is known as the 5-7 plan (5 north-south and 7 east-west routes). The targeted cities are all provincial capitals, cities with urban populations of more than 500,000 and border crossings. The fiscal expansion programme in China was prompted by the 1997 Asian financial crisis. The enactment of the Highway Law in 1997 granted local governments the authority to participate in the process of choosing routes for the National Transportation Highway System. Subsequently, the pace of the NTHS building significantly accelerated. As of the start of 2003, 80% of the building of the NTHS had been finished and construction was completed in early 2008.

The 5-7 plan, approved in 1992, could not meet increasing road transport demand. In 2004 the State Council approved a follow-up plan for the National Expressway Network (NEN), which is also called the 7-9-18 plan, with 7 radial expressways from the capital, 9 north-south routes and 18 east-west routes. The aim of 7-9-18 is to connect cities with urban registered populations of over 200,000 and key nodes such as ports and transportation hubs. The 7-9-18 network is characterized by its exceptional technical quality and is composed only of expressways at the national level.

The third plan, proposed in 2013, extends the 7-9-18 network to a 7-11-18 network, incorporating two additional north-south expressways. This expanded network will connect emerging prefecture capitals and cities with populations exceeding 200,000 and is planned to be completed by 2030. The government places high importance on ensuring that expressways are of excellent quality, allow for faster travel, and have tolls, compared with national and provincial roads.

Chapter 3

The Effects of Highway Access on Within-industry Agglomeration

3.1 Introduction

The level of industrial agglomeration in China has increased over time. The economic reform in 1978 dramatically changed geographical economic activities. Many newly established firms selected their locations driven by the market. Spatial concentration occurred during that period, and for example, the coastal areas where there were more opportunities with natural advantages attracted more export firms. Wen (2004) finds that Chinese manufacturing was more spatial concentrated from 1980 to 1995 with many industries highly concentrated in some coastal areas in 1995. Lu & Tao (2009) also find that the industrial agglomeration level gradually increases during the sample period 1998-2005. What drives the fast expansion of industrial agglomeration? Does transport infrastructure such as the highway network contribute to agglomeration? The development of highways, the increasing level of agglomeration and the importance of transport costs on agglomeration motivate this study.

Highway construction in China experienced a fast development period, with the 5-7 plan in 1992, the 7-9-18 plan in 2004, and the 7-11-18 plan in 2013. The rapidly built highway networks significantly reduced transport time and costs. Regions connected with highways are expected to attract more firms. This chapter investigates the impact of highway construction

on within-industry agglomeration and its channels. The highways and industrial agglomeration are related, as highway networks are crucial for transporting manufacturing goods and affect the geographic distributions of manufacturing industries. This chapter contributes to four main aspects.

First, it investigates the impact of the highway network on within-industry agglomeration for the manufacturing sector, which is important but has not been thoroughly examined in the literature. Industrial agglomeration is important to the development of economies. Rising spatial concentration generates external economies and facilitates the development of industries. A rise of agglomeration has substantial effects on productivity (Ciccone & Hall 1996, Ciccone 2002, Lin et al. 2011, Hu et al. 2015). Agglomeration plays a crucial role in attracting foreign direct investment (Barrell & Pain 1999, Guimaraes et al. 2000) and promotes the exploitation of innovation (Feldman 1999, Antonietti & Cainelli 2011, Zhang 2015).

Some determinants of agglomeration include national advantages (Ellison & Glaeser 1999*a*, Roos 2005) and Marshallian externalities of agglomeration (Rosenthal & Strange 2001, Jofre-Monseny et al. 2011) have been researched substantially. Additionally, large-scale road investment is proved to affect concentration and the effects vary in different industries (Holl 2004). Wen (2004) finds that transport costs affect spatial concentration. The effect of rapid highway construction in China on industrial agglomeration has not been specifically investigated. This chapter provides empirical results showing that highway access has positive impacts on industrial agglomeration and the results are robust to the instrumental variables estimation.

Second, this chapter introduces a novel input-output adjusted measure of highway access to better capture the varying use of highways by different industries, a gap that has not been addressed in existing literature. Industries rely on highways to transport goods from suppliers and to customers, and improved infrastructure lowers transportation costs, fostering agglomeration and localisation as firms seek to maximise profits by concentrating in high-profit areas (Marshall 1890). Previous measures of highway access often ignore the variations in industry size and transportation volumes. By using input-output tables, this study's measure accounts for these variations and provides a more precise understanding of highway utilisation.

Third, the use of micro-economic data enables the calculation of Marshall's three mechanisms of agglomeration and agglomeration measures and makes the results more accurate compared to the macro-level data. The GIS highway data that depict explicit highway locations are used to capture the distance from each firm to the highway network and then group them at the industry level. The firm-level data and the GIS highway data allow this study to research heterogeneity, which enables it to go in-depth and makes the results more convincing.

Additionally, this chapter obtains the relationship between highways and within-industry agglomeration by using novel methods. This study uses time-variant instruments for the highway variable to address the endogeneity problem. The instrument for highways upgrades with the highway construction plan, which makes it a stronger instrument.

The structure of the chapter is as follows. Section 2 reviews both the theoretical and empirical literature on agglomeration, including its drivers, effects, and the specific research on agglomeration in China. Section 3 outlines the hypotheses developed for this study. Section 4 details the data and methodology used, along with a description of the key variable. Section 5 presents the empirical results of the baseline model. Section 6 addresses the instrumental variable estimations used to manage issues of omitted variables and reverse causality. Section 7 provides heterogeneity analysis. Finally, Section 8 provides the conclusions and discusses the limitations of the study.

3.2 Literature Review

3.2.1 Theory of Agglomeration

Agglomeration or clusters is the geographical concentration of economic activities. The concept of agglomeration is broad, with different compositions and many geographical levels. The research target can be a cluster of people, firms in the same industry or different industries, and geographical levels can be a small neighbourhood, industrial district, city, or even country. For example, the agglomeration identified by Marshall (1890) is the spatial concentration of the specialized industry.

Agglomeration theory has been developed over one hundred years. Although a variety of theoretical concepts have been developed, a unified theoretical framework has not emerged. Typical agglomeration theories include but are not limited to Marshall's agglomeration theory, location theory (Weber 1909), Dynamic externalities (and more specifically knowledge spillovers) using the Endogenous growth theory (Romer 1986) and Marshall–Arrow–Romer (MAR) externality (Glaeser et al. 1992), industrial clusters (Porter 1990), and new economic geography (Krugman 1991, Fujita 1988).

Though they are different, some of these theories have points in common. For instance, Marshall's external economies of agglomeration are at the basis of Dynamic externality theories and transportation costs are key factors in both the new economic geography of Krugman (1991) and the location theory of Weber (1909).

Marshall's agglomeration theory

Agglomeration theory is developed by Marshall (1890) in the chapter on Industrial organization-the concentration of specialized industries in particular localities in *Principle of Economics*. Marshall defines the localization of industry as the spatial concentration of many small businesses of a similar character. He discusses the cause of localization of industry and the external economies which arise from it.

The Causes of the Localization of Industry. Marshall (1890) explains the formation and causes for the localization of industry, which existed even in early stages of civilization. One chief cause is natural resources, such as the character of the soil and climate, nearby water access, mines and quarries. Court patronage that attracted skilled workmen to meet the demand for high-quality goods has been another chief reason for the development of a specialized industry. Rulers commonly deliberately invited specialized artisans from elsewhere and grouped them together. Local households are also able to learn from these artisans, and gradually a specialized industry becomes more localized.

External Economies and Internal Economies. Marshall (1890) first divides the economies that are generated from an increase in the scale of production into two categories: external economies and internal economies. Internal economies arise from the increasing scale of the firm itself; for instance, a firm can achieve lower costs when its scale of production increases. A large firm benefits from efficient management, can reduce production costs by producing specialized goods on a larger scale, has lower average costs for the raw materials from its suppliers, and enjoys stronger bargaining power with financial institutions to obtain a cheaper rate. On the other hand, external economies rely on the general expansion of the whole industry outside the specific firm. The growing industry shares a specialized workforce, ideas, and bargaining power with suppliers, all of which are positive for the development of firms in that specialized industry. However, negative effects also appear when the scale of the industry grows, a phenomenon which is known as external diseconomies. Severe competition, especially destructive competition, can cause a group of firms to disappear.

External Economies of Localization of Industries. Marshall (1890) discusses the advantages of the localization of industries. External economies appear when businesses of a similar character are concentrated in a particular locality. The advantages of localized industries discussed by Marshall have been classified and summarized by later scholars into three categories: knowledge spillover, labour pooling, and input sharing.

Knowledge Spillovers. Marshall (1890) indicates that localized industries enable people to acquire the same skills in the neighbourhood gradually, since when an industry has chosen a locality, it tends to stay there for the long term. The skills are no longer mysteries but learned by the people in the locality. The specialized skills are inherited by children who have easy access to them and unconsciously learn some of them. Good ideas and knowledge are appreciated and shared in the locality, and then combined with others' suggestions to form further new ideas. Inventions and innovations keep updated with the joint effect.

Input Sharing. Marshall (1890) states the importance of the subsidiary business for the localized industry. The localized industry has subsidiary businesses in the neighbourhood that provide it with materials and implements for production, and deliver goods and other activities that support the economy in the neighbourhood. Marshall (1890) also points out that the large aggregate production of similar goods in a district can make expensive machinery available to subsidiary industries. Subsidiary industries, which focus on a small part of the production process work for many other similar industries in the neighbourhood, and are able to use highly specialized machinery. The high costs of buying machinery and rapid depreciation are not big problems for them due to the fact that they can make income from their business and pay for the expenses. Advanced machines facilitate the efficiency and innovation of the industry.

labour Pooling. Another advantage of a localized industry is that the labour market offers special skills. The localization of industry offers a constant labour pool with the special knowledge that it requires. Firms located in a thick labour market have the advantage of employing skilled workers and tend to locate in the place where a lot of workers with the required skills are located. If an employer is in an isolated plant, even though they have access to a large general labour force, the lack of specialized workers will make production much more difficult, and thus, they will still need to make efforts to hire specialized workers. Therefore, firms tend to move towards places in which skilled workers are located. On the other hand, the spatial concentration of firms attracts talents and promotes workers' mobility in the industry or among related industries. Workers who seek employment are likely to move from a distance and settle in that place where their special skills are in demand.

Marshall (1890) also mentions the disadvantages of a district that only relies on one kind of worker. For example, an industry in which the work can only be done by strong men (e.g., mining industry), leaving women unemployed. However, this problem can be solved by the development of supplementary industries in that area to provide work opportunities for women.

Additionally, the disadvantage of a district that only depends on one specific industry is also discussed by Marshall, who points out that the district may experience extreme depression if there is a large decrease in demand or a failure in the supply of materials. The problem can be avoided by creating large industrial districts where several different industries are developed. If one of them fails for a time due to low demand or the shortage of materials, the other industries can mitigate its depression by supporting it indirectly. The labour market is more sustainable in that neighbourhood in this way.

Demand of Customers. In addition to discussing from the point of view of the economy of production, Marshall (1890) also points out that customers' demands and their convenience affect the localization of shops. He indicates that customers tend to travel some distance in order to find the special goods shops for their purchase of important goods, while choosing the nearby shops to buy less important goods. Consequently, shops that sell expensive goods are likely to gradually agglomerate, while ordinary convenience shops do not congregate. The localization of industry enables firms to share a larger customer base. The special goods that congregate together in order to meet customer demands attract customers from further away. When a place is well-known for trade in particular goods, it can attract customers and suppliers worldwide.

In summary, agglomeration theory discussed by Marshall (1890) is an initiative. Marshall provides the causes of the localization of industry: natural advantages and the demand of previous rulers. External economies through localization of industry, including knowledge spillovers, input sharing, and labour pooling, are the benefits from localized industry but also promote the localization of industry in their ways to some extent.

Weber's theory of industrial location

Another branch of industrial agglomeration known as location theory is developed by Weber (1909). Weber claims that agglomeration emerges in spatially advantageous locations. Compared with that of Marshall (1890), Weber's theory takes a different perspective to investigate agglomeration.

The 'locational factors'. Weber (1909) puts forward the 'locational factors' which determine the location of industries with a minimum cost of production. The cost of production is different in various areas, and thus industries locate where they have the minimum cost of production. The locational factors are classified as general factors and special factors. The former affect every industry though their influence can be more or less in different industries, such as the cost of labour, transportation, and rent; the latter are the special concerns of only this or that industry, such as perishability of raw materials, the humidity of the air in the factory, the reliance on clean water, etc.

Regional factors, 'agglomerative' or 'deglomerative' factors. Moreover, all locational factors, whether general or special factors, are further divided into regional factors which can be 'agglomerative' or 'deglomerative'. The regional factors affect the regional distribution of industry that creates the primary framework of industrial locations. The 'agglomerative' or 'deglomerative' factors redistribute industry. Among regional factors, transportation costs and labour costs are crucial factors that affect the production costs of industries in different regions. The cost of transportation, labour and agglomerative factors are the most important factors for the location of industries.

Weber's location theory first discusses the idea that firms choose the location that minimizes transportation costs and then attributes the deviation from minimum transportation costs to economy of labour and agglomeration. The transportation cost is determined by the weight of goods and transport distance.

The 'ubiquitous material' and 'localized material'. Weber (1909) distinguishes between the 'ubiquitous material' and 'localized material'. 'Ubiquitous material' is available everywhere in certain areas, such as wood and grain and since they are everywhere, the cost of transportation is considered. Therefore, to minimize transportation costs, firms tend to locate near their customers. On the other hand, some materials are only obtainable in a particular region rather than everywhere, and they are called 'localized material'.

'Pure material' and 'gross material'. Weber (1909) divides localized material into 'pure material' and 'gross material'. Considering the nature of the production process in which raw materials are processed into final products, the process can either leave some residue or none and residues may be used to manufacture other products. 'Pure material' means the total weight of raw materials is used for the product; ''Gross material' means only part of the materials can be used for the product, an extreme case of which is the fuel used for production as the weight of fuel is a loss in the production process. Fuel is the 'weight-losing material', which means its total weight as a residue.

When localized material is used for production, the material index is the proportion of the weight of localized material to the weight of finished goods. The location of a firm is determined by the transportation costs to the market and raw materials, considering the weights of raw materials and finished products. The material index captures the trade-off of the transportation cost when transporting goods from suppliers to customers. For instance, a firm will locate near the market if finished products are heavier than the source materials. The total weight to be moved, called the locational weight, depends on the material index.

The material index of 'pure material' is one, and it is larger than one for 'gross material'. If the residue during production is large, firms tend to locate near the raw materials. Thus, having a larger residue and using fewer 'ubiquitous materials' lead firms to locate near the location of their materials. Weber (1909) summarizes three principles according to the minimum cost of transportation: (1) a firm only using 'ubiquitous material' locates near its customers; (2) when only 'pure material' is used for production, the firm locates freely either where customers are or materials are; (3) when only 'gross material' is used for production, the firm locates in the same area as its materials. Furthermore, when there are two material regions and one market, the location triangle is used by Weber (1909) to determine the optimal location of an industry. When there are multiple regions with raw materials sources and markets, the polygon is used instead. According to the weights of two raw materials and final output sold on the market, Weber uses the methodology of the Varignon Frame that considers weights and pulleys to find the optimal location that achieves the minimum transportation cost, which is at the centre of gravity of the triangle.

The labour cost factor. Weber (1909) indicates that labour costs are affected by population density. The labour cost index is the labour cost of a unit weight of product. A higher labour cost index in an industry implies that it is likely to move to where labour costs are low. He also develops the 'labour coefficient' to measure the labour deviation in the industry that equals the ratio of labour cost to locational weight, which decides whether transportation costs or labour costs have the upper hand in the choice of location. A higher labour coefficient implies that the industry tends to move to a place where labour costs are lower.

'Agglomerative' or 'deglomerative' factors. Weber (1909) defines an agglomerative factor as a cheapening of production or marketing owing to the fact that production is carried out in one place to some extent (centralization of production), while a deglomerative factor is a cheapening of production because of the decentralization of production. Weber (1909) distinguishes two stages of agglomeration. The first stage results from the enlargement of a plant, focusing on the rise in production of an individual firm. A plant with a local concentration of production has advantages over productions scattered in small plants, as large-scale production has economic advantages over small-scale production, such as saving on technical appliances, improving labour organization, and cheap large-scale purchasing. The second stage of agglomeration results from close local association of several firms, which Weber (1909) also calls social agglomeration. Just as with the benefits of the enlargement of a plant, several firms also congregate together to develop technical equipment, improve labour organization, have division of labour, create a more effective marketing situation, conduct large-scale purchasing, and reduce general overhead costs by jointly sharing gas supply, water mains and streets. Weber also breaks agglomeration down into 'accidental agglomeration' and 'pure agglomeration'. 'Accidental agglomeration' is defined as resulting from minimum transportation cost or labour costs, while 'pure agglomeration' (or 'technical agglomeration') is a necessary consequence, that is, the concentration of firms brings the advantages of improvements in technical equipment, labour organization and specialized facilities and thus firms choose to congregate. Weber (1909) asserts that economies of agglomeration shift the locations of firms.

Weber (1909) indicates that agglomeration may lead to opposing tendencies. Expenses increase with a rise in rent, higher general overhead costs due to higher land value, and higher labour costs. These are 'degglomerative factors' that decentralize the location of industries, and greater agglomeration causes stronger degglomeration forces.

Since the definition of agglomeration is abstract in Weber (1909), Hoover (1937) criticizes Weber for not distinguishing the economies of agglomeration. The three categories of economies of agglomeration identified by Hoover (1937) are: large-scale economies, localization economies and urbanization economies. Large-scale economies are internal economies, and localization economies and urbanization economies belong to external economies; these focus on the size of a company, the size of an industry and the size of a city, respectively.

The New Economic Geography

The new economic geography explains the formation of agglomeration in geographical space (Fujita & Krugman 2004), on the basis of the increasing returns to scale and monopolistic competition assumptions. Krugman (1991) and Fujita (1988) develop new economic geography models based on the monopolistic competition model. Dixit & Stiglitz (1977) develop a remarkable monopolistic competition model, which is a workhorse in the models of new trade, new growth and also new economic geography literature. Monopolistic competition refers to the competition between firms with distinct products, but whose products are similar enough to be substituted. Competition between firms producing close substitutes limits monopoly power.

It is widely acknowledged that the core-periphery model developed by Krugman (1991) gives birth to the new economic geography. Different from the external economies mentioned by Marshall (1890), the core-periphery model concerns pecuniary externalities, neglecting all other sources of agglomeration economies. Pecuniary externalities are associated with either demand or supply linkages. The core-periphery model shows how the economic structures of two regions evolve under the interactions between increasing returns, labour mobility and transport costs. It is a two-region model with two production sectors (manufacturing and agriculture sectors), with agricultural production involving constant returns to scale with homogeneous goods.

In this model, farmers are immobile and have the same earnings in both regions. In contrast, manufacturing production involves increasing returns to scale, and workers are freely mobile between the two regions. The utility function of all individuals assumes a specific form where consumption is divided between agricultural goods and a manufactured goods aggregate. The share of expenditure on manufactured goods is a key parameter that determines the convergence or divergence of the region. The elasticity of substitution among goods is another crucial parameter influencing this dynamic.

Transportation costs play a significant role in the model. It is assumed that the transportation of agricultural output between two regions is costless, ensuring that each farmer has the same earnings regardless of location. However, manufacturing goods incur transport costs, modelled using the 'iceberg' form introduced by Paul Samuelson (1952). In this form, only a fraction of the goods shipped arrive at the destination, representing transport costs. The fraction of manufacturing products that successfully arrive after shipment is a key parameter that influences regional convergence or divergence.

Firms in the model are assumed to operate under a certain labour requirement, with workers and farmers evenly distributed between the two regions. Farmers are completely immobile, while workers can move between regions. The supply of workers in each region depends on the model's assumptions about mobility and labour distribution. Each firm in the model produces a single product, and there are many firms in both regions. Firms set prices to maximize profits, considering the wage rates and demand in their respective regions. The relationship between the number of manufactured goods produced and the number of workers in each region is directly proportional.

The model examines regional differences in nominal and real wages of workers between two regions. If workers in the region with lower real wages migrate towards the other region, it leads to either regional convergence, where the worker-to-farmer ratio equalizes, or divergence, where workers concentrate in one region. The equilibrium of the model is determined by the complex interaction of multiple factors, including regional incomes, price indices of manufactured goods, and nominal wages of workers.

The key conclusion of the model is that when transportation costs are high, the relative real wage difference between regions decreases as the share of manufacturing labour increases, leading to regional convergence. Conversely, when transportation costs are low, an increase in the share of manufacturing labour leads to higher relative real wages, promoting regional divergence. Ultimately, if transport costs are sufficiently low, goods are sufficiently differentiated, and expenditure on manufacturing goods is large enough, the economy may develop into a core-periphery pattern, with all manufacturing concentrated in one region.

In the core-periphery model, there is a trade-off between two opposing effects - centripetal forces and centrifugal forces - that determines the formation of agglomeration. The home market effect and price index effect promote agglomeration. The home market effect means that when increasing return to scale and transportation cost are present, a country with large domestic demands will have higher output and export goods to other countries. A region with more workers (meaning more customers) has a larger market, which implies that it is more profitable to produce in this region because of the transport cost. The home market effect leads to the fact that the wage rate is higher in a large market. The home market effect is the force working towards regional divergence. The price index effect indicates that the low price for manufactured goods results from the large manufacturing sectors. Price index and home market effects imply that higher real wages arise in regions where larger manufacturing sectors exist.

Krugman (1998) also indicates some forces that are important for spatial concentration. The centripetal forces include market-size effects (linkages), thick labour markets and pure external economies (e.g., knowledge spillovers). The immobile factors (certain land and natural resources, and people), land rents and pure external diseconomies (e.g., congestion) are centrifugal forces.

Summary of agglomeration theory

In summary, agglomeration theory is rich but not unifying. Marshall's agglomeration theory plays a central role in sufficient theories. Marshall mainly discusses external economies (i.e., the factors that affect firms come from something outside the firm) and explains three external economies of agglomeration: knowledge spillover, input sharing, and labour market pooling.

Weber's location theory and Krugman's core-periphery model both indicate that transport cost plays a major role in industrial agglomeration. The new economic geography involves increasing returns and monopolistic competition since constant returns to scale to some extent fail to explain the formation and growth of agglomeration (Krugman 1991, Fujita 1988). Heterogeneous agents have also recently been incorporated to make the model more rigorous. Table 3.1 displays the Key points of the typical theories in agglomeration.

Table 3.1: Key points of the typical theories in agglomeration.

Theory	Unit of agglomeration	Key points and conclusions
Marshall (1890) External economies	Agglomeration of small businesses of a similar character	External economies of location of industry: input sharing, knowledge spillover, labour market pooling.
Weber (1909) Location theory	Agglomeration of any firms or any industries	The cost of transportation, labour and agglomerative factors determine the location of industry. The economies of agglomeration shift the location of firms. The economies of agglomeration are discussed abroad but not distinguished enough.
Porter (1990) Industry cluster, Competitiveness theory	interconnected institutions and firms, such as suppliers, customers, financial, training institutions, etc.	Clusters foster regional or national competitiveness in the global economy. Competition in clusters increases productivity, and innovation and promotes the formation of new businesses. Factors determine the competitive advantage: firm strategy, structure and rivalry, demand conditions, factor conditions, related and supporting industry.
Krugman (1991) New Economic Geography; The core-periphery model	Two sections: manufacturing and farm. Agglomeration of the whole manufacturing sector in the model.	Under the framework of increasing returns and imperfect competition. Only focuses on pecuniary externalities (changes in price). The formation of divergence of two regions is when the transport cost is low, goods are sufficiently differentiated and expenditure on manufacturing goods is large.
Glaeser et al., (1992); Jacobs (1969) Dynamic Externalities	MAR externality (in the same industry); Jacobs externality (diversity, different industries)	Based on endogenous growth theory. The dynamic pattern of agglomeration mainly focuses on knowledge spillover in the process. The build-up of knowledge in the long term through agglomeration.

3.2.2 Drivers and Effects of Agglomeration

Drivers of agglomeration

The causes of agglomeration include physical conditions like natural advantages, home market effects, and externalities of agglomeration that encourage firms to agglomerate such as knowledge spillovers, labour market pooling, and input sharing.

Ellison & Glaeser (1999*a*) argue that natural advantage is important for agglomeration. They use the costs of 16 proxies to reflect the natural advantages. They show that natural advantages explain about 20 percent of the agglomeration and conjecture that more than half of agglomeration results in natural advantage due to the natural advantage proxy being imperfect. Roos (2005) regresses a number of variables capturing geographic features on several measures of agglomeration. The geographic features include coal, river, lignite, mountain, national borders, etc., finding that about one third of the agglomeration measure can be attributed to geography.

Holl (2004) examines the effect of large-scale road investment and agglomeration economies on the creation of firms using municipality-level data in Portugal over the period 1986-1997. He regresses the share of firm creation on proxies of transport improvement and agglomeration. He finds that transportation infrastructure improvements affect firm-birth concentration varies in different industries; diversity encourages most firm creation and there is little evidence for the benefits of local specialization on firm-birth.

Rosenthal & Strange (2001) use firm-level data containing over 12 million U.S. firms in manufacturing industries. They adopt the Ellison & Glaeser (1997) measure of spatial concentration and regress it on proxies for knowledge spillovers, labour market pooling, input sharing, natural advantages that affect input shipping costs, and product shipping costs at the zip code, county and state levels, respectively. They find evidence that shows the importance of all these determinants for agglomeration. Some more detailed results indicate that proxies for labour market pooling positively affect agglomeration at all spatial levels; proxies for knowledge spillovers positively influence agglomeration only at the zip code level; manufactured inputs or natural resources and proxies for product shipping costs have a positive effect at state level but little effect at lower geographical levels.

Using Spanish manufacturing firm data Jofre-Monseny et al. (2011) explore the importance of the three mechanisms of agglomeration suggested by Marshall (1890). They regress the count of new firms on labour market pooling, input sharing, and knowledge spillovers, where labour market pooling is measured by labour similarity, that is the extent to which the distribution of labour by occupation is similar to another industry; input sharing is measured

by share of the inputs that one industry purchases from other industries (data are from the Catalan Input-output Table); and knowledge spillover is measured by technology similarity in different industries. The result shows evidence of all three agglomeration mechanisms and their incidence differs in analysis with different spatial scales.

Jofre-Monseny et al. (2014) use city-level data and firm-level manufacturing data in Spain and regress firms' profits on local employment within the industry and in other industries to obtain estimates that reflect the importance of localization economies and urbanization economies. They find that localization and urbanization effects are important in many industries. Then, they regress the two types of estimates on labour market pooling, input sharing, and knowledge spillovers. They shed light on the causes of localization and urbanization economies, finding that urbanization effects are higher in knowledge-advanced industries and localization effects are higher in industries where more industry-specific workers' skills are required and a pool of specialized workers is shared.

The Effect of Agglomeration

Effects of agglomeration on productivity. Ciccone & Hall (1996) explore the effect of state-level agglomeration (employment density) on productivity (gross state output for the US). Using US data they build two models: the increasing return of externalities model and the increasing return variety of intermediate products model, finding that a rise in employment density increases labour productivity. Ciccone (2002) also investigates the effects of regional agglomeration (measured by employment density) on average labour productivity of some European countries: France, Germany, Italy, Spain and the UK. Their estimation is based on two spatial agglomeration models of increasing returns by Ciccone & Hall (1996). They find substantial agglomeration effects in these countries and the effects are not significantly different between these countries.

Balat et al. (2018) investigate the impact of agglomeration on productivity in Colombia with a firm-level panel that includes input and output data for the period 2005–2013 and a panel of municipalities that contains information on city characteristics. They use a control function approach to estimate productivity for each firm, extending the work by Olley and Pakes (1992). The measures of agglomeration include the scale of a city, the degree of sector special-

ization that captures localization economies or intra-industrial spillovers, industrial variety by pseudo-Herfindahl indices that capture urbanization economies or cross-industry spillovers, and level of competition. They find that scale economies do not seem to affect firms' productivity. Additionally, they also show that industrial specialization has a positive impact on productivity while industrial variety hinders productivity in Colombia.

Agglomeration can also hinder productivity. A dense degree of agglomeration may generate congestion and fierce competition, and fiercer competition in product and factor markets can suppress productivity through lower product prices and higher input costs. Lall et al. (1999) find that geographic concentration of the same industry lowers the productivity of Indian firms, while access to markets through improvement in inter-regional infrastructure has a positive effect.

Effects of agglomeration on FDI. Barrell & Pain (1999) explore the effect of host country labour market institutions and agglomeration on the location decisions in Europe of investment by US multinational firms. The geographical distribution of US manufacturing foreign affiliates in EU countries varies, for example, flexible labour markets and lower regulatory burdens attract inward investment to the UK but cannot account for the rapid growth of inward investment in France and Germany since the single market programme started (Barrell & Pain 1998). Barrell & Pain (1999) agree with Krugman's theory of new ergonomic geography that emphasizes the increasing return of scale, differentiated products and trade costs in the location of industry. Barrell & Pain (1999) hold the view that both the firm's internal economies of scale and external economies between firms are reasons for agglomeration, as evolved technology within a firm affects the number of locations used, and technology spillovers among firms shift the location decision of firms. Barrell & Pain (1999) estimate the importance of labour costs and agglomeration on US manufacturing FDI in Europe. In their paper, the strength of agglomeration effects has two measures: the ratios of host country output (and R&D) to EU output (and R&D). They find that these two agglomeration variables both have significant positive effects on the stock of FDI. The potential for agglomeration attracts inward investments.

Guimaraes et al. (2000) investigate the location decisions of foreign-owned manufacturing firms in Portugal. Their paper is germane to the theory of urban service agglomeration by Rivera-Batiz concerning agglomeration economies in consumption and production under increasing returns and monopolistic competition. They regress all investments with foreign participation on factors like agglomeration, labour costs, population density and education, finding that agglomeration plays a crucial role in the location selection of FDI.

Effects on innovation. Feldman (1999) reviews innovation and agglomeration studies and summarizes four strains regarding spillovers in the empirical literature: innovation production function, patent citation linkages, the mobility of skilled labour and spillover embodied in traded goods. Baptista & Swann (1998) examine the impact of industry clusters on innovation with data from 248 manufacturing firms from 1975 to 1982. They find that localization of specialized industries promotes innovation probably due to technology spillover, while local diversity of industries does not have a significant effect on innovation.

Antonietti & Cainelli (2011) explore the relationship between agglomeration and innovation in Italian manufacturing firms. They use the model that identifies the factors underlying the intensity of R&D investments, R&D capital and innovation output, TFP and export performance. They find that agglomeration of both specialized industry and a variety of industries has positive effects on innovation but in different ways. Local diversity promotes R&D and creates new ideas, while specialization promotes the exploitation of innovation and leads to higher TFP.

Dynamic agglomeration externalities and location choice. Head et al. (1995) investigate the positive agglomeration effects on the location choice of firms with data from 751 Japanese manufacturing firms in the US. They find that MAR externalities significantly affect location choice. A rise in the degree of agglomeration increases the future selection of that place. Firms tend to be located where other firms in the same industry are located. They explain that the externalities of intermediate inputs and technology spillover of the same industry attract firms to agglomerate.

Henderson et al. (1995) investigate the dynamic externalities and life cycle of industries with data for eight manufacturing industries in the US between 1970 and 1987. They test both MAR externalities (associated with own industry employment) and Jacobs externalities (employment in different industries in the locality). They find that only MAR externalities for mature industries, that is, production of mature industries decentralizes to smaller and more specialized cities; both MAR and Jacobs externalities for new high-tech industries, that is, new high-tech industries prosper in large, diverse metropolitan areas. Neffke et al. (2011) investigate the agglomeration externalities and life cycle (young, intermediate, and mature) of 12 Swedish manufacturing industries. Their results are similar to those of Henderson et al. (1995): a rise of MAR externalities occurs when industries are more mature, and for young industries, Jacobs externalities (benefits of local diversity) are positive, but they decrease when industries mature.

Dumais et al. (2002) examine the dynamic process of geographic concentration. They view the agglomeration of industries as a dynamic process in which the plant life cycle from birth to expansion/contraction and then closure contributes to spatial concentration. The agglomeration index by Ellison & Glaeser (1997) is extended with two decompositions of concentration changes: industry mobility and plant life cycle. They use US manufacturing industry data from the Census Bureau's Longitudinal Research Database and find that there is a large amount of movement for many spatially concentrated industries, though the industry agglomeration level seems fairly constant and has declined only slightly over the last quarter century in the US. They find that plant birth reduces agglomeration as new firms are generally located away from established centres. They imply that the effect of industry spillover on new plants is not sufficiently strong to attract them.

3.2.3 Research on Agglomeration in China

Drivers of agglomeration in China

The drivers of agglomeration in China have also been investigated from different aspects, including increasing return to scale, transport costs and foreign direct investment. Local protectionism on the other hand is found to hinder industrial agglomeration (Bai et al. 2004, Lu & Tao 2009). Wen (2004) examines the trend and drivers of spatial concentration of Chinese

manufacturing from 1980 to 1995, using data from national industrial censuses, and finding that both transaction and transport costs and increasing return to scale result in industrial agglomeration in China. Gini coefficients are used to measure spatial concentration. Their results show that Chinese manufacturing is more spatially concentrated in 1880, 1985 and 1995 and many of them are highly concentrated in some coastal areas in 1995.

In respect of FDI, Hsu et al. (2019) investigate the effect of foreign direct investment on industrial agglomeration by exploiting the plausibly exogenous relaxation of FDI regulations between 1997 and 2002. The degree of industrial agglomeration is measured using the method of Ellison & Glaeser (1997). Hsu et al. (2019) use panel data on industrial firms from the National Bureau of Statistics of China (NBS) Annual Survey of Industrial firms from 1988-2007 and a difference-in-differences estimation approach, finding that the FDI deregulation in 2002 on average led to special dispersion of industries. They develop a theory of a hump shape between FDI and industrial agglomeration, which indicates that as the scale of the economy grows to a threshold, FDI induces dispersion due to the high competition pressure and decreasing diffusion of technology.

Ge (2009) sheds light on the relationship between industrial agglomeration and globalization in China by investigating the effects of foreign trade and FDI. Ge uses the Ellison–Glaeser index to measure industrial agglomeration for two-digit industry during 1998–2005 with the NBS firm dataset and aggregate manufacturing sector statistics from the China Statistical Yearbook. Ge estimates the fixed effect model and system GMM model and provides three pieces of empirical evidence: first, the increasing degree of industrial agglomeration from 1985 to 2005, where exporters and foreign-invested firms are more spatially concentrated. Second, two determinants of industry location, including foreign trade and FDI, are examined, and prove to significantly affect firms to locate in regions with access to foreign markets (easy access to sea transportation). Third, export intensity and foreign investment, the two determinants of industrial agglomeration, are investigated and proved to positively affect industrial agglomeration.

Regarding the effects of local protectionism, Bai et al. (2004) investigate regional specialization and local protectionism. They construct a measure of regional specialization called the Hoover coefficient of localization based on the location quotient with respect to output. They find that local protectionism discourages industrial agglomeration. The overall time trend for regional specialization drops in the mid-1980s, and experiences a significant boost in later years. Lu & Tao (2009) adopt the Ellison and Glaeser measure of agglomeration with firm data from the Annual Survey of Industrial Firms conducted by China's National Bureau of Statistics over the period 1998-2005, finding that industrial agglomeration in China increased consistently during that time. Moreover, local protectionism is indirectly measured by the share of state-owned enterprises in employment (a higher share presents a higher incentive for local protectionism). Their OLS estimation shows that industrial agglomeration is negatively affected by the share of state-owned enterprises in employment, suggesting local protectionism obstructs industrial agglomeration in manufacturing sectors. Their result is robust to instrumental variable estimation with the share of state-owned enterprises in 1985 being the instrumental variable.

The Effects of Agglomeration in China

Agglomeration and productivity. Empirical research on the effects of agglomeration in China finds that agglomeration increases productivity. Ke (2010) investigates the relationship between industry agglomeration and urban productivity across Chinese cities. The cross-sectional analysis uses data for 617 cities and data for the number of highways and capacity of airports in 2005. The results show that agglomeration and urban productivity are mutually and causally related; employment density has a negative effect on productivity, especially in prefectures. Their spatial model indicates that productivity in neighbouring cities positively affects others within a 100km scope.

Lin et al. (2011) use an unbalanced panel dataset for Chinese textile firms over the period 2000-2005 to investigate the dynamics of industrial agglomeration and the effect of agglomeration on productivity in the textile industry. They follow the Ellison-Glaeser measure of spatial concentration and find a slightly decreasing trend of agglomeration in the textile industry, but the degree of agglomeration of the textile industry in China is high. Additionally, their analysis shows that industrial agglomeration positively affects productivity.

Hu et al. (2015) investigate the effect of industrial agglomeration on productivity. They conduct panel data analysis with NBS annual survey firm data for three-digit industries between 2000 and 2007 and a cross-section analysis using 2004 census data that incorporate the clustering of small firms. Their agglomeration degree is measured by the index of the number and average size of firms in each county. They use the input-output tables of China (two-digit level) to calculate the industry-level input coefficients that measure the degree of agglomeration of the upstream industry. Their estimation model relates the firm's TFP to the agglomeration of the same industry and upstream industries. They find that the agglomeration of upstream industries positively and considerably contributes to the TFP of Chinese firms and the co-location of large firms significantly promotes the TFP. On the other hand, they also present empirical evidence that the increase in the number of firms in a county suppresses the TFP, suggesting that severe congestion and competition offset the benefits of agglomeration.

Wan & Zhang (2018) investigate the effect of infrastructure on productivity in direct and indirect ways, indirect way meaning that infrastructure affects productivity through the agglomeration channel. Agglomeration is simply measured by the provincial share of sales, while asset share and the share of the number of firms are used as the other two agglomeration indicators for the robustness check. The infrastructure in their research incorporates road, telecommunication servers and cable with data collected from provincial statistical yearbooks in China. Firm-level data come from the annual survey of industrial firms data by China's National Bureau of Statistics over the period 2002-2007. They find that both the direct and indirect effects of infrastructure on productivity are positive and significant. Their results generated from the two-stage least squares estimation imply that roads and telecommunications help promote agglomeration.

The effect of agglomeration on exports. Ito et al. (2015) use the Annual Survey of Industrial Firms over the period 2000-2007 to examine the hypothesis that industrial agglomeration of exporters lowers the productivity threshold for exporters and promotes the likelihood that a firm will export. Their semi-parametric quantile regression indicates that the productivity

advantage of exporters against non-exporters is smaller in agglomerated regions defined as counties where the number of firms is larger than the 95th percentile of the distribution. Additionally, a parametric estimation of the model of export entry shows that the agglomeration of incumbent exporters contributes to export entry but its magnitude is limited.

The effect of agglomeration on firm size. Li et al. (2012) explore the effect of agglomeration on firm size. They adopt annual surveys of manufacturing firms by the National Bureau of Statistics of China from 1998 to 2005. The measure for industrial agglomeration and the measure for firm size follow the specification of Holmes & Stevens (2002). The historical population in 1986 is used to instrument the degree of industrial agglomeration. Their results indicate that industrial agglomeration has a significantly positive effect on firm size and imply that a firm tends to be larger through spatial concentration with many large firms rather than with a large number of firms.

3.2.4 Research Gap

Transport costs play a key role in agglomeration (Krugman 1991, Weber 1909). Highways shorten the time required to transport intermediate goods and reduce transport costs. Additionally, the construction of highways may lead firms to locate close to access to highways, and thus they are more likely to be spatially concentrated near them.

There is not much research investigating agglomeration and highways. Song et al. (2012) investigate the relationship between transport accessibility and industrial agglomeration in metropolitan Seoul and publish their research in the Journal of Geographical Systems. The two-digit industry classification is adopted and only 11 industries among 18 two-digit industries remain in the analysis. Time accessibility of the subway network, road density and the distance to the national highway are used to measure transport accessibility. They find that transport networks are positively associated with industrial agglomeration in general. Wan & Zhang (2018) find that infrastructure positively impacts productivity both directly and indirectly, with the indirect effect occurring through the agglomeration channel. Their study shows that roads and telecommunications significantly promote agglomeration, as indicated by the positive relationship between infrastructure and agglomeration metrics such as the

provincial share of sales, asset share, and the number of firms. However, most literature investigating the effect of infrastructure relies on investments or road mileage due to the lack of GIS data. This study uses GIS data to analyze the explicit location of highways, which is expected to produce more accurate and detailed results.

3.3 Hypothesis Development

3.3.1 Overall Effect of Highway Access

The relationship between highway access and agglomeration is indeed multifaceted and depends on various factors. While overall, improved highway access tends to positively influence agglomeration, there is significant heterogeneity in its effects based on industry characteristics and other factors. The impact of reduced transport costs on agglomeration might follow an inverted U-shaped pattern. Initially, as transport costs decrease, firms might cluster together to maximize the benefits of agglomeration. However, beyond a certain point, further reductions in transport costs might encourage firms to disperse, seeking to avoid the downsides of high-density locations, this trend also aligns with the theory by Krugman (1979), who states when transport cost keeps falling, the concentration increase and when transport cost is zero, the location does not matter.

Firms' locations are jointly affected by the places of their supplier, customers, and other firms in the same industry. Firms are supposed to choose to locate where they obtain the most profits. The effects of lower transport costs on agglomeration can be an inverted U shape. When transport costs decrease, they can move to where they benefit the most, in this case, agglomeration provides more benefits, and thus firms can locate together. However, due to land price and congestion, agglomeration may not most benefit firms. Weber (1909) suggests that as industries cluster together, costs may rise due to increased rent, higher overheads from elevated land values, and greater labour expenses. These factors encourage industries to decentralise their locations. Thus the effects of lower transport costs may achieve a peak. When transport costs are zero, even when labour mobility and knowledge spillover can be

instantly transferred, firms can locate anywhere, although this situation cannot happen in reality. Considering the time period of the sample (1998-2007), China experienced fast development during this period, thus the overall effects of highway access on agglomeration are hypothesised to be positive.

3.3.2 Supplier Effect and Highway Access

In addition to the overall positive effects of highway access on agglomeration, this thesis provides different situations of how highway access can affect agglomeration. Weber (1909) argues that a firm's location is influenced by transportation costs related to both the market and raw materials, taking into account the relative weights of raw and finished products. The material index quantifies the trade-off in transportation costs between moving goods from suppliers to customers. Capturing the agglomeration effects of suppliers is challenging with existing data, as industries often have a diverse range of suppliers and complex supply chains. This complexity makes it difficult to accurately measure the impact of agglomeration and the location of all suppliers for each industry. As a result, research often focuses on natural advantages as a significant type of supplier when examining agglomeration (Marshall 1890, Weber 1909, Ellison & Glaeser 1999*b*, Roos 2005, Rosenthal & Strange 2001).

Natural resources represent a specific type of supplier with fixed locations, making their effects easier to capture. Therefore, we investigate natural resources as a proxy to examine the impact of supplier agglomeration. Ellison & Glaeser (1999*a*) argue that natural advantage plays a significant role in industrial agglomeration. By using the costs of 16 proxies to represent natural advantages, they find that these factors account for about 20 percent of agglomeration. They also suggest that the true impact may be even greater, possibly over half, due to the imperfect nature of the proxies used. Similarly, Roos (2005) indicates that roughly one-third of agglomeration can be explained by these geographic characteristics.

Since natural resources are immobile, firms tend to locate near these resources when transport costs are high, resulting in a higher level of agglomeration. However, when transport costs decrease, firms may either continue to cluster together (potentially attracting additional firms) or relocate to other areas where they can achieve greater profitability, possibly due

to higher land prices in agglomerated regions. Therefore, the impact of highway access on agglomeration under these conditions could vary, but it is more likely to be negative. Consequently, we hypothesise that as highway access improves, the influence of natural resources on within-industry agglomeration is likely to diminish.

3.3.3 Customer Effect and Highway Access

The location of customers significantly influences the spatial decisions of industries. Marshall (1890) notes that customer preferences and convenience significantly influence shop locations. He observes that customers are willing to travel farther to access specialized shops for essential goods but prefer nearby stores for everyday purchases. As a result, shops offering high-value products tend to cluster together, while convenience stores remain dispersed. This clustering allows businesses to tap into a larger shared customer base. Weber (1909) suggests that when materials, such as wood and grain, are widely available, transportation costs to customers become crucial. To reduce these costs, firms typically select locations that are close to their customers. Certain industries, such as those producing perishable goods (e.g., fresh food, dairy), service industries (e.g., restaurants) and retail stores, need to be close to their customers to ensure timely delivery and maintain quality (Marshall 1890). Similarly, producers of heavy materials like concrete and bricks often locate near construction sites to reduce transportation costs. These industries typically fall under the category of downstream industries, which produce goods and distribute them to customers.

Due to the dispersed nature of customer locations, which can include both individual consumers and other industries, capturing the agglomeration effects of customers is inherently complex. Empirical literature related to industrial agglomeration has not yet developed a variable specifically designed to capture the effects of customers. To address this gap and explore how highway access influences agglomeration in industries already affected by customer locations, this study uses downstream industries to capture the influence of customers. When highway access improves, firms can be located away from their customers and be able to locate together to obtain the benefits of agglomeration. This study hypothesises that for downstream industries, the effects of highway access on agglomeration are larger.

3.3.4 Input-Output adjusted Highway Access

Industries rely on highways for transporting goods both to and from their suppliers and customers. When the transportation of goods to suppliers and customers is facilitated by improved infrastructure, it reduces the overall transportation costs for reaching other industries or the market. Firms aim to maximise their profits and are thus inclined to locate in areas that offer the greatest profits. This inclination often results in localisation, allowing firms obtain economic externalities of agglomeration, such as improved bargaining power, enhanced knowledge sharing, and greater labour pooling (Marshall 1890). Therefore, reduced transportation costs contribute to increased within-industry agglomeration.

To assess the extent to which industry uses highways for transporting goods from suppliers and to customers, it is possible to construct an input-output based measure of highway access. However, no existing literature has specifically investigated size-adjusted highway access yet. Typical measures of highway access often overlook the variation in industry sizes and the differing volumes of goods transported between industries. As the effects of highway access are homogenous across all industries, this study introduces a novel variable for size-adjusted highway access, derived from input-output tables. This variable represents the weighted sum of highway access based on inputs and outputs among one industry and all other industries. By accounting for the connections between industries via highways, this approach provides a more accurate measure of highway utilisation.

Highways simplify the transport of inputs and outputs, potentially incentivising firms within the same industry to concentrate together. This spatial concentration allows them to benefit from agglomeration externalities and efficiently transport inputs or outputs over long distances to other industries through improved highway access. Therefore, it is hypothesised that input-output size-adjusted highway access positively impacts within-industry agglomeration.

3.4 Data and Methodology

3.4.1 Data and Sample

This study uses a micro-level dataset: the Annual Survey of Industrial Firms (ASIF), which is collected by the National Bureau of Statistics of China based on the annual survey and reports submitted by firms. Firm-level data on the manufacturing industries from 1998 to 2007, which includes employment, firm address, industry classification, etc. are employed in this study.

The ASIF captures industrial enterprises with sales of more than 5 million yuan (around 730,000 US dollars) over the period 1998 to 2007. The ASIF contains incomplete and rough original data. The data cleaning for the ASIF includes 1) deleting companies with missing company ID, revenue, total assets, net fixed asset, number of employees, gross output, district code and firm address; 2) deleting firms whose total assets, fixed assets, output, or paid-in-capital equal zero or less than zero and deleting firms whose number of employees$<$ 8; 3) deleting firms whose total assets are less than current assets, total assets are less than net fixed assets, or accumulated depreciation are less than current depreciation; 4) selecting firms in the manufacturing sector.

Industry code concordances. The industrial classification system in China was modified once over the period 1998-2007. Firms used industrial classification system GB/T 4754-1994 over the period 1998-2002 and GB/T 4754-2002 over the period 2003-2007. There are cases of several new 4-digit industry codes corresponding to one old 4-digit code, and several old 4-digit industry codes corresponding to one new 4-digit industry code. The industry concordances by Brandt et al. (2012) solve this problem by generating 4-digit industry adjustment codes for the old and new industry codes. Thus, this research adopts that list to make the industry code consistent over the sample period.

Region code concordances. The region code in China changed over the period 1998-2007. The county-level codes contain 6 digits; the first 4 digits represent the city-level code, and the first 2 digits are the province-level code. The change of existing county to city, the combination of existing counties and the establishment of new counties all result in the transformation

of county codes over time. Additionally, firms may report the old region code to the ASIF during the sample period. To address these problems, following the idea of Brandt et al. (2012) who construct industry adjustment codes, this study makes a regional concordance list to make the regional code consistent over the sample period. The adjustment county codes are generated and the corresponding county codes (codes for the same place) are converted to adjustment county codes.

Additionally, China input-output table is used, which is mainly for 3-digit manufacturing industries after being converted to the GB2002 industry classification.

The National Trunk Highway System plan approved by the State Council in 1992 is the 5-7 plan (5 north-south and 7 east-west routes). The targeted cities are all provincial capitals, cities with urban population of more than 500,000 and border crossings. This plan was completed ahead of schedule (2020) by the end of 2007. The State Council later approved an follow-up print for the National Expressway Network (NEN) in 2004, which is also called the 7-9-18 plan, with 7 radial expressways from the capital, 9 north-south routes and 18 east-west routes. The aim of 7-9-18 is to connect cities with an urban registered population of over 200,000.

The GIS highway routes are obtained from the ACASIAN Dataset, which is also used by Faber (2014) when investigating the effect of highway networks on economic activity in peripheral regions. The annual GIS highway routes from 1998-2007 are used in this research.

3.4.2 Key Variables

Within-industry agglomeration measures

This study adopts the Ellison and Glaeser agglomeration index for the measurement of within-industry agglomeration. Ellison & Glaeser (1997) develop a model that analyzes spatial concentration, which has been adopted by many studies. This index is based on a theoretical model developed by Ellison and Glaeser in 1997, which is designed to measure the geographic concentration of industries as a result of profit-maximizing location choices made by individual plants. The model assumes that the concentration of an industry arises from a combination of natural advantages and industry-specific spillovers. Natural advantages refer

to factors that make certain locations more suitable for specific industries, such as favourable climates for agriculture. Industry-specific spillovers refer to the benefits that firms receive from locating near other firms in the same industry, such as shared labour markets and knowledge spillovers.

In the model, each firm in an industry chooses its location to maximize its profits. The profitability of locating in a particular area r is influenced by natural advantages, spillovers, and a random component specific to each firm. The profit function for firm k choosing location r is given by:

$$\log \pi_{kr} = \log \overline{\pi}_r + g_r(v_1, \ldots, v_{k-1}) + \varepsilon_{kr}, \tag{3.1}$$

where $\overline{\pi}_r$ is the average profitability of area r, g_r represents the effect of spillovers, and ε_{kr} is an additional random component.

The EG index quantifies the spatial concentration of an industry by comparing observed concentration with a benchmark of random distribution. This involves two key measures. Firstly, the Gini Index (G_i) measures the inequality in the distribution of industry employment across geographic units. It is calculated as:

$$G_i \equiv \sum_r (x_r - s_r^i)^2, \tag{3.2}$$

where x_r is the share of total employment in region r, and s_r^i is the share of employment in industry i in region r. The Gini index is zero when an industry's share of employment in the region r is equal to the share of total employment of all industries in region r. The Gini index is larger when there are a smaller number of firms and a larger size of firms, even if locations are chosen completely at random.

Secondly, the Herfindahl Index (H_i) measures the concentration of firm sizes within the industry. It is calculated as:

$$H_i \equiv \sum_j z_j^2, \tag{3.3}$$

where z_j is the share of employment of firm j in industry i. A higher Herfindahl index indicates a less competitive industry. H_i is between 0 to 1. When the Herfindahl index is 1, it means the employment share of a firm in its industry is 1, which indicates only one firm in that industry.

The EG Index (γ_i) combines these elements to provide a measure of agglomeration that adjusts for the size of the regions and the structure of the industry. It is given by:

$$\gamma_i \equiv \frac{G_i - (1 - \sum_r x_r^2) H_i}{(1 - \sum_r x_r^2)(1 - H_i)}, \tag{3.4}$$

where G_i is the Gini index of the industry's employment distribution, $(1 - \sum_r x_r^2)$ adjusts for the size distribution of regions, and H_i is the Herfindahl index of firm sizes within the industry. The detailed formation of the model is provided in the Appendix.

The value of γ_i indicates the level of spatial concentration. γ_i takes a value of zero, meaning that the industry chooses a random location. A positive γ_i means industries are in excess spatial concentration, while a negative γ_i means that the industry chooses to disperse spatially. A negative EG value appears when the value of G_i is smaller than H_i.

Considering the choice of region r, the EG index can be computed at different geographic levels, such as province, city, and county, to capture varying degrees of spatial concentration. Ellison & Glaeser (1997) calculate the EG index at three geographic levels, including counties, states, and regions in the US, to capture agglomeration patterns across different scales. Similarly, this study calculates the EG index at the province, city, and county levels to account for China's vast and diverse geographic landscape. This multi-level approach is essential for understanding how industrial agglomeration varies across different geographic scales, as concentration patterns may differ significantly depending on the level of aggregation. By estimating γ_i at these three levels, the study ensures a more detailed and accurate analysis of industrial clustering.

This chapter uses data for firms' employment number, the industry categories classification (2-, 3-, and 4-digit), and the regional code (county, city and province level) to compute the EG index. The EG index calculated by employment is the main measure of agglomeration. In addition to the EG index, this study also adopts the Gini index for the robustness check. The Gini index takes the value from zero to one, with zero meaning industry i is equally distributed and close to one meaning the concentration degree of industry i is high.

Highway Access Variable

In this research, the weighted distance between each industry and highway is used to compute the explanatory variable-highway access. The formula for the highway access variable is:

$$highway\ access_i = \frac{1}{distance_{industry_i,highway}} \tag{3.5}$$

where the $distance_{industry,highway}$ is the weighted mean of Euclidean distance between firms of each industry and the nearest highways. The weights are calculated by the share of employment of firms in the industry. The distance from the firm to its nearest highway are computed by ArcGIS. The formula for the weighted distance between industry i and the highway network is as follows:

$$distance_{industry_i,highway} = \sum_{k=1}^{n} distance_{firm_k,nearest\ highway} * weight_k \tag{3.6}$$

where firm k is in industry i. The formula for $weight_k$ calculated by employment is: $weight_k = \frac{employment\ in\ firm_k}{employment\ in\ industry_i}$

The main explanatory variable is highway access. The other form of highway variables used for robustness tests are:

$$logdistance = ln(distance_{industry_i,highway}) \tag{3.7}$$

$$(logdistance)^{-1} = \frac{1}{ln(distance_{industry_i,highway})} \tag{3.8}$$

First, the logarithm of the weighted average distance between industry i and highway is used for robustness tests. The average weighted distances are positive and large. Using logs narrows the range and mitigates the distributions that are skewed. The variable $(logdistance)^{-1}$ is the reciprocal of the $logdistance$.

Input-Output adjusted Highway Access Measurement

The above distance from industry i to the nearest highway, D_i, is

$$D_i = \sum_k w_k dist_k \tag{3.9}$$

where $dist_k$ is the distance from firm k to its nearest highway. w_k is the share of firm k in its industry. This distance measure does not account for variations in industries' actual usage of highways. To address this limitation, the study introduces a size-adjusted highway access measure, termed Input-Output adjusted Highway Access (IOHA). The IOHA quantifies the extent to which each industry relies on highways for transporting inputs and outputs. Figure 3.1 illustrates the input-output relationships between industries. (I) represents that industry i receives inputs from industry j, and (II) represents the output from industry i to industry j. If the input and output sizes are ignored, the highway usage would not be accurate.

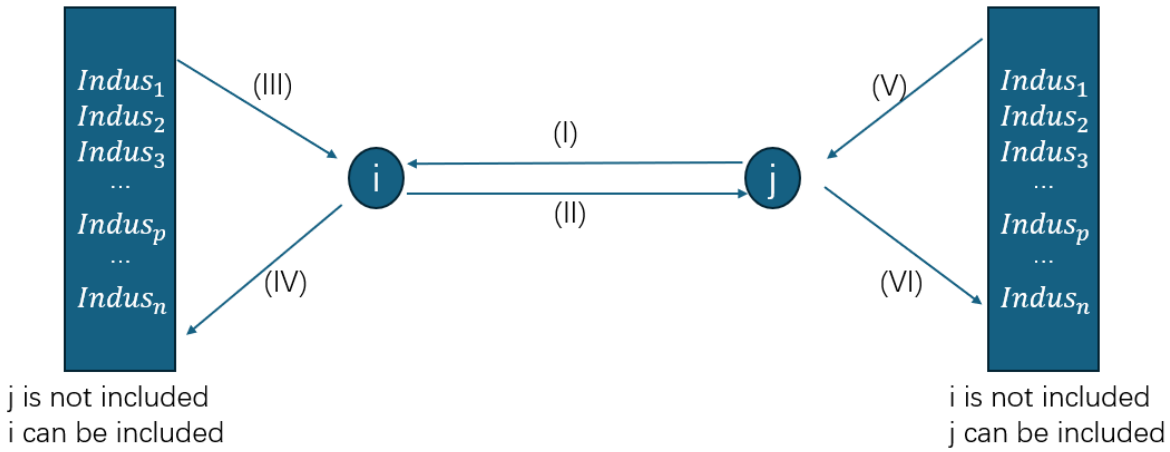


Figure 3.1: Input-Output adjusted Highway Access description

Note: The figure illustrates the input-output relationships among industries. Here, i and j denote industry i and industry j, respectively. The notation (I) indicates that industry i receives inputs from industry j, while (II) signifies the output from industry i to industry j.

The IOHA is defined by the following equation.

$$IOHA_i = \sum_p (s_{i\to p} + s_{p\to i}) * (-\frac{D_i + D_p}{2}) \tag{3.10}$$

where p denotes industry p. Here p can be equal to i. $s_{i\to p}$ is the size of industry i's output to p. $s_{p\to i}$ is the size of industry i's input from p. $s_{i\to p}$ can be the value of industry p's input (x_{ip}) received from industry i, which directly indicates how much input is transported between two industries. Note that in the denominator $\sum_p x_{ip}$, x_{ii} should be double counted. The diagonal of the input-output table that reflects the inputs from the own industry should be double counted in this situation.

The instrumental variables for IOHA are also based on the input-output matrix. For example, the IV of least cost path routes is

$$IOLCP_i = \sum_p (s_{i\to p} + s_{p\to i}) * (-\frac{D_i^{LCP} + D_p^{LCP}}{2}) \tag{3.11}$$

where $IOLCP_i$ is IV for $IOHA_i$. The D_i^{LCP} is the distance from industry i to its nearest least cost path. $s_{i\to p}$ is the size of industry i's output to industry p. $s_{p\to i}$ is the size of industry i's input from industry p.

3.4.3 Model Specification

The baseline regression model for agglomeration is:

$$\gamma_{it} = \alpha + \beta_1 highway\ access_{it} + \sigma X_{it} + \delta_i + \varepsilon_{it} \tag{3.12}$$

where γ_{it} is the EG within-industry agglomeration level of industry i at time t. $highway\ access_{it}$ is the reciprocal of the weighted distance between industry i and the highway. The distance is the weighted mean of Euclidean distance between firms of each industry and their nearest highways. The weights are calculated by the share of employment of firms in the industry. α_i is the constant term or intercept. δ_i is the industry effect. ε_{it} is the error term.

The control variables X_{it} include Marshall's three mechanisms of agglomeration (input sharing, labour pooling and knowledge spillovers), natural advantages and upstreamness level. In the regression, the intermediate input share is the proxy for input sharing; net labour productivity is the proxy for labour market pooling, new product ratio is the proxy for knowledge spillover. The natural advantages have three proxies, which are the agricultural products input share, the fishing products input share and the coal mining products input share, which are calculated by data from the 2002 China Input-Output table. The upstreamness level is calculated by method of Antràs et al. (2012), which is also computed with data from the 2002 China Input-Output table.

Input sharing. Rosenthal & Strange (2001) use Manufactured inputs per $ of shipment and nonmanufactured inputs per $ of shipment (e.g. legal, financial services and insurance) as two proxies for input sharing. Purchased-inputs intensity is used by Lu & Tao (2009) and Holmes (1999) as a proxy for input sharing. It is defined as the ratio of purchased inputs to total output. Jofre-Monseny et al. (2011) use the share of the inputs that one industry purchases from other industries to characterize customer-supplier relations (data are from the Catalan Input-output Table). Lu & Tao (2009) claim that the purchased inputs share reflects the degree of input sharing. Following Lu & Tao (2009), the intermediate inputs share is the proxy for input sharing, which is computed with ASIF data. It is expected to have a positive effect on industrial agglomeration. More volume of input share means that there is more possibility of input sharing activities within the industry.

Upstreamness level. The upstreamness or downstreamness level also indicates the input sharing effect to some extent. An abundant supply of specialized inputs can facilitate geographic concentration of downstream firms (Marshall, 1920). Downstream industries have more input sharing opportunities and thus they are expected to be more agglomerated. The upstreamness level is calculated by the method of Antràs et al. (2012), which is also computed using data from the 2002 China Input-Output table at the 3-digit level. The logic is that when the industry enters the indirect value chain, if the number of stages from final-use production is larger, then its weight for upstreamness is higher. A higher value of the upsteamness measure (U), means there is a higher level of upstreamness.

Labour market pooling. The three proxies in Rosenthal & Strange (2001) for labour market pooling are as follows. Net productivity is the proxy, which is the value of shipments less the value of purchased inputs, divided by employment in the industry. The second is the ratio of management workers, which is the number of management workers divide by the sum of management workers and production workers. The third is worker education, that is the different types of degrees of workers. Lu & Tao (2009) construct a proxy called wage premium, which is the wage premium of an industry over the average wage in a region and then weighted average over all regions. They indicate that a higher wage premium means a higher skill level is required in an industry and labour market pooling is more needed for higher skill level industries.

In this study, net labour productivity is the proxy for labour market pooling. It is calculated as the value-added divided by the number of employees. Rosenthal & Strange (2001) use labour productivity as one of the proxies for labour pooling. Higher labour productivity indicates there are more skilled workers in the industry. More skilled workers in a place can attract other firms which need those skills to that place, and thus leads to a higher level of agglomeration.

Knowledge spillovers. Innovation per $ of shipment is the proxy for knowledge spillover in Rosenthal & Strange (2001). Innovations are captured by the number of new products in 1982. They reject the patent because an innovation can be related to hundreds of patents and US patent data are based on product type instead of industry, so it is difficult to match these two systems. R&D expenditure is not employed by Rosenthal & Strange (2001) either, as they indicate that many innovations come from practice rather than R&D and R&D expenditure is related to knowledge spillover in input rather than output.

Lu & Tao (2009) use new products to output ratio as the proxy for knowledge spillovers. They indicate that the new products in the ASIF database are produced for the first time at least within a province. It is likely to reflect imitation of other regions or countries. The results of Lu & Tao (2009) show that the new product ratio significantly increases the agglomeration level in the pooled OLS but is much weaker in FE. The new product ratio is the proxy for knowledge spillover in this study, following Lu & Tao (2009).

Natural advantages. Ellison & Glaeser (1999*b*) use costs of 16 proxies to reflect the natural advantages, including six proxies for natural inputs, six for labour inputs and four proxies for transport costs. The natural inputs consist of electricity, natural gas, coal, agriculture, livestock and lumber. In their paper, the transport costs of export or import, and transport costs to the market are two types of variables to capture transport costs. They find about 20 percent of observed agglomeration can be explained by their set of advantages. Rosenthal & Strange (2001) control for natural advantages. They use input-output tables from the 1992 Bureau of Economic Analysis. The proxies for the importance of natural advantages are energy per $ shipment for energy input cost, natural resources per $ shipment for cost of natural resources and water per $ shipment measures water-related costs.

Lu & Tao (2009) use the ratio of agricultural product usage and the ratio of mining product usage as two proxies for resource endowments. They use China's 1997 input-output table covering 124 sectors which are between the 2-digit and 3-digit industrial classifications. They match the 124 sectors with the 3-digit industries to regress at 3-digit level. A concordance table of the 124 sectors with the 3-digit industries is used, which explains why their regression analysis is carried out at the 3-digit industry level. They find that the two proxies for natural advantages have positive and mostly statistically significant coefficients.

In this research, the natural advantages have four proxies: input share of agricultural products, fishing products, coal mining products and petroleum products, which are calculated using data from China Input-Output table. If there are more natural product inputs, it implies that the industry is more reliant on natural advantages. Industries that use more natural inputs are likely to be located near the natural resources and are more concentrated. Using the natural inputs ratio in regression can test whether an industry which uses more natural inputs is more concentrated. Thus, their coefficients are expected to be positive.

As a majority number of control variables use data from the IO table, which is mainly at 3-digit industry level, 3-digit industry level is used for regression. The data are an unbalanced panel with 161 3-digit industries in 1998-2000 and 2002-2007, and 157 industries in 2001.

3.5 Summary Statistics and Baseline Results

3.5.1 Summary Statistics of Baseline Variables

The EG index that captures the agglomeration level of each industry is the response variable. This study uses employment data to calculate the EG index, following Ellison & Glaeser (1997). The weighted means of the EG index are calculated by the sum of the EG index that is multiplied by the employment share of each industry. The weighted means are presented in Table3.2.

Table 3.2: Weighted means of the EG index in China's manufacturing industries

EG index in employment Industry and region	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
2-digit industry										
county	0.0020	0.0017	0.0019	0.0022	0.0025	0.0032	0.0040	0.0042	0.0043	0.0032
city	0.0043	0.0044	0.0049	0.0057	0.0063	0.0076	0.0092	0.0095	0.0095	0.0096
province	0.0169	0.0183	0.0204	0.0221	0.0242	0.0283	0.0316	0.0326	0.0323	0.0322
3-digit industry										
county	0.0047	0.0039	0.0043	0.0049	0.0057	0.0069	0.0080	0.0084	0.0084	0.0072
city	0.0094	0.0095	0.0105	0.0114	0.0130	0.0153	0.0176	0.0181	0.0181	0.0181
province	0.0331	0.0342	0.0374	0.0385	0.0432	0.0488	0.0529	0.0551	0.0545	0.0541
4-digit industry										
county	0.0066	0.0059	0.0068	0.0076	0.0090	0.0109	0.0120	0.0125	0.0126	0.0113
city	0.0125	0.0130	0.0147	0.0160	0.0186	0.0216	0.0243	0.0248	0.0250	0.0250
province	0.0409	0.0426	0.0467	0.0485	0.0549	0.0616	0.0665	0.0695	0.0695	0.0696

As shown in Table 3.2, the weighted means of the EG index of all industries in general increase from 1998 to 2006, while decreasing in 2007. The average level of agglomeration tends to increase over the sample period at county, city and province levels. As $G_i \equiv \sum_r (x_r - s_r^i)^2$, G_i is larger when a country is divided into fewer geographic regions and γ_i increases with G_i. The agglomeration degree at province level is the largest and at the county level is the smallest. The agglomeration degree of 4-digit industry is the largest and the 2-digit industry is the smallest, showing that narrower industry classification has a higher degree of agglomeration. For example, if all firms are in one industry category, $G_i = 0, H_i$ approaches zero, and the EG index is very close to zero.

Ellison and Glaeser define $\gamma > 0.05$ as highly spatially concentrated, $0.02 \leq \gamma \leq 0.05$ is somewhat concentrated, and $\gamma < 0.02$ is not very concentrated. $\gamma = 0$ means the location is randomly allocated; $\gamma > 0$ implies industries are in excess spatial concentration. Table 3.3 summarizes the agglomeration levels computed by employment from 1998 to 2007 and

compares them with that of previous literature, based on 4-digit industry and county level. The percentage of industries that are very concentrated almost doubled from 1998 to 2007, and the percentage of industries that are somewhat concentrated also increased almost 1.5 times. The percentages of the EG index for not very concentrated, somewhat concentrated and very concentrated in this study are similar to those of Lu & Tao (2009) who also investigate agglomeration in China. Additionally, Table 3.3 shows that the percentage of spatially concentrated 4-digit industries is lower in China than in the US and UK.

Table 3.3: Percentage of the EG index in employment at 4-digit industry and county level

	Country	year	Percentage of industries that are		
			Not very concentrated	Somewhat concentrated	Very concentrated
This study	China	1998	86.32%	10.14%	3.54%
		1999	88.65%	9.69%	1.65%
		2000	85.07%	11.85%	3.08%
		2001	84.88%	11.46%	3.66%
		2002	85.38%	10.14%	4.48%
		2003	80.90%	13.44%	5.66%
		2004	78.54%	15.80%	5.66%
		2005	76.42%	17.69%	5.90%
		2006	76.42%	17.22%	6.37%
		2007	77.36%	15.57%	7.08%
literature					
Lu&Tao(2009)	China	2005	75.98%	16.20%	7. 82%
Ellison&Glaeser (1997)	US	1987	10.00%	65.00%	25. 00%
Devereux et al. (2004)	UK	1992	65.00%	19.00%	16. 00%

Note: $EGindex < 0.02$ is not very concentrated; $0.02 \leq EGindex \leq 0.05$ is somewhat concentrated; $EGindex > 0.05$ as highly spatially concentrated.

Table 3.4 shows the top ten concentrated 3-digit industries at province, city and county levels respectively. The EG index values are the average EG values for 3-digit industries from 1998 to 2007. The industry code in Table 3.4 is from the GB/T 4754-2002 classification. The top ten concentrated industries at province, city and county levels are not the same but there are some overlaps, which shows that industries that are concentrated at one geographic level have a higher possibility of being concentrated at the other two geographic levels. Additionally, taking the province level as an example, the top ten concentrated 3-digit industries are from different sectors, as their broader industry categories (2-digit industries) are different. Coking, sugar and the aquatic product industry rely more on natural resources; toy manufacturing and stationery manufacturing industries may be labour-intensive industries; electronic computer manufacturing may need more skilled workers and a higher level of technology.

Table 3.4: Top ten concentrated 3-digit industries at province, city and county levels

Top ten concentrated industries at province level Industry name	Industry code	EG index (province)
Coking	252	0.2682
Toy manufacturing	244	0.2304
Home audio-visual equipment manufacturing	407	0.2277
Sugar manufacturing	134	0.2069
Watch and timing instrument manufacturing	413	0.1913
Cultural and office machinery manufacturing	415	0.1706
Silk spinning and finishing	174	0.1628
Aquatic product processing	136	0.1460
Other plastic products manufacturing	309	0.1309
Refractory product manufacturing	316	0.1247
Top ten concentrated industries at city level **Industry name**	**Industry code**	**EG index (city)**
Motorcycle manufacturing	373	0.0839
Aquatic product processing	136	0.0757
Rare rare earth metal smelting	333	0.0756
Electronic computer manufacturing	404	0.0581
Silk spinning and finishing	174	0.0571
Home audio-visual equipment manufacturing	407	0.0570
Refractory product manufacturing	316	0.0532
Watch and timing instrument manufacturing	413	0.0484
Toy manufacturing	244	0.0431
Stationery manufacturing	241	0.0420
Top ten concentrated industries at county level **Industry name**	**Industry code**	**EG index (county)**
Aquatic product processing	136	0.0352
Home audio-visual equipment manufacturing	407	0.0344
Silk spinning and finishing	174	0.0327
Toy manufacturing	244	0.0307
Refractory product manufacturing	316	0.0276
Rare earth metal smelting	333	0.0273
Electronic computer manufacturing	404	0.0255
Amusement equipment and entertainment products manufacturing	245	0.0250
Ceramic products manufacturing	315	0.0224
Bicycle manufacturing	374	0.0215

The weighted distance from industry to highways is used to compute the highway access variable. Table 3.5 shows the summary statistics of weighted distance from 3-digit industries to the highway. The unit of distance in the table is metre. It can be found that the weighted distance gradually reduces over the period.

Table 3.5: The summary statistics of weighted distance from industry to highway

year	mean	sd	min	max
1998	26266	13403	4693	74828
1999	24948	13766	6141	86722
2000	19529	10932	4938	64145
2001	17913	10050	4240	60614
2002	15071	9135	4449	55517
2003	17435	12204	3923	75451
2004	13674	9326	3602	48158
2005	10563	7027	3502	44759
2006	10312	6478	3582	42473
2007	9964	6754	3199	45868
Total	16564	11602	3199	86722

Note: This table shows the weighted distance from 3-digit industries to the highway network from 1998 to 2007. The unit here is metre.

Summary statistics of variables used in the baseline model are displayed in Table 3.6. EG indices are calculated at county, city and province levels. Highway access is the research focus and a larger value represents more highway accessibility.

Table 3.6: Summary statistics of variables used in the baseline model

Variable	Obs	Mean	Std. Dev.	Min	Max
EG index(county)	1,606	0.049	0.051	-0.086	0.327
EG index(city)	1,606	0.016	0.017	-0.023	0.137
EG index(province)	1,606	0.007	0.009	-0.014	0.085
highway access	1,606	0.090	0.055	0.012	0.313
intermediate ratio	1,606	0.703	0.106	0.279	1.000
net labour productivity	1,606	73.601	54.282	0.000	391.572
new product ratio	1,606	0.053	0.041	0.000	0.336
agriculture input share	1,603	0.035	0.097	0.000	0.645
fishing input share	1,603	0.004	0.034	0.000	0.525
coal input share	1,603	0.008	0.040	0.000	0.553
petroleum input share	1,603	0.008	0.059	0.000	0.657
upstreamness	1,603	2.859	1.114	1.121	6.588

3.5.2 The Baseline Results

The pooled OLS and fixed effect estimators are used to estimate the baseline model. As shown in Table 3.7, columns (1), (2) and (3) show the results of pooled OLS for the highway access variable at province, city and county levels respectively. Columns (4)-(6) and (7)-(9) show the pooled OLS results for *logdistance* an $(logdistance)^{-1}$ respectively, for the robustness check. The estimated coefficient of the highway access variable is statistically significant with positive signs of pooled OLS at all three geographic levels, meaning that improved highway access is associated with a higher level of agglomeration for manufacturing industries.

Comparing the estimated coefficients of highway access at three geographic levels, the coefficients are largest at province level and smallest at county level. The results indicate that an increase in highway access by 0.1 (with highway access ranging from 0.012 to 0.313 in the sample) is associated with increases in the EG indices of 0.033, 0.011, and 0.006, respectively, while holding other factors constant. Table 3.2 in the above section shows that the weighted means of the EG index increase greatly as the geographic level increases from county level to province level, as from county to province level, a country is divided into fewer geographic regions, and G_i used to calculate the EG index is larger, which leads to a higher EG index value. Thus, in Table 3.7, the estimated result for the EG index at province level is much larger than that at county level, which does not imply that highway access has more impact on within-industry agglomeration levels at province level than city and county level.

The estimated coefficients of highway access and $(logdistance)^{-1}$ (for the robustness test) are expected to be positive and the estimated coefficients of *logdistance* (for the robustness test) are expected to be negative, that is, the increase in the weighted average distance from industry to highway increases is associated with the decrease in agglomeration level.

The FE estimation results for highway access, *logdistance* and $(logdistance)^{-1}$ are shown in Table 3.8 in Columns (1)-(3), (4)-(6) and (7)-(9), respectively. The coefficients estimated by FE for all highway variables are statistically significant. An increase in highway access by 0.1 is associated with increases in the EG indexes of 0.011, 0.008 and 0.003 at province, city and county levels, respectively, holding other factors fixed. The estimated coefficients

of $logdistance$ and $(logdistance)^{-1}$ have expected signs and are consistent with that of the pooled OLS estimation. Regarding $logdistance$, the results indicate that the weighted distance from industry i to a 1% increase in the distance to highways is associated with a decrease in the EG index value by 0.00013, 0.00006, and 0.00003 at the province, city, and county levels respectively, while holding other factors constant.

Regarding the performance of Marshall's three mechanisms, the coefficients for intermediate ratio and net labour productivity (proxies for input sharing and labour pooling respectively) are positive and statistically significant at province level with the fixed effect estimate. The new product ratio has positive but statistically insignificant coefficients in the FE estimation.

With respect to the preformance of natural advantages, agriculture input share does not show a significant association with economic growth at any regional level. In contrast, fishing product input share is positively and significantly associated with within-industry agglomeration at the city and county levels. Coal and petroleum input shares are strongly associated with higher levels of agglomeration at the province level, with petroleum also showing a positive association at the city level. The overall positive and significant coefficients of these natural resources variables mean that industries which rely more on natural resources are more agglomerated as they need to be located near them. This is consistent with the findings by Lu & Tao (2009) and Ellison & Glaeser (1999*a*).

Regarding the year effect, preliminary analyses have shown that the inclusion of year fixed effects leads to insignificant results. This is likely because the year-fixed effects capture broad temporal trends, including the general increase in agglomeration over time due to infrastructure developments. Since these broad trends are not the primary focus of the study, their inclusion could dilute the effect of the factors directly influencing agglomeration, making it difficult to discern the actual impact of highways. The approach taken aligns with the methodology used in related studies, such as Lu & Tao (2009), which investigates the effect of local protectionism on agglomeration in China excluding year effects.

Table 3.7: Pooled OLS results for three forms of highway variables

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
highway access	0.3326***	0.1148***	0.0591***						
	(0.0785)	(0.0243)	(0.0118)						
logdistance				-0.0214***	-0.0070***	-0.0043***			
				(0.0069)	(0.0025)	(0.0011)			
$(logdistance)^{-1}$							0.1650***	0.0568***	0.0296***
							(0.0402)	(0.0128)	(0.0061)
intermediate_ratio	-0.0127	-0.0027	-0.0021	-0.0134	-0.0027	-0.0026	-0.0137	-0.0031	-0.0023
	(0.0473)	(0.0161)	(0.0070)	(0.0501)	(0.0172)	(0.0074)	(0.0477)	(0.0163)	(0.0070)
net_labour_productivity	-0.0001	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0001	-0.0000	-0.0000
	(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0000)	(0.0000)
new_product_ratio	-0.1870**	-0.0339	-0.0180	-0.1635*	-0.0244	-0.0156	-0.1870**	-0.0338	-0.0182
	(0.0859)	(0.0209)	(0.0142)	(0.0918)	(0.0223)	(0.0152)	(0.0869)	(0.0212)	(0.0143)
agriculture	0.1048*	0.0074	-0.0037	0.0984*	0.0043	-0.0036	0.1060*	0.0078	-0.0034
	(0.0548)	(0.0130)	(0.0058)	(0.0574)	(0.0146)	(0.0065)	(0.0553)	(0.0132)	(0.0059)
fishing	0.1707***	0.1179***	0.0571***	0.1574***	0.1128***	0.0553***	0.1691***	0.1173***	0.0569***
	(0.0423)	(0.0142)	(0.0061)	(0.0445)	(0.0152)	(0.0064)	(0.0425)	(0.0143)	(0.0061)
coal	0.2106	0.0202	0.0074	0.2005	0.0164	0.0059	0.2088	0.0196	0.0071
	(0.1848)	(0.0215)	(0.0062)	(0.1843)	(0.0214)	(0.0061)	(0.1845)	(0.0214)	(0.0061)
petroleum	0.0154	0.0021	0.0006	0.0115	0.0008	-0.0001	0.0144	0.0018	0.0004
	(0.0272)	(0.0093)	(0.0035)	(0.0270)	(0.0092)	(0.0035)	(0.0274)	(0.0093)	(0.0035)
upstreamness	-0.0006	-0.0006	-0.0011**	-0.0008	-0.0007	-0.0011**	-0.0005	-0.0006	-0.0010**
	(0.0034)	(0.0011)	(0.0005)	(0.0036)	(0.0011)	(0.0005)	(0.0034)	(0.0011)	(0.0005)
_cons	0.0392	0.0117	0.0085	0.1224**	0.0392**	0.0250***	0.0019	-0.0011	0.0018
	(0.0363)	(0.0126)	(0.0058)	(0.0513)	(0.0188)	(0.0084)	(0.0358)	(0.0123)	(0.0056)
N	1603	1603	1603	1603	1603	1603	1603	1603	1603
r2_a	0.132	0.147	0.155	0.091	0.097	0.126	0.128	0.142	0.153

Note: Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 3.8: FE estimates results for three forms of highway variables

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
highway access	0.1068**	0.0756***	0.0316***						
	(0.0430)	(0.0189)	(0.0089)						
logdistance				-0.0129***	-0.0063***	-0.0027***			
				(0.0041)	(0.0017)	(0.0009)			
$(logdistance)^{-1}$							0.0594***	0.0397***	0.0165***
							(0.0222)	(0.0098)	(0.0046)
intermediate_ratio	0.0522*	0.0101	0.0042	0.0522*	0.0104	0.0043	0.0523*	0.0102	0.0043
	(0.0281)	(0.0114)	(0.0063)	(0.0284)	(0.0116)	(0.0064)	(0.0281)	(0.0114)	(0.0063)
net_labour_productivity	0.0002***	0.0000	0.0000	0.0001***	0.0000	0.0000	0.0001***	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
new_product_ratio	0.0396	0.0682	0.0207	0.0345	0.0670	0.0201	0.0390	0.0680	0.0206
	(0.1106)	(0.0475)	(0.0291)	(0.1113)	(0.0481)	(0.0294)	(0.1105)	(0.0475)	(0.0291)
agriculture	0.0022	-0.0073	-0.0110	0.0064	-0.0045	-0.0098	0.0028	-0.0067	-0.0108
	(0.0261)	(0.0083)	(0.0076)	(0.0244)	(0.0079)	(0.0075)	(0.0259)	(0.0082)	(0.0075)
fishing	-0.0340	0.0180***	0.0161***	-0.0303	0.0192***	0.0166***	-0.0335	0.0182***	0.0162***
	(0.0246)	(0.0055)	(0.0035)	(0.0207)	(0.0041)	(0.0027)	(0.0241)	(0.0052)	(0.0033)
coal	0.2357***	0.0353	-0.0015	0.2114***	0.0256	-0.0058	0.2322***	0.0333	-0.0023
	(0.0530)	(0.0218)	(0.0102)	(0.0444)	(0.0179)	(0.0085)	(0.0520)	(0.0211)	(0.0099)
petroleum	0.2258***	0.0401**	-0.0013	0.2029***	0.0316**	-0.0051	0.2223***	0.0382**	-0.0020
	(0.0451)	(0.0189)	(0.0088)	(0.0385)	(0.0159)	(0.0076)	(0.0443)	(0.0184)	(0.0086)
upstreamness	0.0047	0.0003	0.0008	0.0036	-0.0002	0.0006	0.0046	0.0002	0.0008
	(0.0038)	(0.0012)	(0.0006)	(0.0039)	(0.0013)	(0.0006)	(0.0038)	(0.0012)	(0.0006)
_cons	-0.0275	-0.0048	-0.0026	0.0205	0.0193	0.0078	-0.0417*	-0.0142	-0.0064
	(0.0239)	(0.0107)	(0.0055)	(0.0325)	(0.0140)	(0.0070)	(0.0227)	(0.0106)	(0.0054)
N	1603	1603	1603	1603	1603	1603	1603	1603	1603
r2_a	0.165	0.160	0.079	0.172	0.144	0.073	0.167	0.162	0.079

Note: Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. This table shows the FE results without the year effect. When adding the year effect, the results are not significant because highways expand and the trend of agglomeration increases, the year effect explains the increase in the level of agglomeration, which makes no economic sense. Additionally, Lu & Tao (2009) do not include the year effect when investigating the effect of local protectionism on agglomeration.

3.5.3 Robustness Tests for the Baseline Model

Gini index. The Gini index is used as a robustness check for the EG index. Table 3.9 illustrates the pooled OLS and FE estimates results for the Gini index. The highway access variable has positive and significant coefficients with pooled OLS and less significant coefficients with FE, which indicates that in general, improved highway access is associated with spatial concentration calculated using the Gini index.

Table 3.9: Pooled OLS and FE results for the Gini index calculated by employment

	Pooled OLS			FE		
	Gini (province) (1)	Gini (city) (2)	Gini (county) (3)	Gini (province) (4)	Gini (city) (5)	Gini (county) (6)
highway access	0.3148***	0.1276***	0.0731***	0.0797*	0.0653**	0.0230
	(0.0748)	(0.0349)	(0.0268)	(0.0447)	(0.0278)	(0.0264)
intermediate_ratio	-0.0171	-0.0095	-0.0086	0.0443	0.0050	-0.0009
	(0.0457)	(0.0222)	(0.0160)	(0.0270)	(0.0182)	(0.0196)
net_labour_productivity	-0.0001	-0.0001*	-0.0001**	0.0001**	-0.0000	-0.0000
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
new_product_ratio	-0.1948**	-0.0581	-0.0426	-0.0687	-0.0420	-0.0877
	(0.0962)	(0.0472)	(0.0464)	(0.1000)	(0.0655)	(0.0561)
agriculture	0.0873*	-0.0018	-0.0129	-0.0026	-0.0118	-0.0156*
	(0.0509)	(0.0151)	(0.0102)	(0.0241)	(0.0099)	(0.0091)
fishing	0.1443***	0.1025***	0.0428***	-0.0306	0.0210***	0.0193***
	(0.0394)	(0.0162)	(0.0102)	(0.0187)	(0.0043)	(0.0041)
coal	0.1875	0.0155	0.0026	0.1859***	0.0125	-0.0236*
	(0.1613)	(0.0212)	(0.0173)	(0.0486)	(0.0222)	(0.0140)
petroleum	0.0117	0.0011	-0.0006	0.1695***	0.0085	-0.0320**
	(0.0304)	(0.0178)	(0.0129)	(0.0413)	(0.0192)	(0.0132)
upstreamness	-0.0009	-0.0010	-0.0014	0.0034	-0.0003	0.0002
	(0.0035)	(0.0020)	(0.0017)	(0.0036)	(0.0014)	(0.0012)
_cons	0.0550	0.0317*	0.0284**	0.0005	0.0211	0.0236
	(0.0358)	(0.0184)	(0.0142)	(0.0221)	(0.0160)	(0.0174)
N	1603	1603	1603	1603	1603	1603
r2_a	0.115	0.068	0.037	0.092	0.027	0.019

Note: Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

EG index calculated by 2004 economy census data. The first economy census was conducted in 2004 in China and the second, third and fourth economic censuses were in 2008, 2013 and 2018, respectively. The economy census and industry census data have the advantage of containing small firms while the ASIF dataset records only firms with sales of more than 5 million yuan. Additionally, the ASIF is at the firm level but the industry census data is at the plant level. This study uses the 2004 economy census data to capture the agglomeration degree of manufacturing industries.

Table 3.10 shows the empirical results for the baseline model using the 2004 economy census data. In terms of the highway access variable, all coefficients are statistically significant and have expected signs, indicating that after including small firms, improved highway access is still associated with higher within-industry agglomeration levels. The estimated coefficients using the 2004 economy census data are larger than those using the ASIF dataset in Table 3.7, indicating that improved highway access is associated with more spatial concentration when small firms are included.

Table 3.10: OLS results with 2004 economy census data

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
highway access	1.0685***	0.4489***	0.3573***						
	(0.3361)	(0.1171)	(0.1075)						
logdistance				-0.0574**	-0.0274***	-0.0226***			
				(0.0229)	(0.0089)	(0.0080)			
$(logdistance)^{-1}$							0.5173***	0.2196***	0.1755***
							(0.1698)	(0.0587)	(0.0534)
intermediate_ratio	0.0102	0.0390	0.0373	0.0229	0.0425	0.0396	0.0107	0.0390	0.0373
	(0.0662)	(0.0533)	(0.0541)	(0.0740)	(0.0552)	(0.0555)	(0.0676)	(0.0536)	(0.0543)
net_labour_productivity	0.0001	0.0003	0.0003	0.0002	0.0003	0.0003	0.0001	0.0003	0.0003
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
new_product_ratio	-0.4210***	-0.1944**	-0.2016**	-0.3250**	-0.1663*	-0.1821**	-0.4103**	-0.1913**	-0.1995**
	(0.1593)	(0.0933)	(0.0921)	(0.1458)	(0.0926)	(0.0918)	(0.1585)	(0.0932)	(0.0920)
agriculture	0.1080*	-0.0168	-0.0297	0.0750	-0.0237	-0.0336	0.1074*	-0.0163	-0.0291
	(0.0574)	(0.0445)	(0.0454)	(0.0616)	(0.0459)	(0.0462)	(0.0588)	(0.0447)	(0.0455)
fishing	0.0836**	0.0295	-0.0022	0.0611	0.0229	-0.0069	0.0810**	0.0287	-0.0027
	(0.0383)	(0.0305)	(0.0304)	(0.0428)	(0.0322)	(0.0318)	(0.0391)	(0.0308)	(0.0306)
coal	0.7499***	0.2930***	0.1644**	0.6921***	0.2887***	0.1656**	0.7561***	0.2975***	0.1685**
	(0.1340)	(0.0736)	(0.0712)	(0.1542)	(0.0839)	(0.0784)	(0.1404)	(0.0753)	(0.0725)
upstreamness	-0.0062	-0.0067	-0.0057	-0.0059	-0.0066	-0.0056	-0.0062	-0.0067	-0.0056
	(0.0065)	(0.0044)	(0.0043)	(0.0073)	(0.0045)	(0.0044)	(0.0066)	(0.0044)	(0.0043)
_cons	-0.0211	-0.0365	-0.0391	0.1963**	0.0655*	0.0446	-0.1406*	-0.0874	-0.0798
	(0.0533)	(0.0486)	(0.0504)	(0.0903)	(0.0390)	(0.0369)	(0.0751)	(0.0588)	(0.0603)
N	161	161	161	161	161	161	161	161	161
r2_a	0.237	0.198	0.202	0.099	0.141	0.161	0.216	0.189	0.195

Note: Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

3.6 Endogeneity and IV Estimation

3.6.1 Identification and Endogeneity

Highway networks are not constructed randomly and are likely to be located where population density is higher. Some omitted factors are correlated with the highway access variable. Additionally, with industrial agglomeration and highway networks, it may be a case of reverse causality, as agglomeration may lead to the construction of highway routes. In order to address these endogeneity problems, this study uses three types of instrumental variables for the highway access variable, including the historical routes, least cost path minimum spanning tree network and the Euclidean straight line minimum spanning tree networks. Time-variant IVs are adopted to fit better with the highway independent variables and panel data analysis.

Instrumental Variables

Historical roads. Duranton & Turner (2012) first applied a historical route IV approach that used colonial routes or 'caminos reales' and the 1938 road network as time-invariant instruments. Later, Garcia-López et al. (2015), Martincus et al. (2017) and Baum-Snow et al. (2017) also use time-invariant historical roads as instruments. Baum-Snow et al. (2015) use the 1962 road network as an instrument for the 2010 highway network in China, based on the idea that the 1962 roads primarily served as connections from agricultural areas to nearby cities to move agricultural goods to local markets. The 1962 roads could be upgraded to modern highways at a lower cost than establishing new highways. They indicate that locations with more 1962 roads also had more highways in 2010. Baum-Snow et al. (2017) also use road and railways in 1962 as instruments for modern infrastructure.

Holl (2016) uses a time-variant historical road IV approach, following the strategy of Hornung (2015). He generates a 10 km corridor along the highways that open to traffic during a given year. Then, 1760 postal routes and Roman roads that fall within the 10 km corridor (buffer) along the highways are selected as instruments. Holl (2016) uses log (distance to the nearest highway) as the explanatory variable, and this study follows Holl's time-variant IV approach.

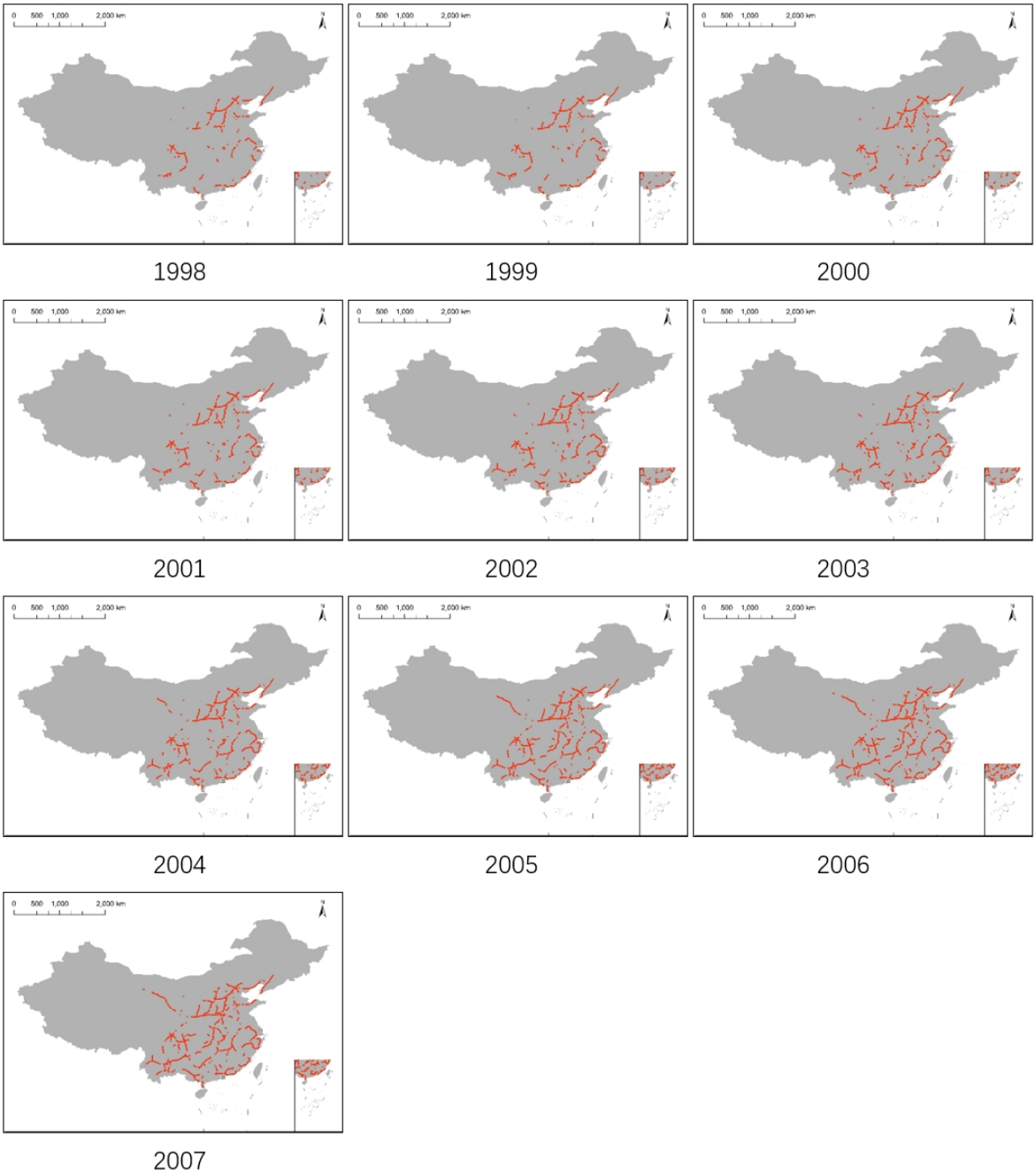


Figure 3.2: Intersections of 10km highway buffers and Ming routes from 1998 to 2007
Note: The network in red depicts the intersections of 10km highway buffers and Ming routes from 1998 to 2007.

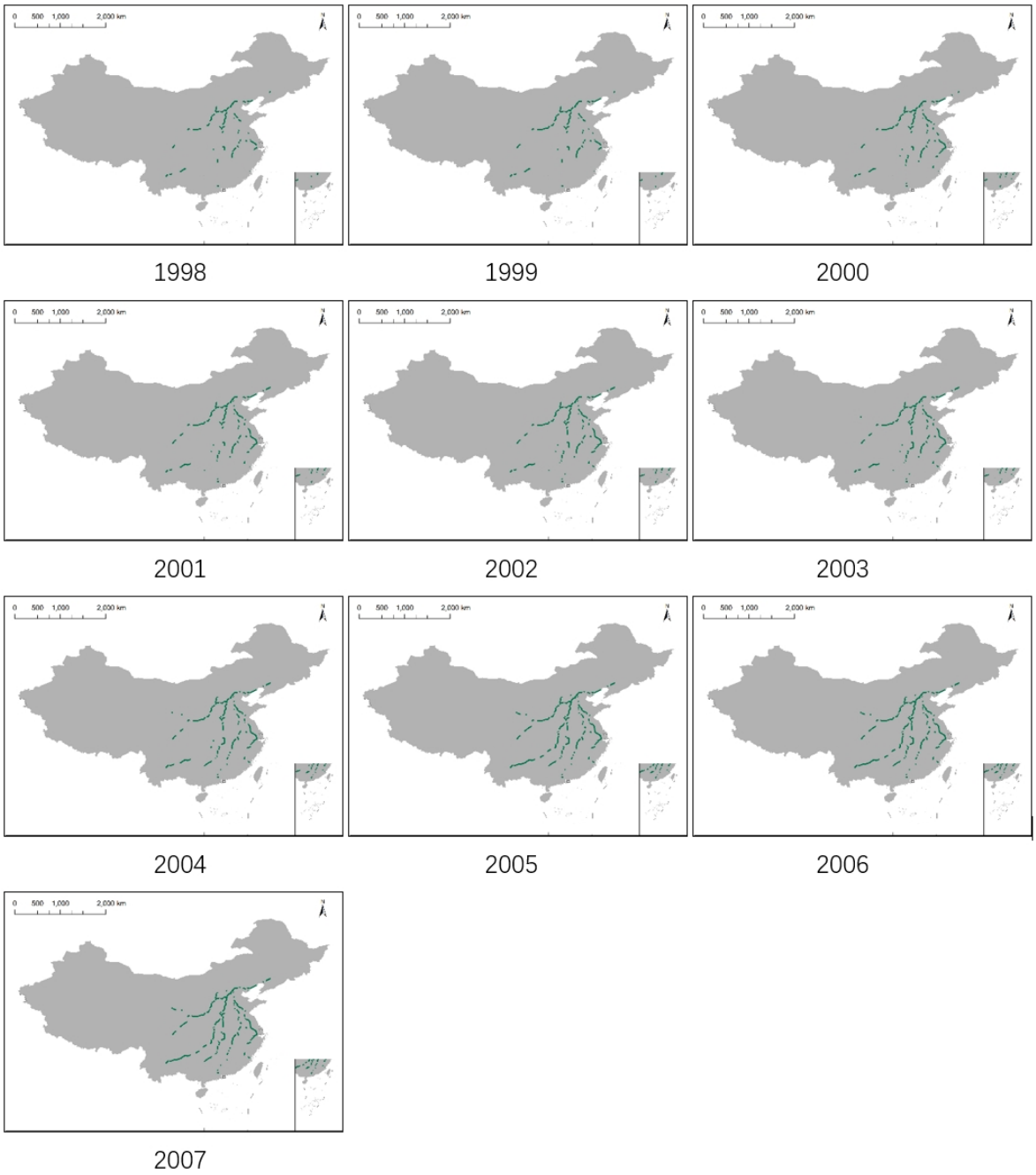


Figure 3.3: Intersections of 10km highway buffers and Qing routes from 1998 to 2007
Note: The network in green depicts the intersections of 10km highway buffers and Qing routes from 1998 to 2007.

This research uses three kinds of historical routes to construct the historical routes instrumental variable, including routes in the Ming Dynasty (1364—1644) and Qing Dynasty (1636-1912), and the combination of these two routes. Ming Dynasty Courier Routes used in this research come from Harvard Dataverse provided by Berman & Zhang (2017) and Qing Courier routes come from Skinner et al. (2008). According to Holl (2016) for constructing the time-variant variable for highway variables, taking Ming routes as an example, this study first

creates a 10km corridor along highways constructed per year from 1998 to 2007. Second, the Ming routes located within the 10km buffer are selected and thus, there are ten intersections of highway buffers and Ming routes from 1998 to 2007. This study also constructs 5km and 20km corridors. Then, the weighted distance from industry to the time-variant historical road is the instrument for highway variables. Figures 3.2 and 3.3 present the intersections of 10km highway buffers and Ming routes, and 10km highway buffers and Qing routes from 1998 to 2007.

Table 3.11 shows the summary statistic of the weighted distance (in metres) from industry to the intersections of the 10km highway buffer with Ming routes, Qing routes and the combination of these two. The trends of the mean distance in these three routes are similar and in general show decreasing distance over time.

Three forms of historical routes IVs for each route are shown:

$$historical\ accessibility = \frac{1}{distance_{industry,historical\ route}} \tag{3.13}$$

For the robustness checks:

$$loghistorical = ln(distance_{industry,historical\ route}) \tag{3.14}$$

$$(loghistorical)^{-1} = \frac{1}{ln(distance_{industry,historical\ route})} \tag{3.15}$$

In summary, the historical routes IV includes Ming routes, Qing routes and the combination of Ming and Qing routes. The intersections with 5, 10 and 20km corridors along highways are used for each route.

Least cost path minimum spanning tree network. This research follows Faber (2014) to construct the Least Cost Path Minimum Spanning Tree network (LCP-MST) network as an instrumental variable for the highway variable. The LCP-MST network is correlated with the highway variable due to the construction of highways depending on the cost of land use and slope, and LCP-MST is purely about that. The target cities are selected according to the highway construction plan in China. The NTHS plan approved by the State Council in 1992

Table 3.11: Summary statistics of the distance from industry to historical routes

year	mean	sd	min	max
Ming routes				
1998	68813	35258	16806	217937
1999	64873	32229	20706	203379
2000	59062	30080	22117	213358
2001	56404	28908	20833	177981
2002	55034	31787	9527	227071
2003	59938	35054	10364	227728
2004	53219	28666	11770	200952
2005	49296	25489	7940	169252
2006	48665	24932	8757	162096
2007	48962	26335	6992	159261
Qing routes				
1998	111220	45750	40821	377317
1999	111614	44292	48519	390190
2000	105065	43512	50382	380677
2001	102031	42537	45732	380734
2002	101790	46335	51137	379093
2003	106771	45725	30745	383566
2004	104692	43036	20445	370147
2005	102530	41912	22664	371908
2006	103709	42652	24063	387031
2007	104003	45084	21038	392893
Combination of Ming and Qing routes				
1998	67319	35201	15861	218180
1999	63040	32066	19911	199912
2000	56938	29930	21743	214076
2001	54434	28693	20578	178342
2002	53160	31597	9302	227274
2003	57706	34639	10043	227630
2004	50917	28125	11516	197987
2005	47092	24925	7625	165559
2006	46265	24339	8446	158106
2007	46537	25771	6641	155289

Note: This table shows the weighted distance from industry to the intersections of the 10km highway buffer with Ming routes, Qing routes and the combination routes. The unit here is the metre.

is the 5-7 plan (5 north-south and 7 east-west routes). The targeted cities are all provincial capitals, cities with an urban population of more than 500,000 and border crossings. The State Council approved a follow-up plan for NEN in 2004 (7-9-18 plan, with 7 radial expressways from the capital, 9 north-south routes and 18 east-west routes). The aim of 7-9-18 is to connect cities with a registered urban population of over 200,000.

The targeted nodes are selected from a report written by the Ministry of Transport of China, which shows the NTHS plan approved in 1992. The NTHS was constructed based on it. The report written by the Ministry of Transport displays 114 targeted nodes, which are used to construct the least cost path spanning tree network instruments. Moreover, in 2004, more cities were in the NEN plan, and thus this study extends the targeted nodes to 323.

The least cost paths between targeted nodes on the basis of remote sensing data on land cover and elevation are computed. The land cover and elevation data of China in 2000 are collected from DIVA-GIS. This research uses a construction cost function by Faber (2014), which assigns higher construction costs to land parcels with steeper slope gradients and land cover classified as water, wetlands, or built-up areas. The cost function is:

$$cost_i = 1 + slope_i + 25water_i + 25wetland_i + 25developed_i \tag{3.16}$$

where $cost_i$ is the cost of crossing a pixel of land i. $Slope_i$ is land i's average slope gradient. $water_i$, $wetland_i$ and $developed_i$ indicate whether the pixel is covered by water, wetland or built-up area. This study uses the remote sensing data on land cover and elevation in 30 arc seconds (approximately 0.82x0.82 km²) to compute the cost for a continuous grid of land parcels.

The minimum spanning tree algorithm is used to identify the subset of routes that connect all targeted nodes based on the global minimum construction cost. Figure 3.4 illustrates the LCP-MST (in blue) constructed with 114 targeted nodes. The network in red depicts the highway network in 2007. Figure 3.5 displays the LCP-MST constructed with 323 targeted nodes, which has more routes than LCP-MST with 114 targeted nodes. The two LCP-MST networks

have fewer routes than the highway network as the minimum spanning tree algorithm connects all targeted nodes on a single continuous network subject to global construction cost minimization, and thus does not allow circles in the network, while the NTHS and NEN networks allow circular routes to connect cities.

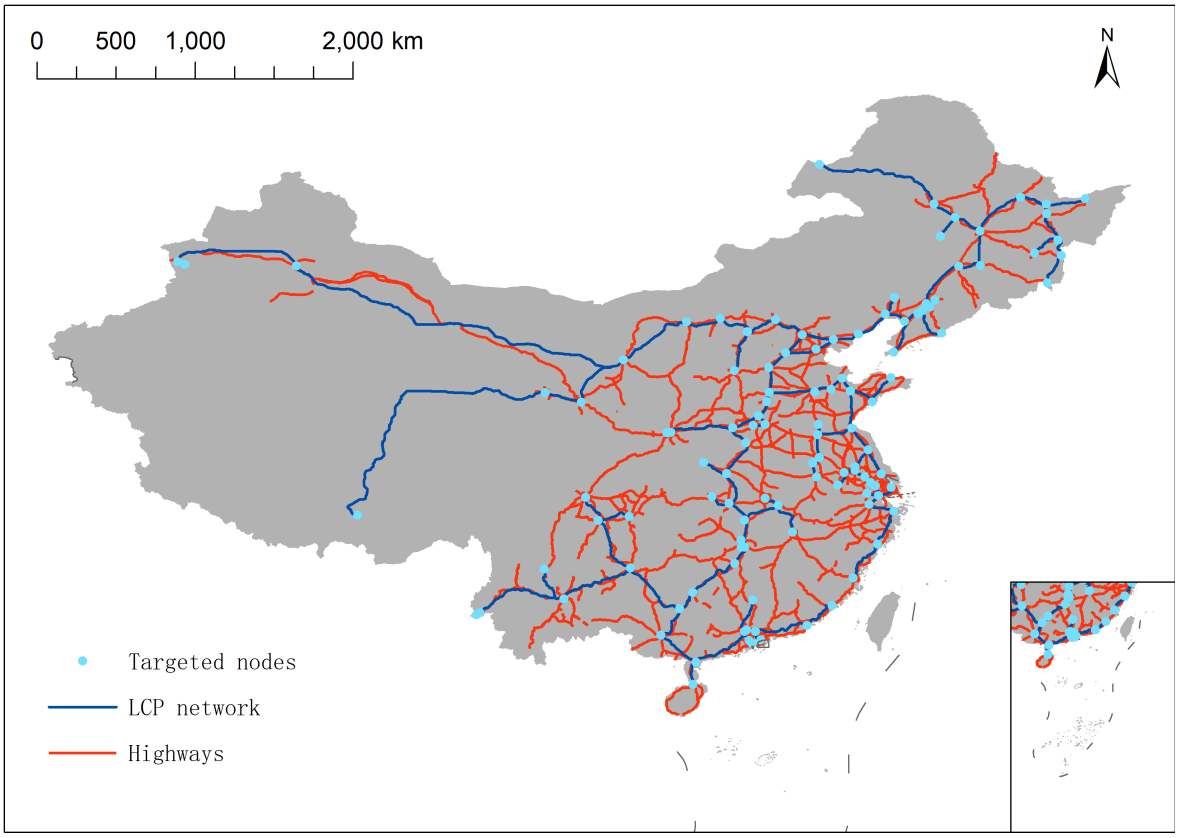


Figure 3.4: Least cost path spanning tree network with 114 targeted nodes

Note: The network in blue depicts the least cost path spanning tree network with 114 target nodes. The network in red depicts highways in 2007. The light blue points are the targeted nodes.

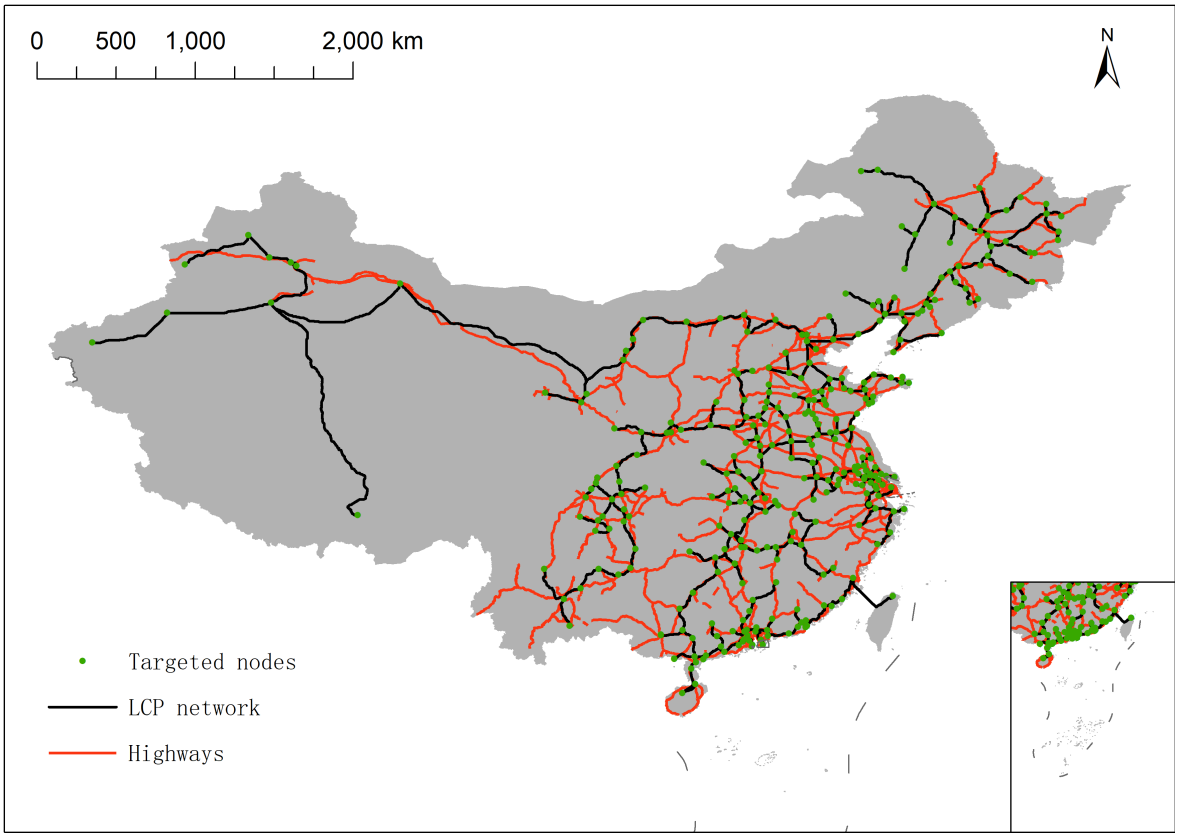


Figure 3.5: Least cost path spanning tree network with 323 targeted nodes

Note: The network in black depicts the least cost path spanning tree network with 323 target nodes. The network in red depicts the highway network in 2007. The green points are the targeted nodes.

Time-variant LCP-MST is constructed at the intersections between 5, 10, 20km highway corridors and LCP-MST. The statistical summary of least cost path spanning tree network instruments by 10km corridor and 114 target nodes is shown in Table 3.12. It describes the weighted distance from industry to the least cost path spanning tree network from 1998 to 2007. The mean distance for 5, 10, and 20km highway corridors does not decrease over the period 1998-2007, which is different from the trends of the historical routes variable and the highway variable. The mean distance to the LCP-MST is much lower than to historical routes, which makes sense as Ming routes do not cover the west and the most northern parts of China.

Table 3.12: Summary statistics of LCP-MST with NTHS nodes

year	mean	sd	min	max
1998	38464	17726	7799	101685
1999	36682	16903	7391	97202
2000	35110	15998	10874	90074
2001	33293	15538	4486	84576
2002	31932	15065	4956	91045
2003	39244	20255	10972	127274
2004	36719	18659	11992	99836
2005	35687	17662	12110	94209
2006	36186	17495	13947	96267
2007	36350	17742	14001	101241
Total	35973	17443	4486	127274

Note: This table shows the weighted distance from industry to the least cost path spanning tree network by 10km corridor and 114 target nodes from 1998 to 2007. The LCP-MST variables constructed by 5km and 20km are not displayed, as they have a similar trend to that of 10km from 1998 to 2007. The unit here is the metre.

Table 3.13: Summary statistics of LCP-MST with mixed target nodes

year	mean	sd	min	max
1998	38464	17726	7799	101685
1999	36682	16903	7391	97202
2000	35110	15998	10874	90074
2001	33293	15538	4486	84576
2002	31932	15065	4956	91045
2003	39244	20255	10972	127274
2004	22440	14644	6959	105822
2005	20956	13662	5660	90234
2006	21235	13550	5830	90338
2007	21477	14264	6054	92942
Total	30075	17438	4486	127274

Note: This table shows the weighted distance from industry to the least cost path spanning tree network by 10km corridor and 114 target nodes from 1998 to 2003, 323 target nodes from 2004 to 2007. The LCP-MST variables constructed by 5km and 20km are not displayed, as they have a similar trend to that of 10km from 1998 to 2007. The unit here is the metre.

In addition to 114 and 323 targeted nodes LCP-MST networks, this research also builds a mixed nodes LCP-MST network. It has 114 targeted nodes from 1998 to 2003 and 323 nodes from 2004 to 2007 in order to extend the targeted cities for the NEN plan after 2004 and upgrade target nodes for highway construction. The statistical summary of least cost path minimum spanning tree network variable with 10km highway buffer and mixed target nodes is shown in Table 3.13. The trend of the mean decreases over time in general, which is expected to fit the highway variable better than that of the 114 target nodes.

Euclidean spanning tree networks IV. The Euclidean spanning tree network is constructed following the method of Faber (2014). It uses the same targeted nodes as the least cost path spanning tree network. Euclidean straight line bilateral connections are used to connect them. Then, the minimum spanning tree network selects the routes subject to minimizing the global network length. Figures 3.6 and 3.7 display the constructed Euclidean spanning tree networks with target nodes in the NTHS plan and NEN plan respectively. The network in black depicts the Euclidean spanning tree networks. The network in red depicts the highway network in 2007. The Euclidean straight line spanning tree network is less precise regarding the actual construction of highway routes and is expected to be weaker than that of LCP-MST IVs.

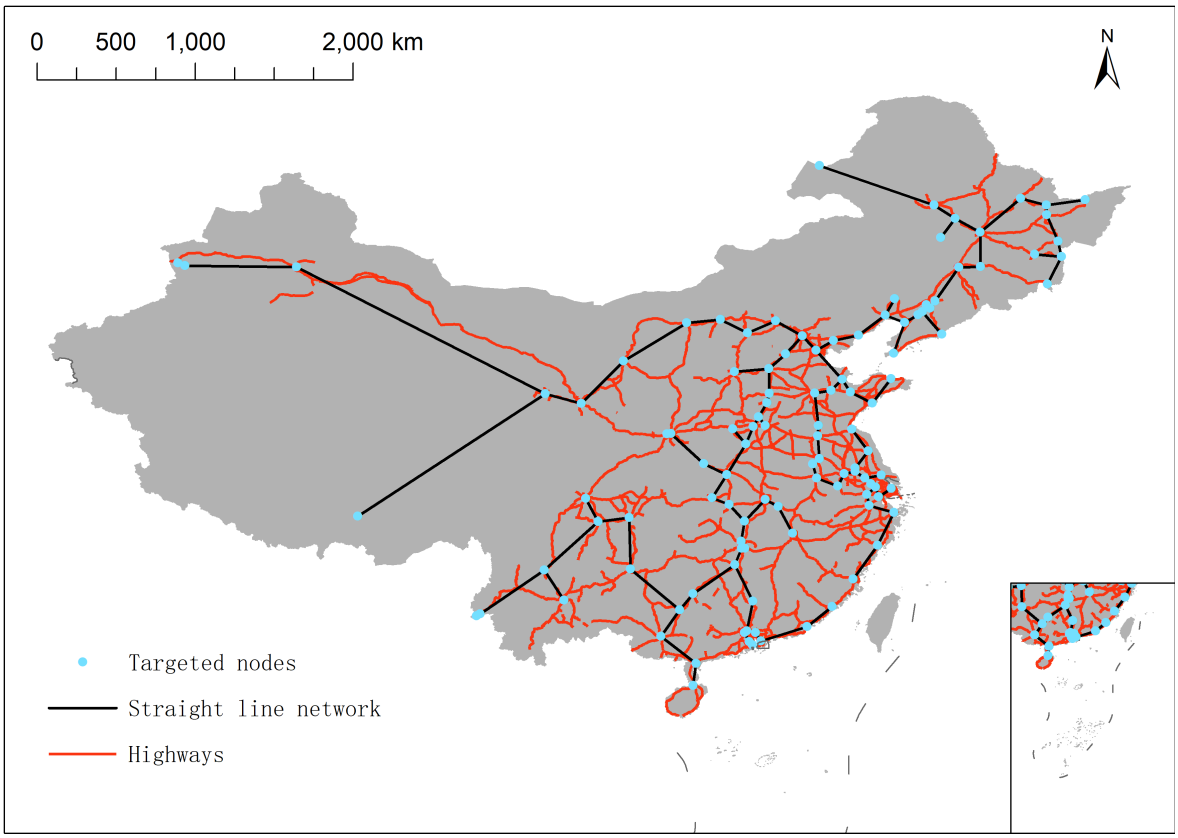


Figure 3.6: Euclidean spanning tree network with target nodes in NTHS plan

Note: The network in black depicts the Euclidean spanning tree network with target nodes in the NTHS plan. The network in red depicts the highway network in 2007. The light blue points are the NTHS targeted nodes, including 114 nodes.

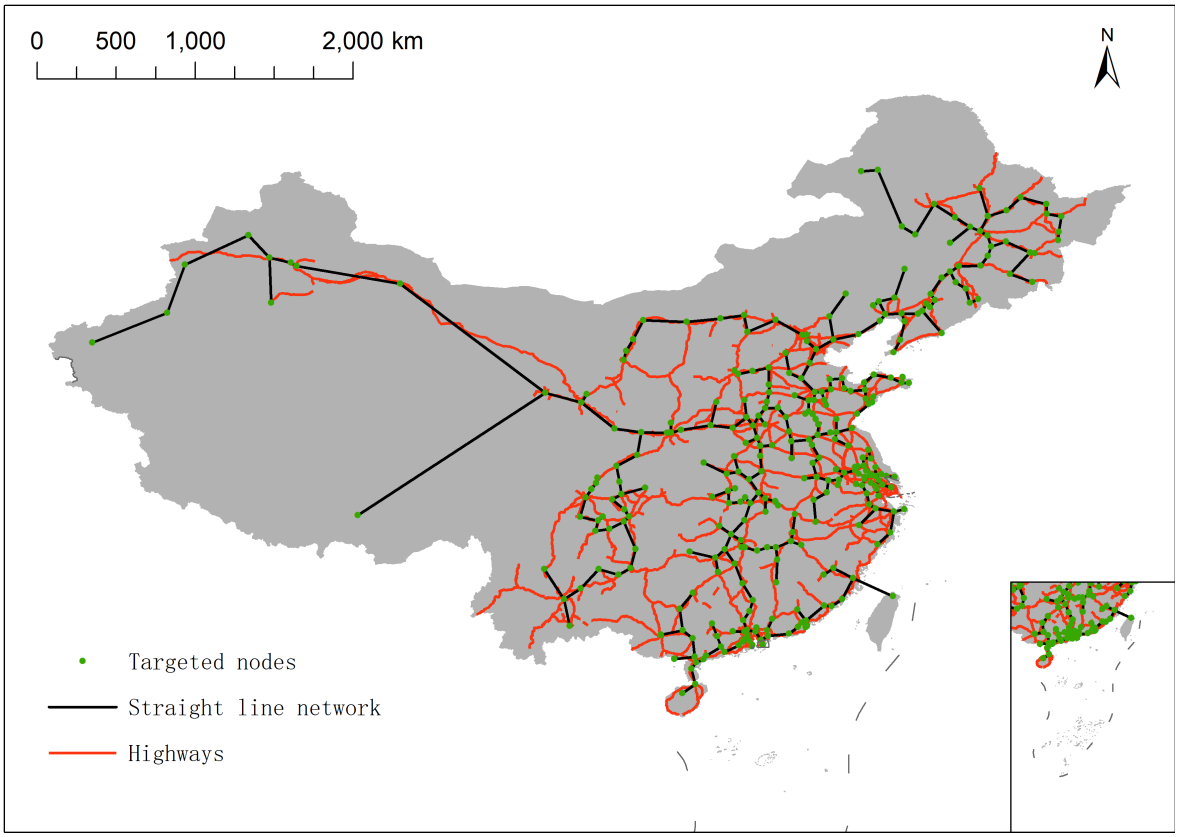


Figure 3.7: Euclidean spanning tree network with target nodes in NEN plan

Note: The network in black depicts the Euclidean spanning tree network with target nodes in the NEN plan. The network in red depicts the highway network in 2007. The light green points are the NEN targeted nodes, including 323 nodes.

Time-variant Euclidean-MST IVs are constructed with 5, 10 and 20km highway corridors and NTHS, NEN and the mixed target nodes in this study. The forms of Euclidean-MST or LCP-MST have three types corresponding to three types of highway variables.

$$MST\ accessibility = \frac{1}{distance_{industry,MST\ network}} \tag{3.17}$$

For the robustness checks:

$$logMST = ln(distance_{industry,MST\ network}) \tag{3.18}$$

$$(logMST)^{-1} = \frac{1}{ln(distance_{industry,MST\ network})} \tag{3.19}$$

Comparison of three types of instruments. The instrument variable needs to satisfy the Relevance Condition and Exclusion Restriction.

$$corr(z,x) \neq 0 \tag{3.20}$$

$$corr(z,e) = 0 \tag{3.21}$$

The instrument must be correlated with x, and can only affect y through the channel of x.

Table 3.14: Pros and cons of three types of IVs

	Historical routes	Least cost path spanning tree	Euclidean spanning tree
Relevance Condition	May be weak	More accurate	Less accurate
Adjustable	NO	YES select targeted nodes, add routes	YES select targeted nodes, add routes
Exclusion Restriction	More exogenous	Less exogenous due to targeted nodes	Less exogenous due to targeted nodes
Time-variant	YES	YES	YES

Historical routes, the LCP-MST network and Euclidean-MST network have their pros and cons, which are displayed in Table 3.14. First, historical routes and the Euclidean routes can be weaker than the LCP-MST. Second, the targeted nodes in the least cost path and Euclidean spanning tree networks can be adjusted to construct better IVs. We can also add other routes to these two networks, for example, routes based on the NTHS plan (5 north-south and 7 east-west routes). However, the target nodes in the least cost path and Euclidean spanning tree are the developed cities with higher populations, which has the possibility of making the instrument correlated to the error term.

3.6.2 IV Estimation Results

The estimation results of FE-2SLS with the historical routes instrument, least cost path spanning tree instrument, and Euclidean spanning tree instrument are described in three sections. These three types of IVs are all time-variant.

Historical Routes IV

Table 3.15 displays the estimation results using the Ming routes instrument (Columns 1-3) and Qing routes instrument (Columns 4-6) and the combination of Ming routes and Qing routes instrument (Columns 7-9). The variable *highway access* is the research focus. As shown in Table 3.15, which displays results with the Ming routes IV, the estimated coefficients of *highway access* obtained from the fixed effect-2SLS estimator are positive and statistically significant at the province level. The Ming route IV is significantly correlated with the highway access variable. The Kleibergen-Paap rk Wald F statistic is used to test the weak identification, and its results indicate that the weak instrument null hypothesis is rejected. The Kleibergen-Paap rk LM statistic and its p-value are also reported, and indicate that it passes the under-identification test. Additionally, the estimated coefficients of the intermediate ratio are statistically significant and positive in FE-2SLS estimation at province levels, indicating a higher level of input sharing leading to a higher level of agglomeration.

The estimation results show that the Qing routes instrument is a weaker IV than the Ming routes instrument. The Kleibergen-Paap rk Wald F statistic indicates that the weak instrument null hypothesis is not rejected. For Qing routes IV, the estimated coefficients of *highway access* obtained from the fixed effect-2SLS estimator are also positive and statistically significant at province levels. The estimated coefficients of the intermediate ratio are positive and insignificant. The coefficients of net labour productivity are negative and insignificant. The proxy for knowledge spillovers has positive but insignificant coefficients. As shown in Columns (7)-(9), the combination of Ming routes and Qing routes IV obtains results consistent with those of Ming routes, including the coefficients for highway access and control variables.

Table 3.15: FE-2SLS results with historical routes IVs

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
Panel A: Second stage of FE-2SLS									
highway access	0.3379***	0.0792	0.0157	0.7957***	0.1985	0.0810	0.3135***	0.0750	0.0126
	(0.080)	(0.054)	(0.030)	(0.251)	(0.128)	(0.078)	(0.083)	(0.052)	(0.029)
intermediate_ratio	0.0494*	0.0100	0.0044	0.0439	0.0086	0.0036	0.0497*	0.0101	0.0044
	(0.030)	(0.011)	(0.006)	(0.040)	(0.013)	(0.007)	(0.029)	(0.011)	(0.006)
net_labour_productivity	0.0000	0.0000	0.0000	-0.0002	-0.0000	-0.0000	0.0000	0.0000	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
new_product_ratio	0.0276	0.0680	0.0215	0.0038	0.0618	0.0181	0.0289	0.0682	0.0217
	(0.107)	(0.047)	(0.029)	(0.120)	(0.048)	(0.029)	(0.107)	(0.047)	(0.029)
agriculture	-0.0049	-0.0074	-0.0105	-0.0189	-0.0110	-0.0125	-0.0041	-0.0072	-0.0104
	(0.030)	(0.008)	(0.008)	(0.042)	(0.010)	(0.008)	(0.029)	(0.008)	(0.008)
fishing	-0.0281	0.0181***	0.0157***	-0.0165	0.0212***	0.0174***	-0.0287	0.0180***	0.0156***
	(0.024)	(0.005)	(0.004)	(0.023)	(0.006)	(0.004)	(0.024)	(0.005)	(0.004)
coal	0.2153***	0.0350	-0.0001	0.1749***	0.0244	-0.0059	0.2174***	0.0353	0.0002
	(0.056)	(0.022)	(0.010)	(0.065)	(0.025)	(0.013)	(0.055)	(0.022)	(0.010)
petroleum	0.2005***	0.0397**	0.0005	0.1505**	0.0267	-0.0067	0.2032***	0.0402**	0.0008
	(0.048)	(0.020)	(0.009)	(0.061)	(0.024)	(0.013)	(0.048)	(0.020)	(0.009)
upstreamness	0.0043	0.0003	0.0008	0.0034	0.0000	0.0007	0.0043	0.0003	0.0008
	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)
Observations	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
Number of industry	161	161	161	161	161	161	161	161	161
Panel B: First stage of FE-2SLS									
Ming road access	2.2843***	2.2843***	2.2843***						
	(0.3776)	(0.3776)	(0.3776)						
Qing road access				2.5403***	2.5403***	2.5403***			
				(0.6445)	(0.6445)	(0.6445)			
combine access							2.1919***	2.1919***	2.1919***
							(0.3293)	(0.3293)	(0.3293)
intermediate_ratio	0.0262	0.0262	0.0262	0.0118	0.0118	0.0118	0.0267	0.0267	0.0267
	(0.0362)	(0.0362)	(0.0362)	(0.0397)	(0.0397)	(0.0397)	(0.0362)	(0.0362)	(0.0362)
net_labour_productivity	0.0004***	0.0004***	0.0004***	0.0005***	0.0005***	0.0005***	0.0004***	0.0004***	0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
new_product_ratio	0.1091	0.1091	0.1091	0.0965	0.0965	0.0965	0.1013	0.1013	0.1013
	(0.0808)	(0.0808)	(0.0808)	(0.0911)	(0.0911)	(0.0911)	(0.0794)	(0.0794)	(0.0794)
agriculture	0.0356	0.0356	0.0356	0.0383	0.0383	0.0383	0.0361	0.0361	0.0361
	(0.0311)	(0.0311)	(0.0311)	(0.0352)	(0.0352)	(0.0352)	(0.0315)	(0.0315)	(0.0315)
fishing	-0.0133	-0.0133	-0.0133	-0.0212*	-0.0212*	-0.0212*	-0.0126	-0.0126	-0.0126
	(0.0108)	(0.0108)	(0.0108)	(0.0116)	(0.0116)	(0.0116)	(0.0106)	(0.0106)	(0.0106)
coal	0.0536	0.0536	0.0536	0.0782**	0.0782**	0.0782**	0.0482	0.0482	0.0482
	(0.0419)	(0.0419)	(0.0419)	(0.0371)	(0.0371)	(0.0371)	(0.0443)	(0.0443)	(0.0443)
petroleum	0.0705*	0.0705*	0.0705*	0.0997***	0.0997***	0.0997***	0.0660	0.0660	0.0660
	(0.0397)	(0.0397)	(0.0397)	(0.0355)	(0.0355)	(0.0355)	(0.0415)	(0.0415)	(0.0415)
upstreamness	0.0014	0.0014	0.0014	0.0029	0.0029	0.0029	0.0011	0.0011	0.0011
	(0.0042)	(0.0042)	(0.0042)	(0.0043)	(0.0043)	(0.0043)	(0.0042)	(0.0042)	(0.0042)
N	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
r2_a	0.508	0.508	0.508	0.428	0.428	0.428	0.508	0.508	0.508
Underidentification test	11.12	11.12	11.12	4.829	4.829	4.829	11.89	11.89	11.89
p-value	0.0009	0.0009	0.0009	0.0280	0.0280	0.0280	0.0006	0.0006	0.0006
Weak identification test	36.61	36.61	36.61	15.54	15.54	15.54	44.33	44.33	44.33

Note: This table shows fixed effect 2SLS results with Ming routes IV, Qing routes IV and the combination of Ming routes and Qing routes IV constructed by the 10km highway corridor. The results for IV constructed by 5 and 20km highway corridors are consistent with that of the 10km highway corridor. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The under-identification test shows the Kleibergen-Paap rk LM statistic and its p-value; Kleibergen-Paap rk Wald F statistic is presented as the weak identification test. The equation is exactly identified - one IV is used for one endogenous variable.

Table 3.16: FE-2SLS results with LCP-MST IVs

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
Panel A: Second stage of FE-2SLS									
highway access	0.1523	0.0573	0.0163	0.1621***	0.0815***	0.0325***	0.1478***	0.0955***	0.0413***
	(0.113)	(0.052)	(0.031)	(0.054)	(0.021)	(0.011)	(0.052)	(0.019)	(0.010)
intermediate_ratio	0.0517*	0.0103	0.0044	0.0515*	0.0100	0.0042	0.0517*	0.0098	0.0041
	(0.028)	(0.011)	(0.006)	(0.028)	(0.011)	(0.006)	(0.028)	(0.012)	(0.006)
net_labour_productivity	0.0001*	0.0000	0.0000	0.0001***	0.0000	0.0000	0.0001***	0.0000	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
new_product_ratio	0.0372	0.0691	0.0215	0.0367	0.0679	0.0207	0.0375	0.0671	0.0202
	(0.110)	(0.047)	(0.029)	(0.110)	(0.047)	(0.029)	(0.108)	(0.047)	(0.029)
agriculture	0.0008	-0.0067	-0.0105	0.0005	-0.0074	-0.0110	0.0010	-0.0079	-0.0113
	(0.027)	(0.008)	(0.008)	(0.027)	(0.008)	(0.008)	(0.026)	(0.008)	(0.007)
fishing	-0.0328	0.0176***	0.0157***	-0.0326	0.0182***	0.0161***	-0.0330	0.0185***	0.0164***
	(0.024)	(0.006)	(0.004)	(0.024)	(0.005)	(0.003)	(0.024)	(0.005)	(0.003)
coal	0.2317***	0.0369*	-0.0001	0.2308***	0.0348	-0.0016	0.2321***	0.0335	-0.0024
	(0.054)	(0.022)	(0.010)	(0.053)	(0.022)	(0.010)	(0.053)	(0.022)	(0.010)
petroleum	0.2208***	0.0421**	0.0004	0.2198***	0.0395**	-0.0014	0.2213***	0.0379**	-0.0023
	(0.047)	(0.020)	(0.009)	(0.045)	(0.019)	(0.009)	(0.045)	(0.019)	(0.009)
upstreamness	0.0046	0.0003	0.0008	0.0046	0.0003	0.0008	0.0046	0.0002	0.0008
	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)
Observations	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
R-squared	0.166	0.162	0.079	0.165	0.165	0.084	0.167	0.161	0.082
Panel B: First stage of FE-2SLS									
LCP 1992 NTHS	1.0362***	1.0362***	1.0362***						
	(0.3133)	(0.3133)	(0.3133)						
LCP 2004 NEN				1.7809***	1.7809***	1.7809***			
				(0.2546)	(0.2546)	(0.2546)			
LCP combination							1.4496***	1.4496***	1.4496***
							(0.1686)	(0.1686)	(0.1686)
intermediate_ratio	0.0172	0.0172	0.0172	0.0251	0.0251	0.0251	0.0186	0.0186	0.0186
	(0.0401)	(0.0401)	(0.0401)	(0.0271)	(0.0271)	(0.0271)	(0.0193)	(0.0193)	(0.0193)
net_labour_productivity	0.0005***	0.0005***	0.0005***	0.0003***	0.0003***	0.0003***	0.0002***	0.0002***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
new_product_ratio	0.0893	0.0893	0.0893	0.1576**	0.1576**	0.1576**	0.0661	0.0661	0.0661
	(0.1092)	(0.1092)	(0.1092)	(0.0661)	(0.0661)	(0.0661)	(0.0681)	(0.0681)	(0.0681)
agriculture	0.0424	0.0424	0.0424	0.0404	0.0404	0.0404	-0.0057	-0.0057	-0.0057
	(0.0375)	(0.0375)	(0.0375)	(0.0261)	(0.0261)	(0.0261)	(0.0166)	(0.0166)	(0.0166)
fishing	-0.0252**	-0.0252**	-0.0252**	-0.0373***	-0.0373***	-0.0373***	-0.0424***	-0.0424***	-0.0424***
	(0.0109)	(0.0109)	(0.0109)	(0.0067)	(0.0067)	(0.0067)	(0.0111)	(0.0111)	(0.0111)
coal	0.0758**	0.0758**	0.0758**	0.0024	0.0024	0.0024	0.0368	0.0368	0.0368
	(0.0357)	(0.0357)	(0.0357)	(0.0546)	(0.0546)	(0.0546)	(0.0223)	(0.0223)	(0.0223)
petroleum	0.1035***	0.1035***	0.1035***	0.0278	0.0278	0.0278	0.0506**	0.0506**	0.0506**
	(0.0337)	(0.0337)	(0.0337)	(0.0474)	(0.0474)	(0.0474)	(0.0211)	(0.0211)	(0.0211)
upstreamness	0.0037	0.0037	0.0037	0.0077**	0.0077**	0.0077**	0.0026	0.0026	0.0026
	(0.0044)	(0.0044)	(0.0044)	(0.0034)	(0.0034)	(0.0034)	(0.0028)	(0.0028)	(0.0028)
N	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
r2_a	0.461	0.461	0.461	0.702	0.702	0.702	0.728	0.728	0.728
Underidentification test	11.50	11.50	11.50	44.72	44.72	44.72	53.97	53.97	53.97
p-value	0.0007	0.0007	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Weak identification test	10.95	10.95	10.95	48.96	48.96	48.96	73.94	73.94	73.94

Note: This table shows fixed effect 2SLS results with LCP-MST IV constructed by a 10km highway corridor and NTHS target nodes, NEN target nodes and the combination of NTHS and NEN plans target nodes. The results for IV constructed by 5 and 20km highway corridors are consistent with those of the 10km highway corridor. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The under-identification test shows Kleibergen-Paap rk LM statistic and its p-value; the Kleibergen-Paap rk Wald F statistic is presented as the Weak identification test. The equation is exactly identified - one IV is used for one endogenous variable.

Table 3.17: FE-2SLS results with Euclidean IVs

	EG (province) (1)	EG (city) (2)	EG (county) (3)	EG (province) (4)	EG (city) (5)	EG (county) (6)	EG (province) (7)	EG (city) (8)	EG (county) (9)
Panel A: Second stage of FE-2SLS									
highway access	0.1113	0.0388	-0.0037	0.1532*	0.0616***	0.0141	0.1828***	0.0892***	0.0317**
	(0.140)	(0.040)	(0.025)	(0.086)	(0.023)	(0.015)	(0.066)	(0.019)	(0.012)
intermediate_ratio	0.0521*	0.0105	0.0046	0.0516*	0.0102	0.0044	0.0513*	0.0099	0.0042
	(0.027)	(0.011)	(0.006)	(0.028)	(0.011)	(0.006)	(0.028)	(0.011)	(0.006)
net_labour_productivity	0.0002*	0.0000**	0.0000**	0.0001**	0.0000*	0.0000**	0.0001**	0.0000	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
new_product_ratio	0.0394	0.0701	0.0225	0.0372	0.0689	0.0216	0.0357	0.0675	0.0207
	(0.111)	(0.048)	(0.030)	(0.111)	(0.048)	(0.030)	(0.109)	(0.047)	(0.029)
agriculture	0.0021	-0.0061	-0.0099	0.0008	-0.0068	-0.0105	-0.0001	-0.0077	-0.0110
	(0.026)	(0.008)	(0.008)	(0.027)	(0.008)	(0.008)	(0.027)	(0.008)	(0.008)
fishing	-0.0339	0.0171***	0.0152***	-0.0328	0.0177***	0.0157***	-0.0321	0.0184***	0.0161***
	(0.024)	(0.006)	(0.004)	(0.024)	(0.006)	(0.004)	(0.024)	(0.005)	(0.003)
coal	0.2353***	0.0385*	0.0016	0.2316***	0.0365*	0.0000	0.2290***	0.0341	-0.0015
	(0.054)	(0.022)	(0.010)	(0.053)	(0.022)	(0.010)	(0.054)	(0.022)	(0.010)
petroleum	0.2253***	0.0441**	0.0026	0.2207***	0.0416**	0.0006	0.2175***	0.0386**	-0.0013
	(0.047)	(0.019)	(0.009)	(0.046)	(0.019)	(0.009)	(0.046)	(0.019)	(0.009)
upstreamness	0.0047	0.0003	0.0009	0.0046	0.0003	0.0008	0.0046	0.0002	0.0008
	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)	(0.004)	(0.001)	(0.001)
Observations	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
R-squared	0.169	0.153	0.055	0.166	0.163	0.077	0.161	0.163	0.084
Panel B: First stage of FE-2SLS									
Euclidean N 1992 NTHS	1.1913***	1.1913***	1.1913***						
	(0.3186)	(0.3186)	(0.3186)						
Euclidean N 2004 NEN				1.6831***	1.6831***	1.6831***			
				(0.2172)	(0.2172)	(0.2172)			
Euclidean N combination							1.4439***	1.4439***	1.4439***
							(0.2133)	(0.2133)	(0.2133)
intermediate_ratio	0.0157	0.0157	0.0157	0.0155	0.0155	0.0155	0.0112	0.0112	0.0112
	(0.0398)	(0.0398)	(0.0398)	(0.0313)	(0.0313)	(0.0313)	(0.0215)	(0.0215)	(0.0215)
net_labour_productivity	0.0005***	0.0005***	0.0005***	0.0004***	0.0004***	0.0004***	0.0002***	0.0002***	0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
new_product_ratio	0.0842	0.0842	0.0842	0.1443**	0.1443**	0.1443**	0.0652	0.0652	0.0652
	(0.1088)	(0.1088)	(0.1088)	(0.0675)	(0.0675)	(0.0675)	(0.0712)	(0.0712)	(0.0712)
agriculture	0.0418	0.0418	0.0418	0.0430	0.0430	0.0430	-0.0075	-0.0075	-0.0075
	(0.0368)	(0.0368)	(0.0368)	(0.0296)	(0.0296)	(0.0296)	(0.0181)	(0.0181)	(0.0181)
fishing	-0.0252**	-0.0252**	-0.0252**	-0.0161*	-0.0161*	-0.0161*	-0.0314***	-0.0314***	-0.0314***
	(0.0110)	(0.0110)	(0.0110)	(0.0084)	(0.0084)	(0.0084)	(0.0110)	(0.0110)	(0.0110)
coal	0.0809**	0.0809**	0.0809**	0.0146	0.0146	0.0146	0.0398	0.0398	0.0398
	(0.0365)	(0.0365)	(0.0365)	(0.0560)	(0.0560)	(0.0560)	(0.0258)	(0.0258)	(0.0258)
petroleum	0.1092***	0.1092***	0.1092***	0.0440	0.0440	0.0440	0.0510**	0.0510**	0.0510**
	(0.0346)	(0.0346)	(0.0346)	(0.0487)	(0.0487)	(0.0487)	(0.0234)	(0.0234)	(0.0234)
upstreamness	0.0036	0.0036	0.0036	0.0087*	0.0087*	0.0087*	0.0027	0.0027	0.0027
	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0044)	(0.0037)	(0.0037)	(0.0037)
N	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603	1,603
r2_a	0.458	0.458	0.458	0.622	0.622	0.622	0.689	0.689	0.689
Underidentification test	13.13	13.13	13.13	37.58	37.58	37.58	52.74	52.74	52.74
p-value	0.0003	0.0003	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Weak identification test	13.99	13.99	13.99	60.07	60.07	60.07	45.84	45.84	45.84

Note: This table shows fixed effect 2SLS results with Euclidean IV constructed by a 10km highway corridor and NTHS target nodes, NEN target nodes and the combination of NTHS and NEN plans target nodes. The results for IV constructed by 5 and 20km highway corridors are consistent with those of the 10km highway corridor. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The under-identification test shows Kleibergen-Paap rk LM statistic and its p-value; the Kleibergen-Paap rk Wald F statistic is presented as the Weak identification test. The equation is exactly identified; IV is used for one endogenous variable.

LCP IV

Table 3.16 displays the FE-2SLS regression results with LCP constructed by the NTHS target nodes in Columns (1)-(3), NEN target nodes in Columns (4)-(6) and the combination of these two (NTHS nodes for 1998-2003 and NEN nodes for 2004-2007) in Columns (7)-(9). More nodes in the LCP IV may cause the IV to not be exogenous enough. The combination IV captures the highway access variable the best (see its adjusted R-square) and is the strongest of the three LCP IVs. With respect to the regression results, the combination LCP IV obtains similar estimated coefficients with those of LCP IV with NEN target nodes. The estimated coefficients for highway access are significant at all geographic levels with LCP constructed by NEN target nodes and the combination IV. However, the LCP IV with NTHS nodes does not pass the weak identification tests.

Most coefficients for natural advantages, such as fishing, coal, and petroleum input shares, are positive and significant, highlighting the crucial role of natural resources in firms' location decisions.

Straight-line IV

Table 3.17 shows the FE-2SLS results with Euclidean IV constructed by the NTHS target nodes, NEN target nodes and their combination (NTHS nodes for 1998-2003 and NEN nodes for 2004-2007). Regarding the estimated coefficients of the highway access variable, the results using the Euclidean IV combination are the most robust, indicating that this combination serves as a stronger instrumental variable than the other two. Compared with the LCP IV, the performance of Euclidean IV is weaker due to the accuracy of Euclidean MST being lower. These results show that highway access and the proxies for input sharing and labour market pooling are likely to have positive impacts on within-industry agglomeration.

Robustness Tests

Table 3.18 presents the results of overidentification tests for time-variant instrumental variables using the FE-2SLS estimation method. The table is divided into two panels. Panel A reports the second-stage regression results, examining the effect of highway access on EG within-industry agglomeration across different administrative levels: province, city, and county. Panel B displays the first-stage regression results, focusing on the determinants of highway access.

In Panel A, the coefficient for highway access is positive and statistically significant across all models, indicating a positive relationship between highway access and within-industry agglomeration calculated at different geographical levels. Specifically, the coefficients are 0.189 for provinces, 0.077 for cities, and 0.028 for counties. Panel B highlights the relevance of the time-variant instruments LCP routes with NTHS nodes and Ming road, which are significant predictors of highway access, as indicated by their statistically significant coefficients. The strong results for the underidentification test (with a test statistic of 39.47 and a p-value of 0) and the weak identification test (with a test statistic of 25.54) confirm that the instruments are appropriate and relevant.

The Hansen overidentification test provides evidence regarding the validity of the instruments. With a Hansen statistic of 2.553 (p-value = 0.110) for the provincial level and above higher p-values for city and county levels, the results do not reject the null hypothesis that the instruments are valid. This suggests that the instruments are uncorrelated with the error term and thus satisfy the exclusion restriction.

Table 3.18: Overidentification tests for time-variant IVs

	(1) EG(province)	(2) EG(city)	(3) EG(county)
Panel A: Second stage of FE-2SLS			
highway_access	0.1892***	0.0766***	0.0276**
	(0.0580)	(0.0231)	(0.0130)
knowledge_spillover	0.0235	0.0653	0.0192
	(0.1134)	(0.0480)	(0.0297)
input_sharing	-0.0011	-0.0010	-0.0011
	(0.0123)	(0.0035)	(0.0021)
labour_pooling	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
upstreamness	0.0053	0.0004	0.0009
	(0.0036)	(0.0012)	(0.0006)
agriculture	-0.0006	-0.0073	-0.0108
	(0.0272)	(0.0083)	(0.0076)
coal	0.2493***	0.0406**	0.0028
	(0.0494)	(0.0205)	(0.0091)
fishing	-0.0282	0.0189***	0.0164***
	(0.0213)	(0.0048)	(0.0031)
petroleum	0.2502***	0.0486***	0.0053
	(0.0453)	(0.0186)	(0.0084)
Panel B: First stage of FE-2SLS Dependent variable: highway access			
LCP NTHS	1.3218***	1.3219***	1.3220***
	(0.2532)	(0.2533)	(0.2534)
Ming road	0.9235**	0.9236**	0.9237**
	(0.4418)	(0.4419)	(0.4420)
Controls	Yes	Yes	Yes
r2_a	0.635	0.635	0.635
Underidentification test	39.47	39.47	39.47
p-value	0	0	0
Weak identification test	25.54	25.54	25.54
Hansen statistics	2.553	0	0.292
p-value	0.110	0.985	0.589

Note: The results of FE-2SLS are presented in this table. The instruments for highway access are the LCP IV and historical routes IV. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

3.7 Heterogeneity Analysis

3.7.1 Supplier Effect and Highway Access

It is hypothesised that improved highway access is likely to reduce the impact of natural resources on within-industry agglomeration. To test this hypothesis, this thesis examines whether enhanced highway access diminishes the effects of natural resource inputs on within-industry agglomeration. The regression results presented in Table 3.19 display the second stage of the FE-2SLS estimation for the interaction terms between highway access and petroleum inputs, and between highway access and coal inputs. The LCP IV is employed; however, for coal, the underidentification test is not passed, with a p-value above 0.1.

Furthermore, for petroleum, the estimated coefficients of the interaction with highway access are negative for agglomeration at the provincial and city levels, but positive for agglomeration at the county level, which are -0.77, -0.36 and 0.18, respectively. This suggests that in industries reliant on petroleum, improved highway access may cause some firms to disperse to other cities or provinces, thereby reducing agglomeration at these levels. Conversely, at the county level, lower transport costs may encourage firms to move closer to petroleum resources, increasing the level of agglomeration. As transport costs decrease, firms might either remain concentrated, potentially drawing in more firms, or relocate to areas offering higher profitability, which may occur due to increased land prices in densely agglomerated regions. This suggests that the relative profitability of agglomeration versus dispersion varies depending on the specific conditions and changes in transport costs.

Table 3.19: Natural resources and highway access

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	EG(province)	EG(city)	EG(county)	EG(province)	EG(city)	EG(county)
highway_access	0.1565***	0.0791***	0.0317***	0.1571***	0.0843***	0.0335***
	(0.0511)	(0.0206)	(0.0112)	(0.0513)	(0.0203)	(0.0112)
highway*petroleum	-0.7743	-0.3602*	0.1836			
	(0.6958)	(0.2106)	(0.1409)			
highway*coal				-0.1951	0.8226**	0.4081**
				(1.0646)	(0.3521)	(0.1804)
knowledge_spillover	0.0242	0.0647	0.0193	0.0250	0.0647	0.0189
	(0.1138)	(0.0480)	(0.0296)	(0.1139)	(0.0481)	(0.0297)
interme_sha_industryi	-0.0008	-0.0006	-0.0012	-0.0013	-0.0015	-0.0014
	(0.0124)	(0.0035)	(0.0020)	(0.0125)	(0.0036)	(0.0021)
labour_pooling	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
upstreamness	0.0056	0.0005	0.0008	0.0054	0.0004	0.0009
	(0.0036)	(0.0012)	(0.0006)	(0.0036)	(0.0012)	(0.0006)
agriculture	0.0003	-0.0074	-0.0109	0.0005	-0.0077	-0.0111
	(0.0267)	(0.0083)	(0.0076)	(0.0266)	(0.0083)	(0.0076)
coal	0.2604***	0.0421*	0.0009	0.2517***	0.0573**	0.0106
	(0.0518)	(0.0223)	(0.0081)	(0.0606)	(0.0240)	(0.0098)
fishing	-0.0297	0.0187***	0.0167***	-0.0290	0.0191***	0.0166***
	(0.0219)	(0.0050)	(0.0030)	(0.0215)	(0.0047)	(0.0030)
petroleum	0.2427***	0.0400**	0.0081	0.2624***	0.0345*	-0.0024
	(0.0492)	(0.0197)	(0.0088)	(0.0393)	(0.0192)	(0.0096)
Observations	1,603	1,603	1,603	1,603	1,603	1,603
Underidentification test	45.66	45.66	45.66	2.330	2.330	2.330
p-value	0.000	0.000	0.000	0.127	0.127	0.127
Weak identification test	23.65	23.65	23.65	24.50	24.50	24.50

Note: The results of FE-2SLS are presented in this table. The instrument for highway access is the LCP-MST IV. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

3.7.2 Customer Effect and Highway Access

Customer location plays a critical role in shaping industry spatial decisions, as firms often cluster or disperse based on transportation costs and proximity to customers. For downstream industries, which sell products directly to customers, this study hypothesises that the effects of highway access on agglomeration are more pronounced due to their ability to relocate away from customers. This study employs the interaction term between highway access and downstreamness to test the hypothesis.

In Table 3.20, a higher value of the upstreamness variable indicates that an industry is further upstream in the supply chain. The interaction term 'highway access*Upstream' shows statistically significant negative coefficients across all levels of geographic aggregation (province, city, and county). Specifically, the coefficients are -0.083 at the provincial level, -0.050 at the city level, and -0.027 at the county level. These findings indicate that, as highway access improves, industries that are more downstream in the supply chain tend to experience an increase in their agglomeration levels.

These results support the hypothesis that improved highway access diminishes the impact of natural resources on within-industry agglomeration. For downstream industries, which rely on proximity to customers for timely delivery and service, the impact of highway access is more significant. This indicates that improved highway access allows downstream firms greater flexibility in location choice, enabling them to optimise for other factors while still maintaining efficient access to the market.

Table 3.20: Downstream and highway access

	(1) EG(province)	(2) EG(city)	(3) EG(county)
Panel A: Second stage of FE-2SLS			
highway access	0.2004***	0.1223***	0.0554***
	(0.0441)	(0.0176)	(0.0113)
highway access*Upstream	-0.0828**	-0.0501***	-0.0272**
	(0.0403)	(0.0173)	(0.0116)
intermediate_ratio	0.0683***	0.0126***	0.0038
	(0.012)	(0.005)	(0.003)
net_labour_productivity	0.0002***	0.0000*	0.0000
	(0.000)	(0.000)	(0.000)
new_product_ratio	0.0619	0.0739**	0.0236
	(0.077)	(0.034)	(0.022)
Observations	1,606	1,606	1,606
R-squared	0.142	0.154	0.074
Number of industry	161	161	161
Underidentification test	151.2	151.2	151.2
p-value	0.0000	0.0000	0.0000
Weak identification test	147.4	147.4	147.4

Note: The results of FE-2SLS are presented in this table. The instrument for highway access is the LCP-MST IV. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

3.7.3 Input-Output adjusted Highway Access

This study hypothesises that improved highway access, measured by the efficiency of transporting inputs and outputs, fosters within-industry agglomeration. Firms utilise highways to move inputs and outputs; enhanced highway access reduces transport costs to other industries, facilitating greater within-industry agglomeration. To test this hypothesis, a variable termed IOHA, which reflects inputs and outputs-adjusted highway access, is constructed. The instrumental variable estimation employs different instruments, with results presented in Tables 3.21 and 3.22.

In Tables 3.21, IOHA denotes the input- and output-based highway access, IHA represents input-based highway access, and OHA refers to output-based highway access. The sizes of inputs and outputs are measured by the input values in the IO table. The results with two instrumental variables LCP114nodes and LCP323nodes are reported in this table. The coefficient for IOHA indicates that a one-unit increase in IOHA is associated with a 0.0002 increase in the EG index at the provincial level, a 0.0001 increase at the city level, and a 0.0001 increase at the county level. This suggests that improvements in input-output highway access contribute to increased within-industry agglomeration.

Input-adjusted highway access and output-adjusted highway access exhibit similar results to IOHA. Specifically, increases in IHA and OHA correspond to comparable rises in the EG index across different levels. This consistency suggests that both IHA and OHA also positively impact within-industry agglomeration. Highways make firms easily get inputs and outputs from other industries and agglomerate with other firms in the same industry.

Table 3.22 presents robustness checks for the impact of input-output adjusted highway access on within-industry agglomeration, using other instrumental variables (IVs): straight line IV, historical routes IV, and LCP time-invariant IV. The results show that the coefficient for IOHA remains positive and statistically significant across all IVs, affirming the robustness of the positive relationship between highway access and within-industry agglomeration.

Table 3.21: Input-output highway access with LCP IV

	(1) EG(province)	(2) EG(province)	(3) EG(province)	(4) EG(city)	(5) EG(city)	(6) EG(city)	(7) EG(county)	(8) EG(county)	(9) EG(county)
IV: LCP114nodes									
IOHA	0.0002			0.0001***			0.0001***		
	(0.0001)			(0.0000)			(0.0000)		
IHA		0.0003			0.0004***			0.0002***	
		(0.0003)			(0.0001)			(0.0001)	
OHA			0.0004			0.0002***			0.0001**
			(0.0003)			(0.0001)			(0.0000)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	No	No	No
N	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567
N_g	157	157	157	157	157	157	157	157	157
Underidentification test	39.31	47.66	30.31	39.31	47.66	30.31	39.31	47.66	30.31
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weak identification test	283.8	418.5	208.9	283.8	418.5	208.9	283.8	418.5	208.9
IV: LCP323nodes									
IOHA	0.0002			0.0001***			0.0001**		
	(0.0001)			(0.0000)			(0.0000)		
IHA		0.0003			0.0003***			0.0001**	
		(0.0003)			(0.0001)			(0.0001)	
OHA			0.0003			0.0002***			0.0001**
			(0.0002)			(0.0001)			(0.0000)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	No	No	No
N	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567
N_g	157	157	157	157	157	157	157	157	157
Underidentification test	38.04	49.08	28.85	38.04	49.08	28.85	38.04	49.08	28.85
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weak identification test	548.2	869.1	386.1	548.2	869.1	386.1	548.2	869.1	386.1

Note: The results of FE-2SLS are presented in this table. The instrument for highway access is the LCP IV. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Both the historical routes IV and straight line IV exhibit strong relevance in predicting highway access, as reflected in high values of the weak identification test. These IVs are likely to satisfy the relevance condition, indicating they are well-suited for predicting highway access. Conversely, while the time-invariant IVs maintain robust exclusion restrictions, they might sacrifice some relevance as they capture less dynamic changes in highway access over time. The time-variant IVs offer increased prediction power for highway access but may compromise the exclusion restriction, whereas time-invariant IVs provide stronger adherence to the exclusion restriction criteria. Thus, the overall evidence supports the hypothesis that improved highway access contributes to within-industry agglomeration, with different IVs validating this finding.

Table 3.22: Input-output highway access with historical and straight line IV

	(1) EG(province)	(2) EG(province)	(3) EG(province)	(4) EG(city)	(5) EG(city)	(6) EG(city)	(7) EG(county)	(8) EG(county)	(9) EG(county)
IV: Straight line 114nodes									
IOHA	0.0002			0.0002***			0.0001***		
	(0.0001)			(0.0000)			(0.0000)		
IHA		0.0004			0.0004***			0.0002***	
		(0.0003)			(0.0001)			(0.0001)	
OHA			0.0004			0.0002***			0.0001**
			(0.0003)			(0.0001)			(0.0000)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	No	No	No	No	No	No	No	No	No
N	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567	1,567
N_g	157	157	157	157	157	157	157	157	157
Underidentification test	39.91	47.48	31.15	39.91	47.48	31.15	39.91	47.48	31.15
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weak identification test	236.8	324.8	184.9	236.8	324.8	184.9	236.8	324.8	184.9
IV: Ming and Qing routes									
IOHA	0.0002			0.0001***			0.0001***		
	(0.0001)			(0.0000)			(0.0000)		
IHA		0.0003			0.0003***			0.0002***	
		(0.0003)			(0.0001)			(0.0001)	
OHA			0.0003			0.0002***			0.0001**
			(0.0002)			(0.0001)			(0.0000)
Underidentification test	37.04	47.17	28.36	37.04	47.17	28.36	37.04	47.17	28.36
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weak identification test	299.5	444.3	213.9	299.5	444.3	213.9	299.5	444.3	213.9
IV: LCP time-invariant									
IOHA	0.0002			0.0002***			0.0001***		
	(0.0001)			(0.0000)			(0.0000)		
IHA		0.0004			0.0004***			0.0002***	
		(0.0003)			(0.0001)			(0.0001)	
OHA			0.0005*			0.0002***			0.0001**
			(0.0003)			(0.0001)			(0.0000)
Underidentification test	39.51	46.70	30.71	39.51	46.70	30.71	39.51	46.70	30.71
p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weak identification test	244.6	363.7	175.9	244.6	363.7	175.9	244.6	363.7	175.9

Note: The results of FE-2SLS are presented in this table. The instruments for highway access include the straight line IV, historical routes IV and LCP time-invariant IV. Standard errors clustered at industry levels are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

3.8 Conclusion

This chapter has studied the impact of highway access on the level of within-industry agglomeration. The level of agglomeration is measured by the EG index at three geographic levels (county, city and province). Highway GIS routes are used to calculate the weighted distance from industry to highway networks for forming the highway access variable. Three types of time-variant instruments are employed to address the endogeneity problem for the highway access variable, including the historical routes IV, LCP IV and straight line IV. The historical routes IV consist of Ming Dynasty routes, Qing Dynasty routes and a combination of them. This study also adjusts the target nodes of LCP and straight-line IVs following the NTHS and NEN plans to obtain stronger IVs. The empirical results support the hypothesis that highway access has positive effects on within-industry agglomeration from both the baseline model and the IV estimations.

The control variables include natural advantages, input sharing, labour market pooling, knowledge spillovers and upstreamness levels. This study finds that natural advantages facilitate industrial agglomeration. Firm-level data are used to compute the proxies for input sharing, labour market pooling and knowledge spillovers. Input sharing and labour market pooling have positive effects on within-industry agglomeration after using LCP and straight-line IVs.

This study performs heterogeneity analyses focusing on the supplier effect, customer effect, and size-adjusted highway access. The results indicate that while improved highway access generally reduces the impact of petroleum on within-industry agglomeration at the provincial and city levels, it increases agglomeration at the county level, reflecting a complex interplay between transportation costs and industry location decisions. Additionally, improved highway access significantly enhances agglomeration for downstream industries by increasing their flexibility to optimise location choices. Moreover, the study finds that enhanced highway access, as measured by the input-output adjusted highway access, significantly enhances within-industry agglomeration. Highways reduce the cost of firms' access to inputs and outputs from other industries, thereby promoting within-industry agglomeration.

This chapter has three limitations. First, the proxies for three Marshallian externalities are constructed with data limitations, which makes their results less significant than other variables. Taking the proxy for labour market pooling as an example, if the number of skilled workers in each industry is available, the estimated results could be better. Second, three Marshallian externalities are likely to be endogenous; though they are not the research focus of this study, it would be better to address this problem. Third, the LCP and Euclidean IVs constructed with target cities are potentially endogenous. To eliminate the endogenous issue caused by targeted nodes, targeted cities can be excluded from the regression.

This study presents several policy implications related to industrial agglomeration. The finding that highway access promotes industrial agglomeration indicates that developing countries aiming to enhance industrial concentration and productivity can greatly benefit from investing in transportation infrastructure, such as highway networks. For these countries, particularly those that are still in the early stages of expanding their infrastructure, the risk of overbuilding is relatively low, and the positive impact of improved highway access on economic activity is substantial. During the period covered by this study (1998 to 2007), China's expansion of highway networks aligned well with growing demand, yielding significant economic benefits.

As highway networks mature, however, the marginal benefits of additional highways may begin to diminish, and the costs associated with maintenance and repair will rise. Thus, while continued investment in transportation infrastructure remains crucial, it is important for developing countries to monitor the balance between infrastructure expansion and its economic returns. Additionally, natural advantages, such as water, agricultural products and mining resources, have a strong power that facilitates industrial agglomeration. This study calls for policies that adapt to local conditions for the development of industries, especially the use of sustainable resources and keeping industries active for the long term.

3.9 Appendices for Chapter 3

Appendix 3.A Review of Agglomeration Measures

Researchers have developed indices to capture the degree of agglomeration, including the Herfindahl-Hirschman index, Entropy index, Gini coefficients, Ellison & Glaeser (1997) index and Duranton & Overman (2005) index. The most tractable and straightforward approach is to use the employment of an industry or the number of firms in an industry. For example, Henderson (2003) uses the number of firms in the same industry to capture the localization economies in high-tech industries. The advantage of using the number of firms to measure agglomeration is that it is possible to assess agglomeration across many industries or when over distance. On the other hand, the simple measure has distinct disadvantages in accuracy.

Herfindahl-Hirschman Index. The Herfindahl-Hirschman Index means the sum of the squares of the market shares of a firm in the industry.

$$H = \sum_{i=1}^{n} S_i^2 = \sum_{i=1}^{n} (\frac{X_i}{X})^2 \tag{3.A-1}$$

Where X_i is an indicator of firm i; X is the sum of the indicator of the whole industry. H is between 0 to 1, and larger H is, the higher the level of agglomeration. The disadvantage is that this cannot show the difference between regions.

Entropy Index. The Entropy Index is a measure of diversity, which has been adopted widely.

$$E_{ij} = \frac{q_{ij}/q_j}{q_i/q} \tag{3.A-2}$$

Where q_{ij} is the output (or employment) of industry i in region j; q_j is the total output of region j; q_i is the output of industry i in the country. q is the total output in the country. A higher E_{ij} means a larger share of industry i in region j compared to the whole country. A larger E_{ij} shows a higher degree of agglomeration and specialization.

Gini coefficient. Kim (1995) develops the Gini spatial concentration index and is among the first to measure agglomeration. Kim uses U.S. manufacturing data over a long period, 1860-1987, describing the long-run trends of spatial concentration in the U.S. Kim (1995) finds that regional specialization over that period shows a curve of a hump in the U.S. and industries are more concentrated when regions are more specialized. A Gini coefficient for each industry i is calculated as follows:

$$G_i = \frac{1}{2n^2\overline{s_l}} \sum_{k=1}^{n} \sum_{j=1}^{n} \mid s_{ij} - s_{ik} \mid \tag{3.A-3}$$

Where s_{ij} is the share of industry i in region j, s_{ik} is the share of industry i in region k, n is the number of regions and $\overline{s_l}$ is the mean of shares. When G_i is zero, it means there is no difference in the share of industry i in any region. Thus industry i is equally distributed. When G_i is close to one, it means the probability of concentration of industry i is high. Ellison & Glaeser (1997) point out the disadvantage of this measure: when G_i is larger than zero, agglomeration may not exist. For example, when a country has only one extremely large firm in an industry and G_i is however big, which does not mean the degree of agglomeration is high.

EG index. Ellison & Glaeser (1997) develop a model that analyses spatial concentration and captures both random concentration of a dart-throwing model and additional concentration resulting from localized industry-specific spillovers and natural advantages. They use 459 four-digit manufacturing industries data to describe the patterns of spatial concentration in the U.S. Their empirical results indicate that many industries are highly concentrated and agglomeration is ubiquitous, while the degree of agglomeration of many industries is low. The index of agglomeration proposed by Ellison & Glaeser (1997) has been adopted by many studies. The degree of industrial agglomeration is measured as:

$$\gamma_i \equiv \frac{G_i - (1 - \sum_r x_r^2) H_i}{(1 - \sum_r x_r^2)(1 - H_i)} \tag{3.A-4}$$

where γ_i is the agglomeration level of industry i. G_i is the Gini index, $G_i \equiv \sum_r (x_r - s_r^i)^2$, with x_r the share of total employment of all industries in region r. s_r^i is the share of output or employment of region r in industry i. $H_i = \sum_j z_j^2$ represents the Herfindahl index of industry i which shows the size of firms in the industry, with z_j the output or employment share of a

firm j in industry i. A larger value for the Herfindahl index means a less competitive industry. The advantage of the EG index is that it considers the difference between the share of industry and the size of the region, and industry structure. The EG index generates the agglomeration degree of each industry.

The advantage of the EG index is that it captures the characters of the Gini index and the Herfindahl index. Compared with the Gini index, the EG index has the advantage that it takes the share of firms (industry structure) into consideration. The disadvantage is that it needs data for firms. Additionally, this index symmetrically treats all spatial units, and disregards the distance between spatial regions.

DO index. The continuous agglomeration index by Duranton & Overman (2005) has also been used in the previous literature with the advantage that it contains information about distance. The DO index needs the geographic coordinates of firms to compute the bilateral distance. $K-density$ is the density function of bilateral distance. For industry i with n firms, $\frac{n(n-1)}{2}$ bilateral distance is generated. The estimator of the density at any distance scope (d) is

$$\hat{K(d)} = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j)} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j) f(\frac{d-d_{i,j}}{h}) \tag{3.A-5}$$

where $d_{i,j}$ is the Euclidean distance between firm i and j in an industry. d is the distance that f is the kernel density function and h is the bandwidth. Duranton & Overman (2005) use a Gaussian kernel to smooth the noise in the measurement of distance. $e(i)$ is the number of employees of firm i. e is used as the weight of firms in this index. The advantage of the DO index is that it can detect departures from randomness and consider continuous distance. However, a substantial amount of data is required in this method, which is its major disadvantage and not many papers use it.

Appendix 3.B EG Model Explanation

The Ellison and Glaeser agglomeration index is designed to measure within-industry agglomeration by considering both natural advantages and spillover effects among firms. In the model, each business unit chooses its location to maximize profits, which depends on the area's inherent profitability and potential spillovers from other firms.

Without spillovers, the model resembles a standard logit model where a firm's probability of choosing a location depends on the area's share of overall manufacturing employment. Natural advantages are captured by a parameter γ^{na}, which measures how much these advantages contribute to industry concentration.

When spillovers are included, the model introduces a parameter γ^{s}, representing the likelihood that firms benefit from being close to each other. The combination of natural advantages and spillovers forms the basis for the agglomeration measure.

The EG index is then constructed using a Gini-like measure of geographic concentration, G, which compares the actual distribution of industry employment across regions to the overall distribution of employment. The final EG index, γ, is calculated by adjusting G for the Herfindahl index H (which measures firm size concentration) and the natural employment distribution across regions. The formula for γ effectively isolates the agglomeration effects by accounting for both geographic concentration and industry structure.

In essence, the EG index provides a robust measure of agglomeration by integrating the effects of natural advantages, spillovers, and firm size distribution, offering a comprehensive view of within-industry spatial concentration. The detailed formation of the model is explained as follows.

In the Ellison and Glaeser model (Ellison & Glaeser 1997), it assumes that the kth business unit chooses its location v_k to maximize its profits given that it will receive profits π_{kr} from locating in area r.

$$log\pi_{kr} = log\overline{\pi}_r + g_r(v_1, ..., v_{k-1}) + \varepsilon_{kr} \tag{3.B-1}$$

Where $\overline{\pi}_r$ is the profitability of locating in area r for a firm. g_r is the effect of spillovers created by firms that have previously chosen locations. ε_{kr} is an additional random component for firm k.

Assume that there are no spillovers $g_r \equiv 0 \; for \; all \; r$, then this model turns to a standard logit model and the firms' location choices are

$$prob\{v_k = r|\overline{\pi}_1, ..., \overline{\pi}_M\} = \frac{\overline{\pi}_r}{\sum_j \overline{\pi}_j} \tag{3.B-2}$$

They first assume that

$$E_{\overline{\pi}_1,...,\overline{\pi}_M} \frac{\overline{\pi}_r}{\sum_j \overline{\pi}_j} = x_r \tag{3.B-3}$$

where x_r is area r's share of overall manufacturing employment. Then assume that the joint distribution of natural advantages is such that there is a single parameter $\gamma^{na} \in [0,1]$ for which

$$var(\frac{\overline{\pi}_r}{\sum_j \overline{\pi}_j}) = \gamma^{na} x_r (1 - x_r) \tag{3.B-4}$$

where γ^{na} captures the importance of natural advantage to the industry.

They then add spillovers whose importance is indexed by a parameter $\gamma_s \in [0,1]$, and assume that

$$log\pi_{ki} = log\overline{\pi}_r + \sum_{l \neq k} e_{kl}(1 - u_{lr})(-\infty) + \varepsilon_{kr} \tag{3.B-5}$$

Where e_{kl} are Bernoulli random variables equal to one with probability γ^s that indicate whether a potentially valuable spillover exists between each pair of plants. u_{lr} is an indicator for whether plant l is located in area r. $-\infty$ means that the spillovers are strong enough so that firms k and l will have negative infinity profits if they are located apart.

They construct a measure of an industry's geographic concentration by setting

$$G \equiv \sum_r (s_r - x_r)^2 \tag{3.B-6}$$

Where x_r is the share of aggregate employment in area r. s_r the share of the industry's employment in area r.

$$s_r = \sum_k z_k u_{kr} \tag{3.B-7}$$

where z_k: the kth firm's share of the industry. u_{kr}: an indicator variable equal to one if firm k chooses to locate in area r.

With the above model setting, thus they make a Proposition

$$E(G) = (1 - \sum x_r^2)[\gamma + (1-\gamma)H] \tag{3.B-8}$$

Where G is a simpler measure of an industry's agglomeration. $H \equiv \sum_k z_k^2$ is the Herfindahl index of the industry's firm size distribution. $\gamma = \gamma^{na} + \gamma^s - \gamma^{na}\gamma^s$.

Using the above proposition, they use γ as the indicator for within-industry agglomeration measurement, where

$$\gamma \equiv \frac{G - (1 - \sum_r x_r^2)H}{(1 - \sum_i x_r^2)(1 - H)} \tag{3.B-9}$$

Chapter 4

The Effects of Highway Access on Coagglomeration

4.1 Introduction

The spatial co-location of different industries began in China after the economic reform in 1978 when firms' location choices started to be driven by the market. Geographical economic activities boomed as plants tended to locate where inputs, ideas and workers were easily accessed. Manufacturing industries that have input-output linkages, workers with similar skills and knowledge spillovers are likely to locate together to obtain lower production costs and keep up with the latest technology. The coagglomeration of vertically-linked industries or industries in different value chain positions but related to the same final products is common in China (Dai et al. 2021).

Geographical economic activities have increased over time, and meanwhile, highway construction in China has experienced fast development. Investment in transport infrastructures is increasingly important, and governments hold the view that it provides profound and long-term benefits to the development of the country. The rapidly built highway networks significantly reduce transport time and costs. Regions connected with highways are expected to attract more firms. The locations of different industries might be affected by the rapid expansion of highways in China.

Can the location of different industries be affected by the rapid expansion of highways due to time- and cost-saving effects? Do regions with better highway connections attract more firms from different industries? What are the key factors that drive industrial coagglomeration in China? Does the transport infrastructure such as the highway network contribute to coagglomeration? If so, what are the heterogeneous effects of highways on spatial co-locations of industries with different features? These interesting questions motivate this chapter.

This chapter examines the impact of highway networks on pairwise coagglomeration of manufacturing industries and further explores the heterogeneity. Highways and industrial coagglomeration are closely related for at least three reasons. First, highway networks are crucial for transporting manufacturing goods and affect the geographic distribution of manufacturing industries. Second, highways may attract firms to co-locate near them. Third, the reduction in transport costs as a result of better highway access may provide firms with better opportunities to locate in a place where they can benefit from knowledge spillovers and labour pooling.

This chapter contributes to four main aspects. First, it examines the impact of the highway network on the level of coagglomeration for industry pairs, which is important but not thoroughly investigated in the literature. Industrial coagglomeration is important for the development of economies as firms benefit from external economies brought about by coagglomeration. The benefits that have been researched include productivity (Tokunaga & Kageyama 2008, Barrios et al. 2006), innovation (Connell et al. 2014) and carbon emission reduction (Li et al. 2019). Some determinants of coagglomeration include national advantages (Ellison et al. 2010) and Marshallian externalities (Ellison et al. 2010, Faggio et al. 2017, Diodato et al. 2018, Howard et al. 2016, Mukim 2015) have been extensively researched. Gallagher (2013) finds that shipping costs and information costs affect industrial coagglomeration at the metropolitan level. The relationship between rapid highway construction in China and industrial coagglomeration has not been specifically investigated and this study is the first to do so.

Second, this chapter makes a significant contribution to the literature by introducing the bilateral Input-Output adjusted Highway Access metric, which offers a more nuanced measure of highway access between industry pairs. Unlike previous measures that often overlook variations in industry size and transportation needs, the bilateral IOHA measurement directly captures the input and output transported between industry pairs via highways. This innovative approach not only addresses a critical gap in the existing research but also enhances our understanding of how infrastructure improvements can influence industrial coagglomeration.

Third, compared with those studies using aggregate measures of transport infrastructure such as highway or railway density, our research is based on comprehensive microeconomic data which offer a much deeper analysis of the research question. For instance, the GIS highway data which depict exact highway locations are used to compute the distance from each firm (or the weighted average from each industry) to the nearest highways. Microeconomic firm-level data are used to construct the coagglomeration index and offer the possibility to explore various interesting heterogeneous channels through which highways impact industrial coagglomeration.

Moreover, in order to shed light on causality and deal with the potential endogeneity problem, this study constructs and applies a number of time-varying instruments of the highway measure, including the historical instruments based on the Ming and Qing Dynasties' courier routes, the least cost paths and the straight line network based on the targeted city points outlined in the highway construction planning.

Using a combination of GIS data for highways and a large dataset of Chinese manufacturing firms over the period of 1998-2007, this chapter finds that highway access increases pairwise coagglomeration at the province and city levels, while at the county level, the effects are statistically insignificant. The positive effect of highways on coagglomeration is larger for industry pairs with a lower share of state-owned enterprises, for related industries, and for industry pairs with input-output linkages.

The structure of the chapter is as follows. Section 2 reviews both theoretical literature on coagglomeration and empirical research that investigates the determinants and effects of co-agglomeration. Section 3 provides a description of the data, the key variable and the model specification. The empirical results of the baseline model are presented in Section 4. Section 5 investigates instrumental variable estimations to deal with the endogenous problem. The heterogeneous results across industry pairs are examined in Section 6. Finally, the conclusion and limitations of this study are discussed in Section 7.

4.2 Literature Review

4.2.1 Theory of Coagglomeration

The spatial concentration of an industry group is referred to as coagglomeration. The concentration of a pair of industries is termed to as pairwise coagglomeration. Porter's (1990) definition of clusters is that particular industries are related to one another and tend to cluster together. Porter makes a compelling case for the importance of clusters in firm location decisions and industrial strategy. The term 'coagglomeration' was coined by Ellison & Glaeser (1997) to describe the more general trend of various industries gathering together.

The economic theory of coagglomeration has not been much developed. Porter (1990) proposes the industrial cluster but does not provide a formal analysis of coagglomeration. Urban research mainly focuses on either complete specialization (only one industry in a city) or complete diversification (colocation of all industries in a city), until Helsley & Strange (2014) build a model for coagglomeration of some but not all industries. The model by Helsley & Strange (2014) is the typical model for coagglomeration and some later coagglomeration models have variations. There are enough theories on agglomeration but so far there are not many theories focusing on coagglomeration.

Transport Cost and Coagglomeration

Krugman (1991) develops a core-periphery model that illustrates the formation of agglomeration. The model has two regions and two sectors: manufacturing and agriculture. The agricultural sector is constant return of scale and the manufacturing sector is increasing return to scale. The farmers cannot move between two regions, whereas workers are movable. Krugman assumes the home market exists, that is, the place where workers are located is also the market, which is a crucial force of agglomeration. To simplify the model, pecuniary externalities are proposed in the model, which results from the supply or demand linkage in a region with a high concentration of other producers. The transport cost is an important factor in the model, which is measured as only a fraction of goods arriving in the other region (the 'iceberg' assumption). The main finding in the core-periphery model suggests that a lower transport cost, higher economies of scale (a lower elasticity of substitution) and a higher share of manufactured products expenditure contribute to spatial concentration.

Regarding the role of transport costs in the core-periphery model, when transport cost is high, the relative value of sales of two regions is larger than one: workers are distributed in two regions and there is no concentrated manufacturing. Then when transport cost falls to a certain threshold, the relative value of the sales of two regions falls below one and the concentration in one region appears (an equilibrium that all workers locate in one region). When transport cost keeps falling, the relative value of sales continues falling and after a certain point, it starts to rise. It approaches one when transport cost is zero; at that time, the location does not matter.

Belleflamme et al. (2000) build an equilibrium model for the emergence of clusters. Their model first uses oligopoly with two firms and examines their location decisions in either of two regions. This explores the location choices for big companies. The co-location in one of the regions enables two firms to save production costs. Localization economies mean the effects of the agglomeration of similar firms. The benefits of localization economies are represented by the marginal cost reduction in the model by Belleflamme et al. (2000). Then they use the model of a large number of firms to simulate the emergence of clusters. This enables them to research the emergence of clusters of many small firms.

They find that three factors determine the emergence of clusters, including low transport costs, differentiated products, and substantial localization economies. A high transport cost disperses firms, and a sufficiently low transport cost makes firms locate together. A difference between the model by Belleflamme et al. (2000) and that of Krugman (1991) is that, according to Belleflamme et al. (2000) the dispersed distribution smoothly turns to agglomeration when the transport cost falls. The second force is price competition: fierce competition prevents firms from co-locating too intensively, whereas product differentiation mitigates competition. For example, if different firms produce very differentiated products, then they are not influenced by price competition. Another factor is the scale of localization economies, and stronger localization economies are beneficial for agglomeration.

Porter's Industrial Clusters

Porter (1990) creates a link between clusters and regional competitiveness. Clusters are spatial concentrations of interconnected organizations and enterprises, according to Porter. Clusters encompass suppliers, customers, and producers of complementary items who own associated skills, technologies, or inputs, and may also incorporate governmental organizations, educational institutions, standards-setting agents, trade groups, and other complementary institutions. Clusters encourage both collabouration and competitiveness. Porter (1990, 2000) investigates industry clusters from a competitiveness standpoint. The competitiveness of the enterprise and others that make up the cluster determines a region's competitive advantage. In the global economy, the choice to build clusters promotes regional or national competitiveness. clusters provide three favourable impacts amid competition: (1) enhancing firm productivity; (2) encouraging future productivity through innovation; (3) supporting the establishment of new enterprises, which in turn expands the cluster (Porter et al. 1998).

Porter (1990) introduces the diamond model of competitive advantage to answer two questions: (1) why companies are competitive in a given industry within a country, and (2) why a country is the most competitive globally for a particular industry. Porter (1990) focuses on four key factors in determining the competitive advantage of industries, including firm strategy, structure and rivalry, demand conditions, factor conditions, and related and sup-

porting industries, which are all based on the location advantages that a specific industry in a specific country retains. Government and chance are two additional factors that are usually used for the model. These factors influence the national environment in which businesses are founded and grow.

Factor conditions. The resources of a country, encompassing natural, human and capital resources, are referred to as factor conditions. Two categories of factor conditions are distinguished: basic and advanced. The former incorporate land, climate and natural resources, whereas the latter involve skilled labour, technology, transportation infrastructure, tax policy, etc. Advanced factor conditions can be upgraded on a regular basis and play a more important role in competitive advantages than basic ones. Some advanced factor conditions can be promoted by cluster rewards, such as a skilled labour pool, technological spillover, and a tax policy in the clusters.

Demand conditions. Demand conditions refer to the home-market demand for an industry's products or services. The industry that serves more home-market demand, such as a huge internal market and the demand for high quality, is under greater pressure to improve. The sophisticated demand condition drives the enterprise to address future customer needs and satisfy the global market, and as a result, the industry innovates more quickly to gain sustainable competitive advantages.

Related and supporting industry. A related and supporting industry is critical to the growth of one industry. An industry benefits from a supplier industry that is innovative, provides high-quality intermediate products and distributes timely information. Related and supporting industries have joined forces to locate near one another in order to cut costs on transportation and communicate more effectively. As a result, an industry cluster emerges in a region, which helps industries by allowing them to grow together and sustain competitive advantages.

Firm strategy, structure and rivalry. The development of a firm is influenced by its strategy, such as emphasizing high quality and high margins. Different organizational and management structures, whether high or low hierarchy, have an impact on how a firm is developed and operated. The presence of rivalries forces a firms to create differentiated products and diversifies the products available in the market. A higher level of domestic competition means there is more incentive to innovate and a better possibility of entering the global market.

Coagglomeration model of Helsley & Strange (2014)

Helsley & Strange (2014) develop a model of coagglomeration that is the first to explain the foundation for coagglomeration. Before their research, theoretical literature always distinguished between cities where there is complete specialization (only one industry present in a city) or complete diversity (all industries in a city). The contribution of Helsley & Strange (2014) is that they consider the intermediate scenario of cities that have coagglomeration of some industries rather than all industries. They question the conventional idea that a city's coagglomeration necessarily entails mutual benefit. They find that equilibrium cities will be inefficient in city composition and in city scale.

Their coagglomeration model analyses the emergence of cities, followed by an examination of the characteristics of an efficient city. In the model, agglomeration external economies such as input sharing, knowledge spillovers and labour market pooling contribute to the growth of output per worker of a firm. In the model by Helsley & Strange (2014), the agglomeration externalities come from both intra-industry and inter-industry activities, which allows for both localization effects from the specialization of an industry and urbanization effects from the diversity of industries. Helsley & Strange (2014) assume that intra-industry external economies are larger than inter-industry economies with evidence from Henderson (2003) and other literature. Within-industry agglomeration generates more productivity benefits than cross-industry effects.

The model assumes that workers are mobile and choose cities to maximize their utility. The cost of living in a city depends on land cost and transportation costs. Helsley & Strange (2014) describes that different workers prefer different compositions (clusters) of local employment. Workers prefer the city where their own industry is dominant. Workers have the incentive to

go where the same type of workers are the majority. The inefficiency comes about because of the weakness of migration of workers, no worker is willing to first move to other cities with other types of workers. As a result, even if the coagglomeration of related industries is superior and improves welfare, it may fail to emerge.

Helsley & Strange (2014) also show that industries may coagglomerate even if there is no benefit arising from their co-location. The workers would be in the shared city with other types of workers though they do not benefit each other. Helsley & Strange (2014) provide an example of the coagglomeration of the oil-refining industry and software industry in San Francisco. These two industries are not complementary but they are located together, and the sufficiently strong own-industry agglomeration effects can account for this situation.

In summary, Helsley & Strange (2014) build a model of city composition where transport cost is considered and find that coagglomeration may not generate optimal benefits for the industries that are located together as long as the current city offers sufficient utility for workers, such as the coagglomeration of irrelevant industries. Complementary industries that have mutual benefits may not select to coagglomerate due to individual migration that no worker is willing to be the first to move away from their current specialized city.

Agent-based Coagglomeration Model of O'Sullivan & Strange (2018)

O'Sullivan & Strange (2018) simulate the emergence of coagglomeration using an agent-based model. They start with a city composition model that has multiple equilibria, including all specialized cities, all diverse cities and the coexistence of specialized and diverse cities. Their basic model is like that of Helsley & Strange (2014), to which they add the agent-based model. Firms which are exposed to intra-industry and inter-industry externalities relocate to cities where they maximize profits. They find that inter-industry externalities and coagglomeration have a positive and nonlinear relationship. History, firm size and relocation costs also affect the magnitude of coagglomeration. In their model, the within-industry and cross-industry external economies, which is continuously differentiable and concave. In equilibrium, it is impossible for firms to raise profits by moving to another city. They show that Nash equilibria can be achieved if all cities are specialized, all cities are diverse and the combination of specialized and diverse cities.

O'Sullivan & Strange (2018) find that there is a positive and nonlinear relationship between levels of coagglomeration and cross-industry externalities. They explain that firms are willing to co-locate to obtain a higher profit if there is a higher level of coagglomeration externalities. Larger coagglomeration economies also discourage firms from relocating to the city where their own industries dominate. They also show that initial conditions for cities affect the final locations of firms. They test the initial conditions that all cities are diverse, all cities are specialized, and different elasticities for coagglomeration and agglomeration effects. They find that when the initial condition is all diverse, the requirement for coagglomeration elasticity is lower for generating diverse cities in equilibrium, while for initially all specialized cities, the requirement is much higher.

They use a higher number of workers in firms to test the model for larger firm sizes, and find that a larger firm size reduces the power of within-industry effects and generates a higher level of coagglomeration. Additionally, other key findings are that the equilibrium of coagglomeration is less than efficient level due to cross-industry external economies being assumed to be smaller than within-industry economies. Moreover, a higher relocation cost facilitates coagglomeration as the firm is less willing to relocate to a specialized city with its own industry and stays in the initial city.

Coagglomeration of Producer Services and Manufacturers

Lanaspa et al. (2016) develop a theoretical framework that focuses primarily on coagglomeration of intermediate producer services and manufacturers. Their model is built on a standard theoretical New Economic Geography model called the Footloose Entrepreneur Model. They incorporate intermediate producer services into the New Economic Geography model, which emphasizes the importance of intermediate producer services in the manufacturing sector. They discover that intermediate producer services induce industrial spatial concentration. When intermediate producer services are productive and less differentiated, and more skilled workers in manufacturing are required, the spatial concentration of manufacturing is promoted.

4.2.2 Empirical Research on Coagglomeration Patterns

Coagglomeration pattern in UK. Duranton & Overman (2008) use extensive firm-level data from the Annual Respondent Database in the UK (which underlies the Annual Census of Production) to look at location patterns of manufacturing industries. To measure agglomeration and coagglomeration, they employ the point-pattern methodology developed by Duranton & Overman (2005). They find good evidence of coagglomeration of vertically-linked industries at the regional scale (around 150 km). On the other hand, agglomeration surpasses coagglomeration at small geographical scales, since firms tend to locate nearer to their own industry than to industries with which they have significant input–output ties at small spatial ranges.

Coagglomeration patterns of new and incumbent firms in Germany. Falck et al. (2014) examine coagglomeration of the manufacturing sector in East and West Germany respectively, since they have distinct institutional and economic conditions. The East German economy underwent a dramatic change from a socialist to a market economy. They use data from the German Social Insurance Statistics on new companies from 1998 to 2001, which comprise incumbents in 103 three-digit manufacturing categories. To characterize patterns of spatial processes, they employ the method of Duranton & Overman (2005). Their findings reveal a large disparity between the two parts of Germany. Coagglomeration for new enterprises in West German manufacturing industries is around 40%, while it is only 5% in East Germany.

Coagglomeration of knowledge-intensive business services and multinational firms. Jacobs et al. (2014) research the coagglomeration of knowledge-intensive business services and multinational firms. They employ the measure of Duranton & Overman (2005) and use firm-level data in the Netherlands from 2000 to 2009. They find that knowledge-intensive business services and multinational enterprises are coagglomerated. Multinational corporations have a substantial impact on the emergence of knowledge-intensive business services, although their impact on such start-ups is far smaller than the favourable impact of previously existing knowledge-intensive business services.

Coagglomeration of exporters and non-exporters in China. He et al. (2012) investigate the spatial agglomeration of exporters and coagglomeration between exporters and non-exporters. They use data from the Annual Survey of Industrial Firms in 2002 and 2007 and employ the agglomeration and coagglomeration indexes of Ellison & Glaeser (1997). They compare the EG agglomeration index for exporters and non-exporters, the EG coagglomeration index between exporters and non-exporters and EG coagglomeration index between exporters and foreign enterprises (without using a regression model). They find that compared with non-exporters, exporters are substantially more geographically agglomerated. They also observe extensive coagglomeration between exporters and non-exporters as well as exporters and foreign firms.

He et al. (2016) further explore agglomeration and coagglomeration of exporters and non-exporters in China, using data from the ASIF from 2002 to 2007. They use the methods of Ellison & Glaeser (1997) to measure the levels of agglomeration and coagglomeration. They utilize a regression model to identify factors in the agglomeration of exporters and coagglomeration between exporters and non-exporters. A self-reinforcing process is pointed out in the agglomeration of exporters and coagglomeration between non-exporters and exporters. Their results also indicate that both exporter agglomeration and non-exporter agglomeration have favourable effects on coagglomeration of exporters and non-exporters.

Coagglomeration of Producer Services and Manufacturing in China. Ke et al. (2014) undertake research on coagglomeration of producer services and the manufacturing sector with data that consist of 286 cities gathered from the China City Statistical Yearbook and China Urban Construction Statistical Yearbook from 2003 to 2008. The 2007 China Input-Output Table is used to capture the producer service sector. The synergy impacts of producer services and manufacturing are demonstrated using a simultaneous equation of coagglomeration of these two sectors. They then use the fixed effects instruments estimation, and find the co-agglomeration of producer services and the manufacturing sector. Their results also suggest that intra-industry agglomeration has spillovers on its own industry in nearby cities.

4.2.3 The Determinants of Coagglomeration

The importance of Marshallian externalities in developed countries

Dumais et al. (1997, 2002) are the first to perform empirical research on the trend of industries co-locating. They explore how spatial concentration emerges from dynamic processes, using data from the Census Bureau's Longitudinal Research Database, and they find that the locations of industrial agglomerations vary over time. They employ experiments that focus on coagglomeration patterns with the underlying purpose of seeing if they will be useful in determining which factors generate intra-industry agglomeration benefits. They examine coagglomeration patterns to evaluate three Marshallian mechanisms of agglomeration: input-output linkage, labour market pooling and intellectual spillovers. They find that agglomeration saves transportation costs by being close to input suppliers or consumers, although the effects are slight. The magnitude of the effects of labour pooling is prominent in reinforcing agglomeration.

Important empirical research on coagglomeration has been done by Ellison et al. (2010), who construct pairwise coagglomeration using confidential firm-level data of US manufacturing industries from the US Economic Census. They quantify industry pair coagglomeration in two ways: the EG metric of coagglomeration and the DO continuous index. Then, they regress the two indexes on the proxies for Marshall's three agglomeration externalities and natural advantages. Ellison et al. (2010) claim that coagglomeration can result from natural advantages that draw two industries together. Based on local cost advantages and industry characteristics, they create a spatial distribution.

Ellison et al. (2010) capture input-output linkage using data from 1987 Benchmark Input-Output Accounts by the Bureau of Economic Analysis. $Input_{i\to j}$ represents the share of industry i's inputs bought from industry j, and range from zero to one. $Ouput_{i\to j}$ is the share of industry i's output that is sold to industry j. The proxies for the interconnection in merchandise between two industries are undirectional forms of variable $Input_{ij} = \max\{Input_{i\to j}, Input_{j\to i}\}$ and $Output_{ij} = \max\{Output_{i\to j}, Output_{j\to i}\}$. They also define a combined $InputOutput_{ij} = \max\{Input_{i\to j}, Output_{i\to j}\}$.

Ellison et al. (2010) use data from the 1987 National Industry-Occupation Employment Matrix by the Bureau of labour Statistics to construct the proxy for labour market pooling of industry pairs. The data contain 277 occupational employment at industry level. They compute the correlation of the share of employment in the same occupation in two industries across occupations.

Ellison et al. (2010) base their work on R&D and patents to reflect information flows. They adopt Frederic M. Scherer's (1984) technology matrix that indicates the flows of R&D activity from one industry to another to construct variable $TechIn_{i\to j}$ and $Techout_{i\to j}$ as $Input_{i\to j}$ and $Ouput_{i\to j}$ described above. They also use patent citation by the National Bureau of Economic Research (NBER) to capture the extent to which technology in industry i cite technology from industry j, and vice versa. Using this in the same way as input-output linkage, they construct $PatentIn_{i\to j}$ and $Patentout_{i\to j}$.

Ellison et al. (2010) find that all three Marshallian externalities are positively related to manufacturing coagglomeration and confirm the importance of shared natural advantages. Shared natural advantages have stronger effects on coagglomeration than Marshallian externalities. Of the three Marshallian external economies input-output linkage is the most crucial, closely followed by labour market pooling. Knowledge spillovers are statistically significant, but they are not as strong as the other factors.

Faggio et al. (2017) examine patterns of coagglomeration in the United Kingdom as well as the micro-foundations of agglomeration economies. They investigate prominent heterogeneity traits across industries in the UK using establishment-level data and the coagglomeration measure of Ellison & Glaeser (1997). They then regress the coagglomeration index on labour pooling, input sharing, and knowledge spillovers and further conduct heterogeneous analysis of different industries, which comprises new industries, industries with high technology and high education, industry structure (size of entrants and size of incumbents).

Faggio et al. (2017) find that input sharing, labour pooling, and knowledge spillovers are all positively associated with coagglomeration. Regarding heterogeneous industries, coagglomeration is not merely a high-tech phenomenon, as evidenced by technology orientation and education traits. High-tech sectors, on the other hand, have higher knowledge spillovers, but low-tech industries display a great deal of evidence of input sharing and labour pooling. Smaller enterprises' agglomeration effects are more potent, especially when it comes to input sharing.

Diodato et al. (2018) research the pattern of coagglomeration over time, and analyse the evolution of the importance of three Marshallian externalities to industrial coagglomeration. They examine coagglomeration at three geographic levels using data from the County Business Patterns in the US and Mexico economic censuses for the years 2003 and 2008. These two datasets are utilized to capture recent coagglomeration trends, while IPUMS USA census samples from 1910 to 2010 are used to capture historical coagglomeration patterns.

Diodato et al. (2018) use the measure for coagglomeration of Ellison & Glaeser (1997) and the OLS estimator and the IV technique to regress the level of coagglomeration on input-output connections, labour market pooling, technical similarity, and natural advantages. To construct instruments, they use analogously constructed variables using data from other countries. The input-output links and labour similarity are computed using Mexican economy statistics, whereas technological linkages are computed using patents from inventors outside the United States.

Diodato et al. (2018) find that for manufacturing sectors, input-output sharing and labour market pooling have a greater impact on coagglomeration than knowledge spillovers, whereas for the service sector, labour pooling has a considerably greater impact than input-output links. In terms of the pattern of industrial coagglomeration over time, they highlight that the effects of input-output linkage on coagglomeration decline dramatically over the years, whereas the importance of labour market pooling remains almost constant.

Diodato et al. (2018) then split the sample into manufacturing and service sectors to investigate these disparities, and find that manufacturing sector's coagglomeration is mainly driven by input-output linkages, whereas the service sector's coagglomeration is more dependent on labour market pooling. They further analyse heterogeneity across industries using 27 sub-sectors and find that different sub-sectors have varied results. In industries such as electronics and the medical sector, input-output linkages are critical for coagglomeration, and both input-output linkages and labour market pooling are essential variables in machinery and hardware manufacture. Additionally, there is considerable variance in different service sub-sectors as well.

The Importance of Marshallian Externalities in Developing Countries

Mukim (2015) investigates the coagglomeration of formal and informal industry in India. The informal sector refers to companies operating in the shadows, uncontrolled by the government, and whose data are not collected to the same extent as the formal sector. In India, the informal sector is quite extensive and continues to grow, employing more than two-thirds of the workforce. Mukim uses the 22 two-digit formal and informal manufacturing data in India for the years 2000 and 2005 and adopts the EG coagglomeration index. Mukim then regresses the level of coagglomeration on buyer-seller linkages (input-output linkages), labour market pool and technological linkages. Mukim concludes that externalities such as buyer-seller linkages and technology spillovers are statistically important for formal-informal co-agglomeration. Mukim also shows that formal-informal coagglomeration is crucial for the birth of small and medium formal firms in India.

Aleksandrova et al. (2020) investigate both intra-industry agglomeration and inter-industry agglomeration in Russia. They use firm-level data for the manufacturing sector from the 2014 RUSLANA database and follow the agglomeration (and coagglomeration) measure of Duranton & Overman (2005). They find that around half of the 4-digit industries and two-thirds of the 3-digit industries are considerably agglomerated. Approximately 70% of industry pairs are significantly coagglomerated, primarily across short distances of less than

100 kilometres. Industries associated with better input-output linkages and more knowledge spillovers yield more coagglomeration, whereas industries with more labour similarities are less coagglomerated in Russia. Their findings also demonstrate that decreased transportation costs result in a higher level of coagglomeration.

Howard et al. (2016) establish a coagglomeration index for developing economies and compare it to other measures available in the literature. Compared with the measure of Ellison & Glaeser (1997), their coagglomeration index is based on the number of enterprises rather than employment. They argue that in developing nations like Vietnam, low-skilled workers make up the majority of the workforce, while high-tech companies with significant knowledge spillovers employ a limited number of workers. As a result, they employ the number of enterprises as a source of agglomeration economies in order to calculate their coagglomeration index. To prevent over-weighting clusters in rural areas in Vietnam, their measure also adjusts for overall distribution across the country rather than density in small areas separately. They use their coagglomeration index to regress coagglomeration on Marshallian externalities and natural advantages, and find that knowledge spillovers are the most significant determinant in Vietnam.

Other Factors that Affect Coagglomeration

Shipping costs and information costs. Gallagher (2013) investigates the impact of shipping costs, information costs and other forces on coagglomeration in the US manufacturing sector with data from the 1997 census of manufacturing, 2002 Commodity Flow Survey, 1997 input-output table, NBER Patent Database and 2002 Occupational Employment Survey. They contribute to developing approaches for capturing shipping and information costs based on inter-industry trade. High shipping costs are expected in industries trading with heavy raw resources, while high information costs are anticipated in industries trading with highly engineered products. They use the coagglomeration measure of Ellison & Glaeser (1997) and regress it on shipping costs, information costs, natural advantages, knowledge spillovers, labour pooling, and input sharing. Their findings reveal that at the metropolitan level, both shipping and information costs affect industrial coagglomeration.

Shared knowledge and coagglomeration of occupations. Gabe & Abel (2016) analyse the coagglomeration of workers in various occupations using data from the 2010 IPUMS, which covers 468 occupations in the United States. The coagglomeration of occupations is measured by the method of Ellison & Glaeser (1997). They then regress the level of occupation coagglomeration on knowledge similarity between occupations (with data from US labour's Occupational Information) and control variables. Their findings suggest that occupations with common knowledge are more inclined to coagglomerate. The shared knowledge related to technology, mathematics, arts, etc. has the greatest impact on occupational coagglomeration in metropolitan regions.

4.2.4 The Effects of Coagglomeration

Impacts on production. Tokunaga & Kageyama (2008) examine the effects of agglomeration and coagglomeration on production in the manufacturing sector in Japan. Plant-level panel data in the 1985,1990,1995 and 2000 Japanese Census of Manufactures are used in their research. They apply the agglomeration and coagglomeration index of Ellison & Glaeser (1997), and find that both agglomeration and coagglomeration positively influence manufacturing production in Japan.

Barrios et al. (2006) investigate the effects of coagglomeration of domestic and foreign firms on the productivity of domestic firms. They use plant-level data in Ireland since 1972 and the coagglomeration measure of Ellison & Glaeser (1997) to generate the coagglomeration index. They regress the total factor productivity and labour demand separately on foreign presence in industries and other control variables. They find a high level of coagglomeration of domestic plants and multinational plants during the sample period. Spillover effects from multinational plants benefit domestic plant productivity and employment, but only in industries where foreign and domestic plants are coagglomerated.

Impacts on carbon intensity. Li et al. (2019) investigate the effects of coagglomeration of producer services and manufacturing on carbon intensity. Their panel data comprises yearly data from the National Bureau of Statistics of China for 30 Chinese provinces from 2009 to 2016. They use a method called location entropy to determine the degree of coagglomeration of producer services and manufacturers. They then adopt the threshold regression model to regress carbon intensity on the level of coagglomeration and four control variables, including government supports, technology market maturity, knowledge spillovers, and R&D inputs. Their findings suggest that resource misallocation restricts the impact of producer services and manufacturing coagglomeration on carbon emission reduction. When there is a suitable allocation of resources, coagglomeration can significantly diminish carbon intensity.

Impacts on knowledge sharing and innovation. Zhang (2015) uses firm-level data from the annual survey by the National Bureau of Statistics of China from 1998 to 2007 to investigate the effects of agglomeration on product innovation (new product). Following Martin et al. (2011), Zhang constructs localization economies and urbanization economies variables to measure agglomeration. The empirical results show that urbanization economies rather than localization economies promote firms' new product output.

Connell et al. (2014) investigate the benefits of industrial clusters for innovation and knowledge spillovers. They use the qualitative research method by interviewing key people within two representative industry clusters in Dubai. The Partner Relationship Managers in each of the two free-zone clusters choose the individuals who will be interviewed. Within the two industry clusters, a total of 18 interviews were carried out, ranging from small, medium, and big businesses. They demonstrate that in Dubai, two of the four-diamond factors identified by Porter (1990) are applicable. The competition of other firms, as well as connected and supporting industries, are major determinants. They claim that industrial clusters boost sharing of knowledge and collabourative innovation greatly.

4.2.5 Summary of the Literature Review

The literature review includes the theory of coagglomeration and empirical research on coagglomeration. The empirical research incorporates coagglomeration patterns, the drivers of coagglomeration and its effects. Compared with agglomeration, there is much less theoretical and empirical research on coagglomeration.

Regarding theory, the cluster theory of Porter (1990) is related to coagglomeration, but it does not offer formal explanations of coagglomeration. Helsley & Strange (2014) present a rigorous, comprehensive analysis of coagglomeration with a model of city composition. They demonstrate that coagglomeration may not benefit the industries that concentrate together as long as the current city provides adequate utility for workers. Since no worker is willing to relocate first, complementary industries with reciprocal benefits may not opt to coagglomerate.

O'Sullivan & Strange (2018) build an agent-based model to mimic the emergence of coagglomeration. They also use a city composition model and contribute by adding the agent-based model to it. They suggest a positive, nonlinear relationship between inter-industry externalities and coagglomeration. Their findings imply that coagglomeration is influenced by factors such as history, size of firm, and relocation costs.

With respect to empirical research, some research focuses on general coagglomeration of all industry pairs (Ellison et al. 2010), while some investigates the coagglomeration of certain industries such as exporters and foreign enterprises (He et al. 2016), domestic firms and foreign firms (Barrios et al. 2006), intermediate producer services and manufacturers (Ke et al. 2014), informal industry and formal industry which only exist in some countries (Mukim 2015).

Most research on the drivers of coagglomeration focuses on three Marshallian forces and natural advantages (Ellison et al. 2010, Diodato et al. 2018, Howard et al. 2016). A few studies investigate transport costs or shipping costs (Gallagher 2013). There is not much research on the effects of coagglomeration, including the impact on productivity (Tokunaga & Kageyama 2008, Barrios et al. 2006) and on carbon intensity (Li et al. 2019). Due to the lack of prior

literature on coagglomeration, some important literature is accorded more attention in this literature review. The construction of highways is expected to be related to the spatial concentration of industries. Highways shorten the time needed to transport intermediate goods and reduce transport costs, which may lead to the co-location of industries that share ideas and labour. Additionally, the construction of highways attracts firms to locate near them. Nevertheless, research that investigates the relationship between coagglomeration and highways is scarce, and this study fills the gap by exploring the effects of highway access on pairwise coagglomeration.

4.3 Hypothesis Development

4.3.1 Highway Access and Coagglomeration

Building on the theoretical framework established by Belleflamme et al. (2000), which explores the dynamics of firm location decisions and cluster formation, this study proposes a hypothesis concerning the relationship between highway access and co-agglomeration. Belleflamme et al. (2000) developed an equilibrium model to analyse the conditions under which firms will co-locate in a single region, focusing on the impact of transport costs, product differentiation, and localisation economies on cluster emergence. According to Belleflamme et al. (2000), as transportation costs decrease, firms are more likely to co-locate, which fosters agglomeration and maximizes the benefits of being situated in close proximity to similar firms. This effect is expected to be more pronounced in the presence of substantial localization economies, where reduced transportation costs lead to greater cost savings.

In this model by Krugman (1991), the reduction in transport costs diminishes the relative cost advantage of spreading production across multiple regions, thereby encouraging firms to locate in a single location. Krugman (1991) highlights that, as transport costs decrease, the spatial concentration of manufacturing industries becomes more pronounced, driven by economies of scale and the home market effect. In Krugman (1991), transport cost reductions

initially promote firm concentration by decreasing the relative costs of spreading production, approaching a point where further reductions yield diminishing effects on spatial concentration, whereas Belleflamme et al. (2000) suggests that as transport costs decrease, the transition from dispersed distribution to agglomeration occurs more smoothly and continuously.

In light of this, the hypothesis developed for this study is as follows: Improved highway access, by reducing transportation costs, enhances the likelihood of coagglomeration among firms. Lower transport costs facilitate the clustering of firms, thereby amplifying the benefits of cost-saving.

4.3.2 Input-output Linkages and Coagglomeration

In analysing the impact of highway infrastructure on industry coagglomeration, it is hypothesised that, for industry pairs with higher input-output linkages, improvements in highway access diminish the influence of these linkages on coagglomeration. Input-output linkages refer to the degree of interdependence between firms based on their supply and demand relationships. When these linkages are strong, firms in an industry pair are typically compelled to co-locate initially to minimise transportation costs and ensure timely delivery of inputs and outputs. Consequently, prior to any infrastructural improvements, the need for proximity due to input-output linkages would drive firms to establish themselves in close geographical proximity.

However, as highway access improves, the cost advantages of proximity diminish, allowing firms to decentralise their locations while still maintaining efficient supply chain interactions. Enhanced transportation infrastructure reduces the logistical burden of maintaining close physical proximity, thus decreasing the necessity for firms to co-locate as a result of input-output linkages. This shift suggests that with better highway access, firms can expand their spatial distribution and still benefit from efficient input-output relationships, leading to a reduction in the role of these linkages in determining coagglomeration patterns. Therefore, the hypothesis posits that improvements in highway access reduce the significance of input-output linkages on coagglomeration, as firms gain the flexibility to locate further from their suppliers and still maintain effective operational interactions.

Bilateral Input-output adjusted Highway Access. The proposed hypothesis centres on the relationship between improvements in highway access and the coagglomeration levels of industry pairs, as captured by the Input-Output adjusted Highway Access metric. This metric evaluates how the enhanced connectivity via highways influences the transport costs between industries and, consequently, their spatial agglomeration. Specifically, it is hypothesised that industries which engage in substantial inter-industry transportation of goods will experience a significant reduction in transport costs due to highway improvements. This reduction in costs would, in turn, decrease the necessity for these industries to remain in close proximity to one another. Conversely, for industry pairs that engage in minimal inter-industry transportation, the impact of improved highway access on their spatial relationship would be negligible, as their transport costs are already low.

Thus, the hypothesis posits that a substantial decrease in transport costs resulting from improved highway access will lead to a lower coagglomeration level between industries that transport large volumes of goods between them. This is because the reduction in transport expenses diminishes the economic incentive for these industries to cluster geographically. On the other hand, industries with minimal transport interactions will remain relatively unaffected by highway improvements in terms of their spatial arrangement.

4.3.3 Related Industries and Coagglomeration

This study posits that the impact of highway access on the coagglomeration of related industries is more substantial compared to its effect on less interconnected sectors. Manufacturing industries that share input-output linkages, possess a common skilled labour pool, and benefit from knowledge spillovers are predisposed to co-locate to reduce production costs and stay competitive with technological advancements (Ellison et al. 2010). In China, this trend is evident as industries involved in different stages of the value chain or those using shared inputs often cluster together (Dai et al. 2021). The benefits of such clustering, including enhanced bargaining power and reduced input costs, are well-documented. For example, research on the cashmere sweater cluster in Puyuan shows a high degree of coagglomeration among industries at various stages of production (Ruan & Zhang 2009).

Given these observations, it is hypothesised that highway access will have a more pronounced effect on the coagglomeration of related industries than on less connected ones. Improved highway infrastructure facilitates easier and more cost-effective transportation of goods and resources, thereby amplifying the clustering effect among industries that are already inclined to co-locate due to their shared inputs and value chain positions. Thus, related industries are expected to exhibit a greater tendency to coagglomerate in response to enhanced highway access, reflecting the advantages gained from improved connectivity and reduced logistical barriers.

4.3.4 State-Owned Enterprises and Coagglomeration

In examining the relationship between highway infrastructure and industrial coagglomeration, it is hypothesised that the effect of highways on coagglomeration is diminished in industries with a high share of state-owned enterprises. According to Lu & Tao (2009), local protectionism, as indicated by the share of SOEs, hinders within-industry agglomeration. The hypothesis posits that highways have a relatively minor impact on coagglomeration within industries where SOEs constitute a significant proportion. This is primarily because many SOEs are established prior to the construction of highways, resulting in their locations being largely unaffected by subsequent infrastructural developments. In the ASIF dataset over the period 1998 to 2007, the average age of SOEs is significantly older years compared to non-SOEs, suggesting that SOEs' locations are less responsive to the changes brought about by highways. Consequently, the presence of highways may not significantly influence the coagglomeration patterns in these industries.

Furthermore, non-SOEs, which are typically more recent and flexible in their location choices, tend to benefit more from highway infrastructure, leading to greater coagglomeration. This dynamic is consistent with the observation that non-SOEs are more likely to co-locate in response to improved transportation networks, whereas the established locations of SOEs may deter new agglomerative patterns. The findings of Lu & Tao (2009), based on data from ASIF dataset and the Ellison and Glaeser measure of agglomeration, demonstrate that industrial ag-

glomeration is inversely related to the share of SOEs. This suggests that local protectionism, as represented by a high share of SOEs, is a significant barrier to industrial agglomeration, further supporting the hypothesis that highways have a reduced effect on coagglomeration in SOE-dominated industries.

4.4 Data and Methodology

4.4.1 Data

The summarized dataset used in this chapter is as follows. Two main datasets are used to compute the key variables. First is the Annual Survey of Industrial Firms, which collects firm-level data in the manufacturing industries. It is used to calculate the coagglomeration index. The address information and other data in the ASIF help the calculation of highway access and control variables. The GIS highway routes are obtained from the ACASIAN Dataset, which is also used by Faber (2014) when investigating the effect of highway networks on economic activity in peripheral regions. The ACASIAN contains GIS information on highways in China.

There are other datasets that are used to generate the controls. The 1997, 2002 and 2007 China Input-output tables are used to capture the input-output linkage of industry pairs and the use of natural resources. These input-output tables are mainly for 3-digit industries in manufacturing after being converted to GB2002 industry classification by this study. The China 2000 census from IPUMS is used to capture the labour similarity of industry pairs, which is at 2-digit industry level. The 1999, 2002 and 2007 US occupation-industry matrix contains the information that is used to capture the labour pooling of industry pairs, which are needed to be converted to industry pairs in China. Additionally, the dataset that matches China's State Intellectual Property Office (SIPO) patents for firms in the ASIF from He et al. (2018) is adopted to capture the knowledge spillovers of industry pairs.

4.4.2 Key Variables

Coagglomeration measures

The most used coagglomeration measure is that of Ellison & Glaeser (1997). Pairwise coagglomeration examines the correlation between two industries in the same region. Regions are predefined administrative districts that are not spatially continuous, which is also a limitation of this measurement. The coagglomeration formula for industry pairs (two industries) using by Ellison & Glaeser (1997) is as follows:

$$\gamma^{c}_{i,j} = \frac{\sum_{r=1}^{R}(s_{ri} - x_r)(s_{rj} - x_r)}{1 - \sum_{r=1}^{R} x_r^2} \tag{4.1}$$

Where $\gamma^{c}_{i,j}$ is the EG coagglomeration value of industry i and industry j ($i \neq j$). s_{ri} is the share of industry i's employment in region r. s_{rj} is the share of industry j's employment in region r. x_r represents the aggregate size of region r. x_r is modelled as the mean employment share in region r across all industries by Ellison et al. (2010). This study follows their method for calculating x_r. The EG coagglomeration index takes the region size into account to generate the coagglomeration degree of industry pairs. A higher γ^{c} means the industry pairs are more concentrated together.

To clarify how the coagglomeration formula in Equation 4.1 is applied at the three aggregation levels (province, city, and county), the following approach is used. Regarding the province Level: the formula calculates the coagglomeration index $\gamma^{c}_{i,j}$ by aggregating employment shares s_{ri} and s_{rj} for each industry i and j within each province. The aggregate size x_r is the mean employment share for all industries within the province. Similarly, at the city level, the employment shares s_{ri} and s_{rj} are computed for industries i and j within each city. The aggregate size x_r is the mean employment share in each city. At the county level, the same approach is applied and x_r is calculated as the mean employment share in each county.

Highway access for pairwise industries

The formula for the highway access variable for industry pairs is:

$$highway\ access_{ij} = \frac{1}{\sqrt{distance_{i,highway} * distance_{j,highway}}}, for\ i \neq j \tag{4.2}$$

where $distance_{industry_i,highway}$ and $distance_{industry_j,highway}$ are the weighted mean of Euclidean distance from the nearest highways to firms of industry i and j respectively. The weights are calculated by the share of employment of firms in the industry. The distance from a firm to its nearest highway are computed by ArcGIS. The formula for the weighted distance between industry i and the highway network is as follows:

$$distance_{industry_i,highway} = \sum_{k=1}^{n} distance_{firm_k,nearest\ highway} * weight_k \tag{4.3}$$

where firm k is in industry i. The formula for $weight_k$ calculated by employment is: $weight_k = \frac{employment\ in\ firm_k}{employment\ in\ industry_i}$.

Bilateral Input-Output adjusted Highway Access Measurement

The Input-Output adjusted Highway Access (IOHA) metric is designed to quantify the highway access between industry pairs, taking into account both the size of the transportation flows and the distance to the nearest highway. The Bilateral IOHA, denoted as $IOHA_{ij}^{Bilateral}$, is defined by the following equation:

$$IOHA_{ij}^{Bilateral} = s_{j\to i} \cdot \left(-\frac{D_i + D_j}{2}\right) + s_{i\to j} \cdot \left(-\frac{D_i + D_j}{2}\right) = (s_{j\to i} + s_{i\to j}) \cdot \left(-\frac{D_i + D_j}{2}\right) \tag{4.4}$$

In this equation, $s_{i\to p}$ represents the output size of industry i destined for industry p, indicating the volume of goods transported from industry i to industry p. Conversely, $s_{p\to i}$ reflects the input size of industry i sourced from industry p, quantifying the amount of goods received by industry i from industry p. The distance D_i denotes the proximity of industry i to the nearest highway, and D_j represents the distance of industry j to its nearest highway. The term $-\frac{D_i+D_j}{2}$ adjusts the measure to account for the average distance to the nearest highway for both industries. By incorporating these variables, the Bilateral IOHA metric captures the combined effect of transportation volume and highway proximity on the inter-industry access.

4.4.3 Model Specification

The baseline model

The baseline regression model for investigating the effect of highway access on coagglomeration is:

$$\gamma^{c}_{ijt} = \alpha + \beta_1 highway\ access_{ijt} + \beta_2 X_{ijt} + \delta_{ij} + \partial_t + \varepsilon_{ijt} \tag{4.5}$$

where γ^{c}_{ijt} is the coagglomeration level of industry i and j at time t ($i \neq j$). $highway\ access_{ijt}$ is the reciprocal of square root of the weighted average distance from industry i to highway multiplies the weighted distance from industry j to highway, $1/\sqrt{DiDj}$. The control variables are the proxies for input-output linkage, labour pooling, knowledge spillover and natural advantages. δ_{ij} is the industry effect. ∂_t is the year effect. ε_{ijt} is the error term. The dependent, independent and control variables' name, definition, name in the regression are shown in Table 4.1.

Table 4.1: Variable definition

Variable name	data	Definition	name in regression
EG coagglomeration index (province-level)	ASIF	EG coagglomeration index for 3-digit industry pairs at province level	coagg_province
EG coagglomeration index (city-level)	ASIF	EG coagglomeration index for 3-digit industry pairs at city level	coagg_city
EG coagglomeration index (county-level)	ASIF	EG coagglomeration index for 3-digit industry pairs at county level	coagg_county
highway access for industry pairs	ACASIAN;ASIF	1/sqrt (weighted mean of distance from industry i to highways*weighted mean of distance from industry j to highways)	highway access
Input output linkage	1997, 2002 and 2007 IO table	max(inputij,inputji)	input linkage
Labor market pooling	China 2000 census;1999, 2002 and 2007 US occupation-industry matrix	US occupation correlation at 3 digit industry* China occupation correlation at 2 digit industry	labor_pooling
Knowledge spillovers	SIPO&ASIF	Patent similarity between industry pairs	Knowledge spillovers
Natural advantages:			
agriculture	1997, 2002 and 2007 IO table	$\sqrt{A_i A_j}$	agriculture
petroleum	1997, 2002 and 2007 IO table	$\sqrt{P_i P_j}$	petroleum
coal	1997, 2002 and 2007 IO table	$\sqrt{C_i C_j}$	coal

Control variables construction

Regarding the control variables, input-output linkage, labour pooling, knowledge spillovers and natural advantages are all expected to have positive effects on coagglomeration. These control variables are constructed as follows.

Input-output linkage

This study uses the direct consumption coefficient to measure the input-output linkage of pairwise industry with data from the China IO table. $Input_{i\to j}$ represents the share of industry i's inputs bought from industry j, and range from zero to one. The proxy for input-output linkage between an industry pair is $Input\_output\ linkage_{ij} = \max\{Input_{i\to j}, Input_{j\to i}\}$. There are three (1997, 2002 and 2007) China IO tables that can be used for the sample period 1998-2007. To make the time-variant variables, this study assigns the 1997 IO table for samples in 1998-2000, the 2002 IO table for samples in 2001-2003, and the 2007 IO table for samples in 2004-2007. The China IO table only contains information about three-digit industries, thus this study conducts regression for three-digit industry pairs.

Labour market pooling

Ellison et al. (2010) use data from the 1987 National Industry-Occupation Employment Matrix from the Bureau of labour Statistics to construct the proxy for labour market pooling of industry pairs. They compute the correlation of the share of employment in the same occupation in two industries across occupations. The rationale is that occupation correlations between different industries tend to be similar across countries. In order to use this data in research on China, Ding et al. (2019) multiply correlation value with the weight from China's 2002 Input-Output table to measure the labour-structure similarity of upstream industries and downstream industries.

This study uses 1999, 2002 and 2007 industry-specific Occupational Employment Statistics data from the Bureau of labour Statistics for samples in 1998-2000, 2001-2003 and 2004-2007, respectively. The industry-specific Occupational Employment Statistics data in 1999 employ the SIC system; data in 2002 and 2007 use the 4-digit North American Industry Classification System (NAICS). In order to make the US data concordant with The China industry classification, this study converts the 2002 NAICS code (and 1987 SIC code) to

1997 NAICS code with a 2002 NAICS to 1997 NAICS concordance table (and a 1987 SIC to 1997 NAICS concordance table) from US Census Bureau. Then this study converts 1997 NAICS to China GB2002 industry code with a concordance table from Ma et al. (2014), which is a 6-digit NAICS to 4-digit GB concordance table.

Since the correlation of occupations come from US data, this study adjusts it with the China 2000 Census data from IPUMS, which provides an occupation-industry table of individuals in 2-digit China GB1994 industry code. The correlation coefficients of occupations in 2-digit industry in China is multiplied by the US correlation coefficients.

$$labour\ pooling_{i,j} = US\ occupation\ correlation_{i,j} * China\ occupation\ correlation_{t,k} \quad (4.6)$$

where i, j is the 3-digit industry code ($i \neq j$). t, k is the two-digit industry code. i belongs to t, j belongs to k. The occupation correlation for $industry\ pair_{ij}$ is calculated with the formula

$$Occ\ correlation_{i,j} = \frac{\sum(occ\ share_i - \overline{occ\ share_i})(occ\ share_j - \overline{occ\ share_j})}{\sqrt{\sum(occ\ share_i - \overline{occ\ share_i})^2 \sum(occ\ share_j - \overline{occ\ share_j})^2}} \quad (4.7)$$

Knowledge spillovers

In this chapter, knowledge spillovers are captured by similar patents created by different industries. Ellison et al. (2010) have two measures for knowledge spillovers. One is Frederic M. Scherer's (1984) technology matrix that indicates the flows of R&D activity from one industry to another. The other is the patent citation by the National Bureau of Economic Research to capture the patent citation number of industry pairs. Due to the lack of R&D and patents citation flow between industries as with Ellison et al. (2010), this chapter uses the number of patents that are in the same sub-classification created by industry pairs in each year to measure knowledge spillovers. This study adopts the dataset from He et al. (2018) who match SIPO patents to firms in the ASIF, which also is used by Chen et al. (2018) when exploring the effects of exporting on firm innovation. Chen et al. (2021) also follow the method by He et al. (2018) when dealing with patents data to investigate the benefits of R&D and patents in China.

Patents in China have three major categories: design patents, invention patents and utility model patents. The design patent refers to a new design based on colour, shape, and pattern of products. The utility model patent refers to a new technical solution for a new product regarding its shape or structure or both. The invention patent refers to new technical solutions proposed for new products, methods or improvements, and it is the patent for which it is the most difficult to get the license. Invention patents and utility model patents use the International Patent Classification, and design patents adopt the Locarno Classification. This study uses all three types of categories, and counts the number of patents that are in the same sub-classification (same first-four patent classification number) for an industry pair each year to calculate the knowledge spillovers.

Natural advantages

The joint natural inputs shares of pairwise industries are used to capture the natural resources advantages of industry pairs. If industry i and industry j both rely on the same natural resources, they are likely to co-locate near resources. The proxies include joint agricultural input share, petroleum input share and coal input share using data from the 1997, 2002 and 2007 China IO table for samples in 1998-2000, 2001-2003 and 2004-2007 respectively. Taking agricultural input share as an example, the agricultural input reliance for industry i and industry j is calculated as the $\sqrt{agricultural\ inputs\ share_i * agricultural\ inputs\ share_j}$. A higher value means that the industry pair depends more on natural resources.

4.5 Summary Statistics and Baseline Results

4.5.1 Summary Statistics of Baseline Model Variables

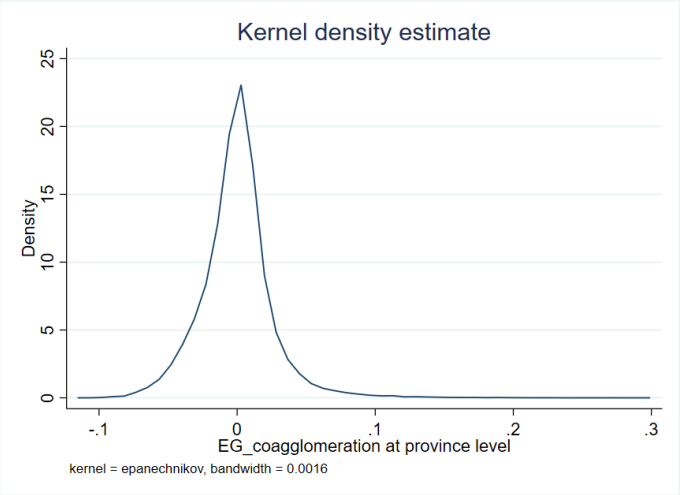
EG coagglomeration analysis

This section shows the summary statistics of baseline model variables. Table 4.2 shows the average EG coagglomeration index of Chinese manufacturing industries over the sample period (1998-2007) for 4-, 3- and 2-digit industries at county, city and province levels, respectively. Table 4.2 also cites the US Pairwise EG coagglomeration at 3-digit industry level of Ellison et al. (2010) with 1987 manufacturing data to compare with the results of this study.

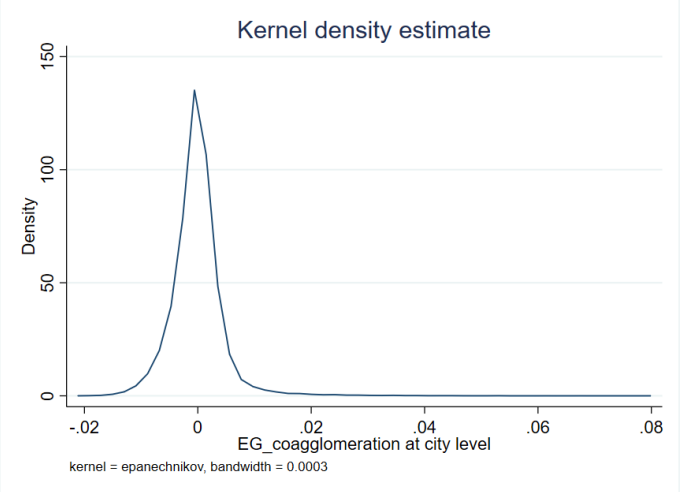
The mean value of EG coagglomeration each year is approximately zero decided by the equation of the EG coagglomeration index. Unlike EG agglomeration within industry, the mean EG coagglomeration is very close to zero, which is reasonable as explained by Ellison et al. (2010). According to the formula of the EG coagglomeration measure, the benchmark is the average size of each region, which is the mean employment share in a region across all industries. The average value of deviations of all industries from the benchmark is approximately zero.

The mean coagglomeration is further explained by the kernel density estimations as shown in Figure 4.1, which presents the kernel density plots of coagglomeration for 3-digit pairwise industries at province, city and county levels. The area under the kernel density plot is 1, which is the total probability. The peaks show that the values of EG pairwise coagglomeration at province, city and county levels concentrate around zero, indicating that the majority of coagglomeration levels are close to zero. This reflects that the mean coagglomeration is approximately zero in the EG coagglomeration measure.

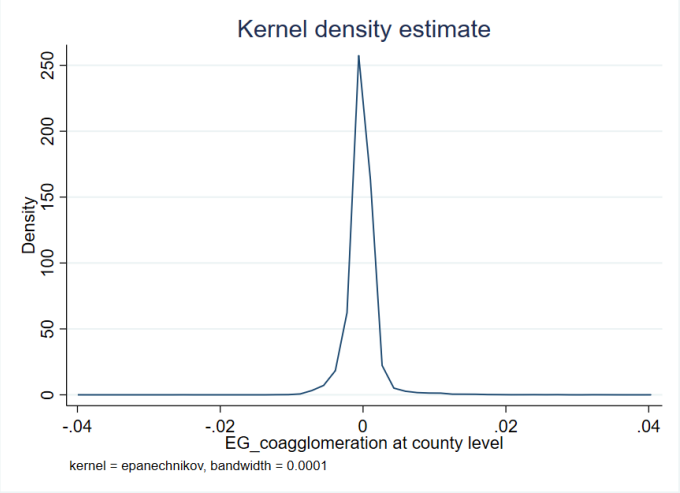
The mean pairwise coagglomeration indices over the sample period do not increase, which is different from the mean EG agglomeration within industry (i.e. the upward trend of mean agglomeration within industry in China from 1998 to 2007). The mean of the coagglomeration index is around zero each year. The coagglomeration index is measured as the deviation of each industry from the benchmark, where the benchmark (average size of each region) adjusts each year. Thus the mean coagglomeration does not change and stays approximately zero over time.



(a) kernel density estimation of coagglomeration at province level



(b) kernel density estimation of coagglomeration at city level



(c) kernel density estimation of coagglomeration at county level

Figure 4.1: Kernel density plots of coagglomeration for 3-digit pairwise industries

Table 4.2: Descriptive statistics for the EG coagglomeration index

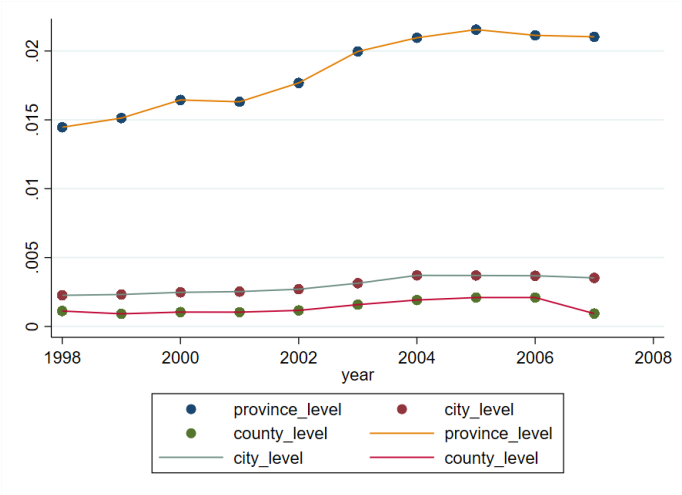
	Mean	Std. Dev.	Min	Max
This study				
4-digit industry				
province	0.000	0.032	-0.168	0.856
city	0.000	0.002	-0.017	0.196
county	0.000	0.001	-0.011	0.213
3-digit industry				
province	0.000	0.026	-0.114	0.298
city	0.000	0.005	-0.021	0.080
county	0.000	0.002	-0.040	0.040
2-digit industry				
province	-0.001	0.019	-0.055	0.095
city	0.000	0.003	-0.010	0.018
county	0.000	0.001	-0.006	0.012
Ellison, Glaeser & Kerr (2010)				
1987 Census of Manufacturers data				
3-digit industry				
state	0.000	0.013	-0.065	0.207
city	0.000	0.006	-0.025	0.119
county	0.000	0.003	-0.018	0.080

The standard deviation of the EG coagglomeration index implies the deviation of the co-agglomeration of industry pairs and the extent to which pairs coagglomerate positively or negatively. As shown in Table 4.2, the standard deviation at province level is larger than at city level, and the standard deviation at the county level is the smallest. This means that at a larger geographic level, the difference of pairwise coagglomeration is more significant. The minimum values are all negative. Ellison et al. (2010) indicate that negative EG coagglomeration values arise when industry pairs are agglomerated in different regions. Regarding different industrial classification levels, the standard deviation of 4-digit industry is the largest and 2-digit is the smallest in general. This reflects that more specific industrial classification generates a larger difference in pairwise coagglomeration.

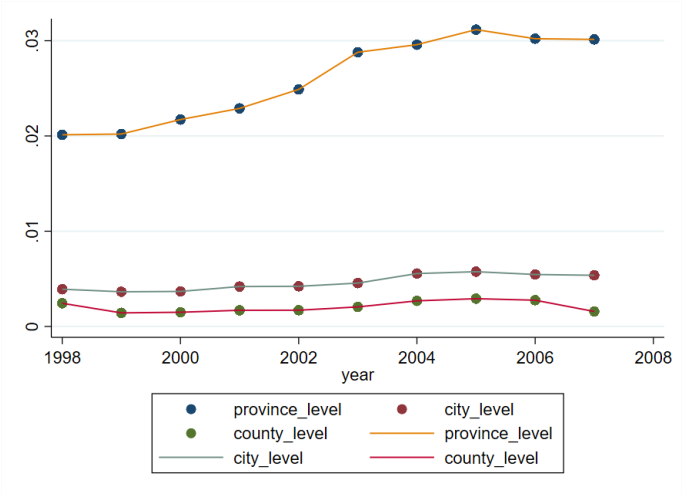
Figure 4.2 shows the standard deviation of coagglomeration for 2-, 3- and 4-digit industry pairs from 1998 to 2007. The standard deviations increase at province level and are stable at city and county levels. This indicates that at province level, industry pairs coagglomerate more or disperse more over time; at city and county level, the pattern does not change from 1998 to 2007. This is probably because the provincial government has more autonomy than the city and county levels. Coagglomeration at the province level is more likely to be affected by the policy of provincial government.

The ten most coagglomerated 3-digit industry pairs at province, city and county level are shown in Tables 4.3, 4.4 and 4.5, respectively. The majority of industry pairs repeatedly appear at three geographic levels. In the top 10 3-digit industry pairs, most industries manufacture electronic equipment, instrumentation, cultural and office machinery and toys. There are overlaps in the industry pairs, for example, among the top ten industry pairs at city level, computer manufacturing (industry code 404) coagglomerates with cultural and office machinery manufacturing (code 415, ranked first), home audio-visual equipment manufacturing (code 407, ranked third), watch and timing instrument manufacturing (code 413, ranked eighth) and other electronic equipment manufacturing (code 409, ranked ninth). Home audio-visual equipment manufacturing (code 407) also coagglomerates with cultural and office machinery manufacturing (code 415), watch and timing instrument manufacturing (code 413) and other electronic equipment manufacturing (code 409). This indicates that these industries are highly likely to co-locate as a group rather than in pairs.

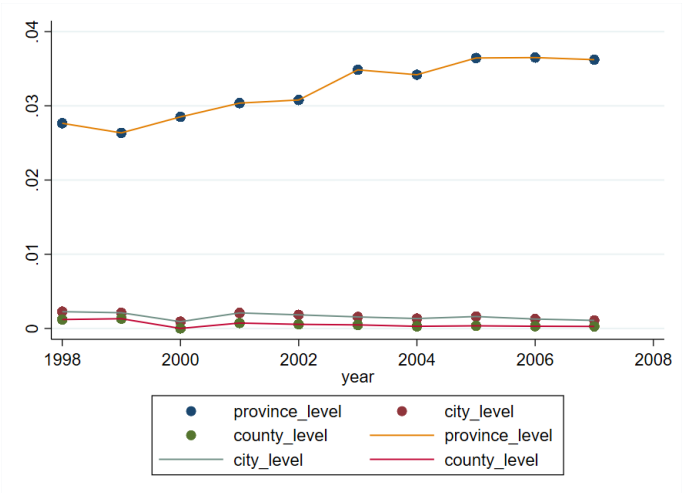
Some industry pairs belong to the same 2-digit categories, such as electronic equipment manufacturers (pairs code 404-407, 407-409 and 404-409); they might have similar inputs, a similar labour pool, and similar technologies. The coagglomeration of toy, other plastic products and some electronic manufacturers is possibly because they have similar inputs such as plastic materials.



(a) SD of coagglomeration for 2-digit industry pairs



(b) SD of coagglomeration for 3-digit industry pairs



(c) SD of coagglomeration for 4-digit industry pairs

Figure 4.2: The standard deviation of coagglomeration from 1998 to 2007

Table 4.3: Highest pairwise coagglomerations at province level

Rank	Industry 1 code	name	Industry 2 code	name	$\gamma^c_{i,j}$
1	244	Toy manufacturing	407	Home audio-visual equipment manufacturing	0.217
2	407	Home audio-visual equipment manufacturing	413	Watch and timing instrument manufacturing	0.202
3	244	Toy manufacturing	413	Watch and timing instrument manufacturing	0.193
4	407	Home audio-visual equipment manufacturing	415	Cultural and office machinery manufacturing	0.190
5	244	Toy manufacturing	415	Cultural and office machinery manufacturing	0.188
6	413	Watch and timing instrument manufacturing	415	Cultural and office machinery manufacturing	0.168
7	309	Other plastic products manufacturing	407	Home audio-visual equipment manufacturing	0.167
8	244	Toy manufacturing	309	Other plastic products manufacturing	0.161
9	309	Other plastic products manufacturing	413	Watch and timing instrument manufacturing	0.151
10	348	Stainless steel and daily-use metal products manufacturing	407	Home audio-visual equipment manufacturing	0.151

Note: the industry code is the industry classification system GB/T 4754-2002.

Table 4.4: Highest pairwise coagglomerations at city level

Rank	Industry 1 code	name	Industry 2 code	name	$\gamma^c_{i,j}$
1	404	Computer manufacturing	415	Cultural and office machinery manufacturing	0.041
2	407	Home audio-visual equipment manufacturing	415	Cultural and office machinery manufacturing	0.040
3	404	Computer manufacturing	407	Home audio-visual equipment manufacturing	0.038
4	407	Home audio-visual equipment manufacturing	413	Watch and timing instrument manufacturing	0.038
5	413	Watch and timing instrument manufacturing	415	Cultural and office machinery manufacturing	0.036
6	244	Toy manufacturing	407	Home audio-visual equipment manufacturing	0.035
7	309	Other plastic products manufacturing	407	Home audio-visual equipment manufacturing	0.032
8	404	Computer manufacturing	413	Watch and timing instrument manufacturing	0.032
9	404	Computer manufacturing	409	Other electronic equipment manufacturing	0.032
10	407	Home audio-visual equipment manufacturing	409	Other electronic equipment manufacturing	0.030

Note: the industry code is the industry classification system GB/T 4754-2002.

Table 4.5: Highest pairwise coagglomerations at county level

Rank	Industry 1 code	name	Industry 2 code	name	$\gamma^c_{i,j}$
1	244	Toy manufacturing	407	Home audio-visual equipment manufacturing	0.023
2	407	Home audio-visual equipment manufacturing	415	Cultural and office machinery manufacturing	0.020
3	404	Computer manufacturing	407	Home audio-visual equipment manufacturing	0.019
4	309	Other plastic products manufacturing	407	Home audio-visual equipment manufacturing	0.018
5	407	Home audio-visual equipment manufacturing	409	Other electronic equipment manufacturing	0.017
6	244	Toy manufacturing	309	Other plastic products manufacturing	0.017
7	309	Other plastic products manufacturing	409	Other electronic equipment manufacturing	0.016
8	182	Textile fabric shoe manufacturing	244	Toy manufacturing	0.016
9	244	Toy manufacturing	415	Cultural and office machinery manufacturing	0.016
10	244	Toy manufacturing	409	Other electronic equipment manufacturing	0.015

Note: the industry code is the industry classification system GB/T 4754-2002.

Highway access for industry pairs

Table 4.6 shows the summary statistics of highway access for 3-digit industry pairs yearly. The mean, minimum and maximum values all increase, which indicates that highway access for industry pairs improved over the sample period. Figure 4.3 more vividly shows the trend of highway access with the mean values from 1998 to 2007.

Table 4.6: Summary statistics of highway access for 3-digit industry pairs

year	mean	sd	min	max
1998	0.0461	0.0176	0.0143	0.1817
1999	0.0489	0.0180	0.0127	0.1545
2000	0.0629	0.0235	0.0163	0.1951
2001	0.0681	0.0249	0.0165	0.2153
2002	0.0823	0.0303	0.0190	0.2214
2003	0.0751	0.0307	0.0133	0.2367
2004	0.0946	0.0369	0.0209	0.2698
2005	0.1186	0.0418	0.0234	0.2777
2006	0.1190	0.0401	0.0249	0.2751
2007	0.1265	0.0452	0.0239	0.3101

Note: This table shows highway access for 3-digit industry pairs from 1998 to 2007.

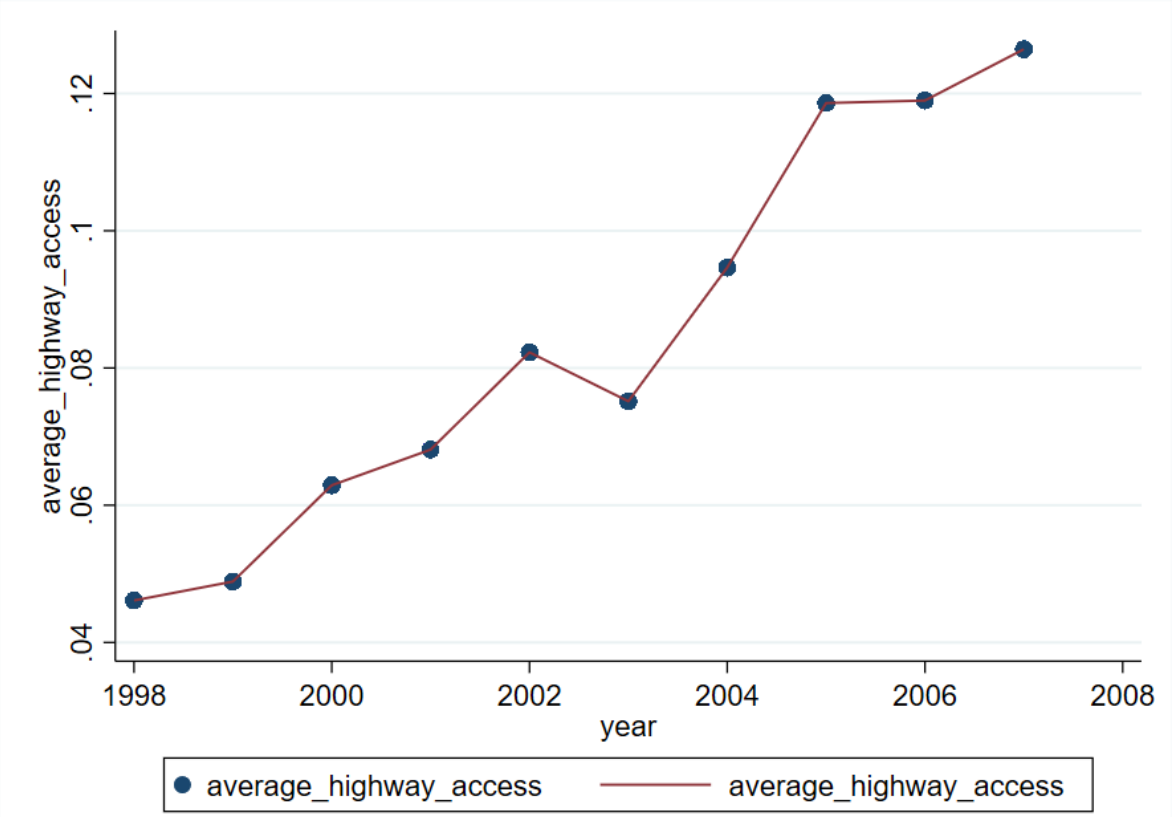


Figure 4.3: Average highway access from 1998 to 2007

Summary statistics for all variables

The correlation matrix is shown in Table 4.7. Besides EG coagglomeration at province, city and county levels (three dependent variables), there are no very correlated variables.

Summary statistics of variables used in the baseline model are displayed in Table 4.8. EG coagglomeration indices are calculated at province, city and county levels.

Table 4.7: Correlation coefficients of all variables in the baseline model

	EG province	EG city	EG county	highway acess	knowledge	input_output	labor_pool
EG province	1.0000						
EG city	0.7153	1.0000					
EG county	0.5589	0.7501	1.0000				
highway_access	0.0559	0.0534	0.0328	1.0000			
knowledge	0.0797	0.1023	0.0990	0.3946	1.0000		
input_ouput	0.0860	0.0944	0.0627	0.0625	0.1199	1.0000	
labor_pool	0.1210	0.1377	0.0888	0.1007	0.2300	0.4259	1.0000
agriculture	0.0741	0.0778	0.0490	-0.1645	0.0760	0.0256	0.1544
petroleum	0.0238	0.0259	0.0224	0.0137	0.0538	0.0349	0.0720

	agriculture	petroleum	coal
agriculture	1.0000		
petroleum	-0.0215	1.0000	
coal	-0.0325	0.1335	1.0000

Table 4.8: Summary statistics of variables used in the baseline model

Variable	Obs	Mean	Std. Dev.	Min	Max
EG coagglomeration (province)	128,166	-0.0004	0.0263	-0.1141	0.2976
EG coagglomeration (city)	128,166	-0.0002	0.0047	-0.0210	0.0797
EG coagglomeration (county)	128,166	-0.0001	0.0021	-0.0399	0.0403
highway_access	128,166	0.0843	0.0427	0.0127	0.3101
knowledge_spillover	128,166	2.1518	2.1156	0.0000	9.8330
input_ouput linkage	127,686	0.0183	0.0450	0.0000	0.6475
labor_pooling	128,166	0.1378	0.2769	-0.0463	1.0000
agriculture	127,686	0.0076	0.0336	0.0000	0.5881
petroleum	127,686	0.0013	0.0076	0.0000	0.6573
coal	127,686	0.0032	0.0067	0.0000	0.5531

4.5.2 The Baseline Results

The pooled OLS and fixed effect estimators are used to estimate the baseline model. In Table 4.9, Columns (1), (2) and (3) show the results of pooled OLS for the highway access variable, at province, city and county levels respectively; Columns (4), (5) and (6) present the results for fixed effect estimations. The variable of interest, highway access, obtains positive and statistically significant estimated coefficients using pooled OLS and FE at all geographic

levels. The coefficients are at the 1% significance level at province and city geographic levels in FE estimation and at all geographic scopes in the pooled OLS. This indicates that highway accessibility for industry pairs is positively associated with the level of coagglomeration for manufacturing industries.

Table 4.9: Pooled OLS and FE results for coagglomeration

	Pooled OLS			FE		
	coagg (province) (1)	coagg (city) (2)	coagg (county) (3)	coagg (province) (4)	coagg (city) (5)	coagg (county) (6)
highway_access	0.0773***	0.0120***	0.0028***	0.0501***	0.0079***	0.0018**
	(0.0107)	(0.0021)	(0.0009)	(0.0065)	(0.0017)	(0.0007)
knowledge_spillover	0.0006***	0.0002***	0.0001***	0.0002***	0.0000***	0.0000***
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
input_output_link	0.0253***	0.0046***	0.0014***	0.0065*	0.0006	-0.0001
	(0.0051)	(0.0012)	(0.0005)	(0.0036)	(0.0009)	(0.0005)
labour_pooling	0.0060***	0.0013***	0.0003***	0.0020	0.0003	-0.0002
	(0.0009)	(0.0002)	(0.0001)	(0.0028)	(0.0007)	(0.0003)
agriculture	0.0661***	0.0113***	0.0030***	0.0313***	0.0074***	0.0036***
	(0.0058)	(0.0008)	(0.0003)	(0.0071)	(0.0011)	(0.0004)
petroleum	0.0618***	0.0117***	0.0047***	0.1104***	0.0174***	0.0063***
	(0.0151)	(0.0028)	(0.0010)	(0.0175)	(0.0027)	(0.0010)
coal	0.3285***	0.0519***	0.0159***	0.1189***	0.0170***	0.0046***
	(0.0679)	(0.0104)	(0.0033)	(0.0205)	(0.0027)	(0.0012)
_cons	-0.0064***	-0.0012***	-0.0004***	-0.0088***	-0.0015***	-0.0004***
	(0.0005)	(0.0001)	(0.0000)	(0.0009)	(0.0002)	(0.0001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
N	127686	127686	127686	127686	127686	127686
r2_a	0.033	0.038	0.020	0.006	0.003	0.001

Note: Columns (1)-(3) are results for pooled OLS estimator and Columns (4)-(6) are for the FE estimator. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

With respect to the control variable from the pooled OLS results, the coefficients for three Marshallian mechanisms (input-output linkages, labour market pooling and knowledge spillovers) all have positive relationships with pairwise coagglomeration at three geographic levels. Regarding the FE estimator, the coefficients of knowledge spillovers are positive and significant at the 1% level. This shows knowledge spillovers are positively associated with coagglomeration. However, the coefficients for input-output linkages and labour pooling are overall not significant, and only input-output linkages have a positive and statistically significant effect

on coagglomeration at province level. The reason is highly likely to be that the construction of these two variables is not precise. These two variables only have the data for three years over the whole ten years. Moreover, the proxy for labour pooling uses US data to compensate for the lack of Chinese data.

Additionally, three proxies for natural advantages (joint agriculture input share, joint petroleum product input share and joint coal input share) also have positive and statistically significant coefficients in both pooled OLS and FE estimations. This indicates that natural advantages have positive relationships with pairwise coagglomeration.

4.5.3 Robustness Tests

Control for within-industry agglomeration

Table 4.10 presents results from both pooled OLS and fixed effects estimations, assessing the impact of highway access on industrial coagglomeration in different administrative levels (province, city, and county) while controlling for within-industry agglomeration. The findings address key questions about whether coagglomeration of industries might occur purely by chance or due to the inherent utility offered by current locations, as suggested by Helsley & Strange (2014).

The results reveal that highway access remains a significant factor influencing industrial coagglomeration even after controlling for within-industry agglomeration. Specifically, the coefficient for highway access is positive and statistically significant across all models. In the pooled OLS estimates, the coefficients are 0.0842 (province), 0.0136 (city), and 0.0035 (county), while in the FE models, they are 0.0491 (province), 0.0082 (city), and 0.0019 (county). These coefficients suggest that improved highway access increases coagglomeration of industries at various administrative levels.

Furthermore, Table 4.10 suggests that within-industry agglomeration can exhibit both positive and negative effects. Specifically, while it may contribute to increased coagglomeration, it may also potentially inhibit the spatial expansion of other industries. Given the limited space available within a region, excessive within-industry agglomeration could lead to a competitive exclusion effect, wherein the saturation of one sector restricts the entry or expansion of additional industries.

In summary, the findings affirm that highway access significantly impacts industrial coagglomeration, even when controlling for the clustering tendencies of industries within the same sector. This underscores the importance of transportation infrastructure in shaping the spatial arrangement of industries, beyond the effects of within-industry agglomeration alone.

Table 4.10: Control for within-industry agglomeration

	Pooled OLS			FE		
	(1)	(2)	(3)	(4)	(5)	(6)
	coagg(province)	coagg(city)	coagg(county)	coagg(province)	coagg(city)	coagg(county)
highway_access	0.0842***	0.0136***	0.0035***	0.0491***	0.0082***	0.0019***
	(0.0101)	(0.0020)	(0.0008)	(0.0068)	(0.0016)	(0.0007)
a_agg(province)	0.0048			-0.0093		
	(0.0064)			(0.0068)		
b_agg(province)	-0.0208***			0.0012		
	(0.0068)			(0.0057)		
a_agg(city)		0.0095***			0.0026	
		(0.0031)			(0.0031)	
b_agg(city)		-0.0163***			-0.0091***	
		(0.0025)			(0.0030)	
a_agg(county)			0.0066**			0.0016
			(0.0026)			(0.0025)
b_agg(county)			-0.0153***			-0.0097***
			(0.0027)			(0.0028)
knowledge_spillover	0.0006***	0.0002***	0.0001***	0.0002***	0.0000***	0.0000***
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
input_output_link	0.0246***	0.0043***	0.0013***	0.0064*	0.0006	-0.0000
	(0.0051)	(0.0011)	(0.0005)	(0.0036)	(0.0009)	(0.0005)
labour_pooling	0.0060***	0.0013***	0.0003***	0.0020	0.0002	-0.0002
	(0.0009)	(0.0002)	(0.0001)	(0.0028)	(0.0007)	(0.0003)
agriculture	0.0672***	0.0113***	0.0030***	0.0315***	0.0072***	0.0034***
	(0.0057)	(0.0008)	(0.0003)	(0.0071)	(0.0011)	(0.0004)
petroleum	0.0609***	0.0112***	0.0044***	0.1119***	0.0173***	0.0061***
	(0.0147)	(0.0027)	(0.0010)	(0.0176)	(0.0027)	(0.0010)
coal	0.3359***	0.0522***	0.0156***	0.1206***	0.0171***	0.0043***
	(0.0705)	(0.0105)	(0.0033)	(0.0207)	(0.0027)	(0.0012)
_cons	-0.0062***	-0.0012***	-0.0004***	-0.0082***	-0.0014***	-0.0004***
	(0.0007)	(0.0001)	(0.0001)	(0.0013)	(0.0003)	(0.0001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
N	127686	127686	127686	127686	127686	127686
r2_a	0.0350	0.0429	0.0251	0.006	0.003	0.001

Note: Columns (1)-(3) are results for pooled OLS estimator and Columns (4)-(6) are for the FE estimator. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Alternative measure for highway variable

The $\log(\text{distance}_i \times \text{distance}_j)$ has been employed as an alternative measure to ensure the robustness of the results. This variable represents the logarithmic form of the average distance between industries i and j, where a larger value indicates poorer access to highways. Thus, the estimated coefficients for $\log(\text{distance}_i \times \text{distance}_j)$ are expected to be negative. Table 4.11 presents the results for $\log(\text{distance}_i \times \text{distance}_j)$ using both the pooled OLS and fixed effects models, indicating that greater distances from highways are linked to reduced coagglomeration. The coefficients are statistically significant at the city and county levels, indicating that increased distance between firms is linked to a decrease in coagglomeration.

Table 4.11: Alternative measure of highway variable

	Pooled OLS			FE		
	coagg (province) (1)	coagg (city) (2)	coagg (county) (3)	coagg (province) (4)	coagg (city) (5)	coagg (county) (6)
logdist_ij	-0.0005	-0.0001*	-0.0001**	-0.0002	-0.0001*	-0.0001**
	(0.0008)	(0.0001)	(0.0000)	(0.0006)	(0.0001)	(0.0000)
knowledge_spillover	0.0008***	0.0002***	0.0001***	0.0003***	0.0001***	0.0000***
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
input_output_link	0.0258***	0.0047***	0.0014***	0.0068*	0.0006	-0.0001
	(0.0053)	(0.0012)	(0.0005)	(0.0037)	(0.0009)	(0.0005)
labour_pooling	0.0072***	0.0015***	0.0004***	0.0007	0.0001	-0.0002
	(0.0009)	(0.0002)	(0.0001)	(0.0028)	(0.0007)	(0.0003)
agriculture	0.0478***	0.0080***	0.0019***	0.0280***	0.0069***	0.0034***
	(0.0056)	(0.0007)	(0.0003)	(0.0068)	(0.0010)	(0.0004)
petroleum	0.0408***	0.0081***	0.0036***	0.0811***	0.0123***	0.0048***
	(0.0135)	(0.0023)	(0.0009)	(0.0162)	(0.0026)	(0.0011)
coal	0.2414***	0.0366***	0.0112***	0.0834***	0.0109***	0.0028**
	(0.0539)	(0.0077)	(0.0025)	(0.0202)	(0.0027)	(0.0013)
_cons	-0.0014	-0.0009**	-0.0006***	-0.0021	-0.0008***	-0.0005***
	(0.0026)	(0.0004)	(0.0002)	(0.0014)	(0.0003)	(0.0001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
N	127686	127686	127686	127686	127686	127686
r2_a	0.026	0.033	0.019	0.002	0.001	0.001

Note: Columns (1)-(3) are results for pooled OLS estimator and Columns (4)-(6) are for the FE estimator. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

4.6 Addressing Endogeneity

The highway network is not constructed randomly and is likely to be located where population density is higher. Some factors correlated with the highway access Variable are omitted. Additionally, with industrial agglomeration and highway networks it can be a case of reverse causality, as agglomeration may lead to the construction of highway routes. In order to address the endogenous problem, this study uses three types of instrumental variables for the highway access variable, including historical routes, least cost path minimum spanning tree network and Euclidean Straight line minimum spanning tree networks. Time-variant IVs are adopted to fit better for the highway independent variables and panel data analysis. For example, the historical routes IVs for industry i and industry j are shown as:

$$historical\ accessibility_{ij} = \frac{1}{\sqrt{distance_{i,historical\ routes} * distance_{j,historical\ routes}}}, for\ i \neq j \tag{4.8}$$

4.6.1 IV Estimation Results

The estimation results for the FE-2SLS model using the historical routes instrument, the least cost path spanning tree instrument, and the Euclidean spanning tree instrument are detailed below. All three types of instrumental variables are time-variant.

Historical Roads IV

The FE-2SLS estimation results using historical routes instruments are presented in Table 4.12, including the Ming routes instrument (Columns 1-3) and Qing routes instrument (Columns 4-6). The combination of the Ming routes and Qing routes instrument (Columns 1-3) is shown in Table 4.13. The variable *highway access* is the research interest. The first stage of FE-2SLS regresses highway access on historical routes IVs and controls. Highway access and three types of historical routes IVs have a positive and statistically significant relationship. The Kleibergen-Paap rk LM statistic and its p-value are reported, which indicates that it passes the underidentification test. The Kleibergen-Paap rk Wald F statistic is used to test weak iden-

tification, and the results in Table 4.12 indicate that the weak instrument null hypothesis is rejected. The value of the Kleibergen-Paap rk Wald F statistic is large, which indicates that the IVs are highly correlated with highway access. The historical routes IVs are significantly correlated with the highway access variable.

The second stage of FE-2SLS estimation results indicate that after using the Ming routes IV, the estimated coefficients of *highway access* obtained are statistically insignificant at three geographic levels. The results using the Qing routes IV obtain negative coefficients for highway access on coagglomeration at county level, while for coagglomeration at province and city levels, the coefficients for highway access are positive and statistically insignificant. The 2SLS estimation shows that the effects of highway access on coagglomeration are insignificant at province and city levels and have negative effects at county-level coagglomeration. Highways make the industry pairs spatially disperse at county level, possibly because highways make it possible for firms to get inputs at an acceptable cost from nearby counties.

Regarding the performance of natural advantages on coagglomeration, the estimated coefficients of agriculture, petroleum and coal resources are statistically significant and positive in general after using 2SLS. Natural advantages are proven to attract industry pairs to locate together if they share the same natural resources. In terms of three Marshallian mechanisms, the estimated coefficients of the proxy for knowledge spillovers on coagglomeration are positive and statistically significant. This indicates a higher level of knowledge spillovers leads to a higher level of coagglomeration. The input-output linkages have positive impacts on coagglomeration at province level. However, the estimated coefficients of the proxy for labour pooling are insignificant.

Table 4.12: FE-2SLS results with historical routes IV

	(1) province	(2) city	(3) county	(4) province	(5) city	(6) county
		Ming road IV			Qing road IV	
Panel A: Second stage of FE-2SLS						
Dependent variables: EG coagglomeration index at different geographic levels						
highway access	0.0096	0.0007	-0.0014	-0.0067	0.0015	-0.0052*
	(0.019)	(0.004)	(0.001)	(0.035)	(0.008)	(0.003)
knowledge_spillover	0.0003***	0.0001***	0.0000***	0.0003***	0.0001**	0.0000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
input_output_link	0.0067*	0.0006	-0.0001	0.0068*	0.0006	-0.0000
	(0.004)	(0.001)	(0.000)	(0.004)	(0.001)	(0.000)
labour_pooling	0.0010	0.0001	-0.0002	0.0005	0.0001	-0.0003
	(0.003)	(0.001)	(0.000)	(0.003)	(0.001)	(0.000)
agriculture	0.0286***	0.0069***	0.0033***	0.0275***	0.0070***	0.0031***
	(0.007)	(0.001)	(0.000)	(0.007)	(0.001)	(0.000)
petroleum	0.0862***	0.0131***	0.0044***	0.0765***	0.0136**	0.0022
	(0.020)	(0.004)	(0.001)	(0.026)	(0.006)	(0.002)
coal	0.0897***	0.0119***	0.0022	0.0780**	0.0124*	-0.0005
	(0.024)	(0.004)	(0.002)	(0.032)	(0.007)	(0.003)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,686	127,686	127,686	127,686	127,686	127,686
Number of pairs	12,880	12,880	12,880	12,880	12,880	12,880
Panel B: First stage of FE-2SLS						
Endogenous variable: highway access						
		for (1)-(3)			for (4)-(6)	
Ming road access		1.6544***				
		(0.0295)				
Qing road access					1.9179***	
					(0.0718)	
knowledge_spillover		0.0023***			0.0023***	
		(0.0001)			(0.0001)	
input_output_link		0.0085*			0.0072	
		(0.0044)			(0.0046)	
labour_pooling		-0.0291***			-0.0282***	
		(0.0038)			(0.0040)	
agriculture		-0.0701***			-0.0669***	
		(0.0093)			(0.0101)	
petroleum		-0.5405***			-0.5936***	
		(0.0502)			(0.0533)	
coal		-0.6690***			-0.7255***	
		(0.0453)			(0.0484)	
_cons		0.0871***			0.1065***	
		(0.0010)			(0.0009)	
Year_FE		Yes			Yes	
N		127686			127686	
r2_a		0.838			0.823	
Underidentification test		964.4			476.4	
p-value		0.0000			0.0000	
Weak identification test		3143			714.1	

Note: This table shows fixed effect 2SLS results with Ming routes IV, Qing routes IV and the combination of Ming routes and Qing routes IV constructed by 10km highway corridor. Year Fixed effects are included in the 2SLS estimations. The results for IV constructed by 5 and 20km highway corridors are consistent with that of the 10km highway corridor. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. Underidentification test shows Kleibergen-Paap rk LM statistic and its p-value; Kleibergen-Paap rk Wald F statistic is presented as the weak identification test. The equation is exactly identified - one IV is used for one endogenous variable.

Table 4.13: FE-2SLS results with Ming and Qing combination routes IV

	(1) province	(2) city	(3) county
	Ming&Qing road IV		
Panel A: Second stage of FE-2SLS			
Dependent variables: EG coagglomeration index at different geographic levels			
highway access	0.0077	0.0003	-0.0018
	(0.019)	(0.004)	(0.001)
knowledge_spillover	0.0003***	0.0001***	0.0000***
	(0.000)	(0.000)	(0.000)
input_output_link	0.0067*	0.0006	-0.0001
	(0.004)	(0.001)	(0.000)
labour_pooling	0.0009	0.0001	-0.0003
	(0.003)	(0.001)	(0.000)
agriculture	0.0285***	0.0069***	0.0033***
	(0.007)	(0.001)	(0.000)
petroleum	0.0851***	0.0128***	0.0042***
	(0.020)	(0.004)	(0.001)
coal	0.0883***	0.0115***	0.0020
	(0.024)	(0.004)	(0.002)
Year_FE	Yes	Yes	Yes
Observations	127,686	127,686	127,686
Number of pairs	12,880	12,880	12,880
Panel B: First stage of FE-2SLS			
Endogenous variable: highway access			
		for (1)-(3)	
combine access		1.5139***	
		(0.0287)	
knowledge_spillover		0.0023***	
		(0.0001)	
input_output_link		0.0091**	
		(0.0045)	
labour_pooling		-0.0297***	
		(0.0038)	
agriculture		-0.0706***	
		(0.0094)	
petroleum		-0.5473***	
		(0.0512)	
coal		-0.6827***	
		(0.0463)	
_cons		0.0885***	
		(0.0010)	
Year_FE		Yes	
N		127686	
r2_a		0.837	
Underidentification test		969.9	
p-value		0.0000	
Weak identification test		2782	

Note: This table shows fixed effect 2SLS results with the combination of Ming routes and Qing routes IV constructed by a 10km highway corridor. Year Fixed effects are included in the 2SLS estimations. The results for IV constructed by 5 and 20km highway corridors are consistent with that of the 10km highway corridor. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The underidentification test shows Kleibergen-Paap rk LM statistic and its p-value; Kleibergen-Paap rk Wald F statistic is presented as the Weak identification test. The equation is exactly identified - one IV is used for one endogenous variable.

LCP IV

Table 4.14 displays the FE-2SLS regression results with LCP-MST constructed by the NTHS target nodes in Columns (1)-(3), and NEN target nodes in Columns (4)-(6). The number of NEN target nodes is about three times larger than NTHS target nodes. More nodes in the LCP-MST IV may cause the IV not to be exogenous. The higher adjusted R-squares and weak identification test values for IV constructed by NEN target nodes indicate that it is a stronger IV that captures the highway access variable better than the LCP-MST IV constructed using NTHS nodes and historical IVs.

With respect to the regression results in Table 4.14, the estimated coefficients of highway access are positive and statistically significant on coagglomeration measured at province level after using LCP-MST IVs constructed with the NTHS target nodes (Column 1) and NEN target nodes (Column 4). Nevertheless, for coagglomeration measured at county levels, the coefficients of highway access are negative and insignificant with LCP IVs.

Lower transport costs may cause a lower level of coagglomeration at the county level. Firms locate where they benefit from within-industry agglomeration and are able to get intermediate inputs from other industries in other regions with a low transport cost. Intra-industry external economies are larger than inter-industry external economies (Henderson, 2003; Helsley and Strange, 2014). Particularly for the county level, a large value of the coagglomeration index means both two industries have large sizes in one county (two within-industry agglomeration industries in one county). However, as a county is smaller and with fewer resources than a city and province, and the benefits from within-industry agglomeration may be larger than coagglomeration, and thus firms choose to satisfy agglomeration first in a county. Conversely, at the city and province levels, lower transport costs can result in a higher coagglomeration value. This is because lower transport costs reduce the economic barriers to co-locating with other industries, allowing firms to benefit from broader coagglomeration externalities.

The performance of control variables is consistent with that using historical IVs. Natural advantages have positive and significant effects on coagglomeration. Knowledge spillovers are important factors that facilitate coagglomeration measured at all geographic levels.

Table 4.14: FE-2SLS results with LCP IV

	(1) province	(2) city	(3) county	(4) province	(5) city	(6) county
Panel A: Second stage of FE-2SLS Dependent variables: EG coagglomeration index at different geographic levels						
highway access	0.0326*	0.0076	-0.0023	0.0390***	0.0054**	-0.0010
	(0.017)	(0.005)	(0.002)	(0.009)	(0.002)	(0.001)
knowledge_spillover	0.0002***	0.0000**	0.0000***	0.0002***	0.0000***	0.0000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
input_output_link	0.0066*	0.0006	-0.0001	0.0065*	0.0006	-0.0001
	(0.004)	(0.001)	(0.000)	(0.004)	(0.001)	(0.000)
labour_pooling	0.0016	0.0003	-0.0003	0.0017	0.0002	-0.0002
	(0.003)	(0.001)	(0.000)	(0.003)	(0.001)	(0.000)
agriculture	0.0301***	0.0074***	0.0033***	0.0306***	0.0072***	0.0034***
	(0.007)	(0.001)	(0.000)	(0.007)	(0.001)	(0.000)
petroleum	0.0999***	0.0172***	0.0039***	0.1037***	0.0159***	0.0047***
	(0.020)	(0.004)	(0.001)	(0.018)	(0.003)	(0.001)
coal	0.1063***	0.0168***	0.0016	0.1109***	0.0152***	0.0026*
	(0.024)	(0.004)	(0.002)	(0.021)	(0.003)	(0.001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,686	127,686	127,686	127,686	127,686	127,686
Number of pairs	12,880	12,880	12,880	12,880	12,880	12,880

Panel B: First stage of FE-2SLS Endogenous variable: highway access		
	for (1)-(3)	for (4)-(6)
LCP with NTHS nodes	1.3648***	
	(0.0306)	
LCP with NEN nodes		1.5666***
		(0.0163)
knowledge_spillover	0.0021***	0.0013***
	(0.0001)	(0.0000)
input_output_link	0.0078*	-0.0019
	(0.0045)	(0.0033)
labour_pooling	-0.0283***	-0.0133***
	(0.0038)	(0.0028)
agriculture	-0.0698***	-0.0586***
	(0.0103)	(0.0068)
petroleum	-0.5466***	-0.3161***
	(0.0478)	(0.0301)
coal	-0.6623***	-0.4140***
	(0.0432)	(0.0291)
_cons	0.0837***	0.0365***
	(0.0012)	(0.0011)
Year_FE	Yes	Yes
N	127686	127686
r2_a	0.848	0.905
Underidentification test	2155	3536
p-value	0.00	0.00
Weak identification test	1990	9267

Note: This table shows fixed effect 2SLS results with LCP-MST IV constructed and the NTHS target nodes and NEN target nodes. Year Fixed effects are included in the 2SLS estimations. The results for IV constructed by 10km highway corridor are displayed. IVs constructed by 5 and 20km highway corridors generate consistent estimation results with that of the 10km highway corridor. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The underidentification test shows Kleibergen-Paap rk LM statistic and its p-value; Kleibergen-Paap rk Wald F statistic is presented as the weak identification test. The equation is exactly identified-one IV is used for one endogenous variable.

Straight-line IV

The FE-2SLS results with Euclidean-MST constructed with the NTHS target nodes and NEN target nodes are shown in Table 4.15. The weak identification test (Kleibergen-Paap rk Wald F statistic) indicates that Euclidean-MST IVs constructed with NTHS and NEN target nodes are highly correlated to highway access, and the correlation is stronger when using more target nodes. The results show that historical, LCP and Euclidean IVs are all very strong in predicting highway access.

Concerning the effects of highway access on coagglomeration, the estimated results using LCP and Euclidean IVs are consistent, which indicates that highway access significantly increases the level of coagglomeration measured at province and city levels while the effects at county level are insignificant. The estimated results of controls are also consistent with those using the other two types of IVs.

4.7 Heterogeneity across Industries

4.7.1 Bilateral Input-output adjusted Highway Access

Table 4.16 presents the results of the fixed effects two-stage least squares (FE-2SLS) estimation of the effect of Input-Output adjusted Highway Access on industrial coagglomeration across different administrative levels (province, city, and county). The primary variable of interest, bilateral IOHA, consistently shows a positive and statistically significant impact on industrial coagglomeration.

The coefficient of bilateral IOHA ranges from 0.0003 to 0.0024 across different specifications, with all estimates being highly significant at the 1% level. For example, in column (1), the coefficient of bilateral IOHA is 0.0022, indicating that a one-unit increase in the IOHA metric leads to an increase of 0.0022 units in industrial coagglomeration at the provincial level. The robustness of these results is confirmed across different instrumental variables (IVs). Specifically, the table includes results using four different IVs: LCP114nodes, Straight line 114nodes, Ming and Qing routes, and LCP time-invariant. The estimated coefficients for

Table 4.15: FE-2SLS results with Euclidean IV

	(1) province	(2) city	(3) county	(4) province	(5) city	(6) county
Panel A: Second stage of FE-2SLS Dependent variables: EG coagglomeration index at different geographic levels						
highway access	0.0312*	0.0089*	-0.0024	0.0245**	0.0045**	-0.0006
	(0.017)	(0.005)	(0.002)	(0.010)	(0.002)	(0.001)
knowledge_spillover	0.0002***	0.0000**	0.0000***	0.0002***	0.0000***	0.0000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
input_output_link	0.0066*	0.0006	-0.0001	0.0066*	0.0006	-0.0001
	(0.004)	(0.001)	(0.000)	(0.004)	(0.001)	(0.000)
labour_pooling	0.0015	0.0003	-0.0003	0.0014	0.0002	-0.0002
	(0.003)	(0.001)	(0.000)	(0.003)	(0.001)	(0.000)
agriculture	0.0300***	0.0075***	0.0033***	0.0296***	0.0072***	0.0034***
	(0.007)	(0.001)	(0.000)	(0.007)	(0.001)	(0.000)
petroleum	0.0991***	0.0180***	0.0039***	0.0951***	0.0154***	0.0049***
	(0.020)	(0.004)	(0.001)	(0.018)	(0.003)	(0.001)
coal	0.1052***	0.0177***	0.0016	0.1004***	0.0146***	0.0028*
	(0.024)	(0.004)	(0.002)	(0.021)	(0.003)	(0.001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,686	127,686	127,686	127,686	127,686	127,686
Number of pairs	12,880	12,880	12,880	12,880	12,880	12,880
Panel B: First stage of FE-2SLS Endogenous variable: highway access						
		for (1)-(3)			for (4)-(6)	
Euclidean IV with NTHS nodes		1.5403***				
		(0.0337)				
Euclidean IV with NEN nodes					1.4821***	
					(0.0150)	
knowledge_spillover		0.0020***			0.0016***	
		(0.0001)			(0.0000)	
input_output_link		0.0071			0.0046	
		(0.0045)			(0.0038)	
labour_pooling		-0.0258***			-0.0150***	
		(0.0039)			(0.0033)	
agriculture		-0.0664***			-0.0568***	
		(0.0100)			(0.0079)	
petroleum		-0.5457***			-0.4157***	
		(0.0488)			(0.0382)	
coal		-0.6689***			-0.5367***	
		(0.0443)			(0.0366)	
_cons		0.0835***			0.0486***	
		(0.0011)			(0.0010)	
Year_FE		Yes			Yes	
N		127686			127686	
r2_a		0.846			0.887	
Underidentification test		2329			3071	
p-value		0.000			0.000	
Weak identification test		2088			9804	

Note: This table shows fixed effect 2SLS results with Euclidean-MST IV constructed by NTHS target nodes and NEN target nodes. Year Fixed effects are included in the 2SLS estimations. The results for IV constructed by the 10km highway corridor are displayed. IVs constructed by 5 and 20km highway corridors generate consistent estimation results with that of the 10km highway corridor. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively. The underidentification test shows the Kleibergen-Paap rk LM statistic and its p-value; Kleibergen-Paap rk Wald F statistic is presented as the weak identification test. The equation is exactly identified-one IV is used for one endogenous variable.

bilateral IOHA remain consistent in sign and significance across all IV specifications, reinforcing the reliability of the findings. The underidentification and weak identification tests for each IV provide strong evidence for the validity of the instruments, and substantial values for the weak identification tests, ensuring that the results are not driven by weak instruments.

The results presented in the table appear to contradict the initial hypothesis that substantial decreases in transport costs, due to improved highway access, would lead to lower coagglomeration levels among industries that transport large volumes of goods between them. According to the hypothesis, a reduction in transport expenses should diminish the economic incentive for these industries to cluster geographically, as the cost savings from improved highways would allow these industries to disperse more freely.

However, the findings indicate a positive relationship between bilateral IOHA and industrial coagglomeration across all administrative levels, contrary to what was anticipated. The coefficients for bilateral IOHA are positive and significant, suggesting that improvements in highway access are associated with increased coagglomeration rather than a decrease. This result implies that rather than reducing the need for industries to cluster, better highway access may enhance their clustering tendencies.

One possible explanation for this unexpected result could be that improved highway access not only reduces transport costs but also increases the benefits of being close to other related industries. Enhanced connectivity might facilitate not only the movement of goods but also the sharing of resources, knowledge, and collabourative opportunities among co-located industries. Consequently, the reduction in transport costs may make clustering more attractive, as the benefits of proximity outweigh the cost savings from improved transport infrastructure.

In summary, the analysis reveals that improved highway access appears to reinforce, rather than diminish, industrial coagglomeration. This finding suggests that the economic advantages of clustering might be augmented by better transportation infrastructure, challenging the initial hypothesis that improved highway access would lead to reduced coagglomeration among industries with significant transport interactions.

Table 4.16: Bilateral highway access

	(1) EG_coag(province)	(2) EG_coag(city)	(3) EG_coag(county)
IV: LCP114nodes			
IOHA_bilat	0.0022***	0.0007***	0.0003***
	(0.0005)	(0.0001)	(0.0001)
Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	120,435	120,435	120,435
N_g	12,090	12,090	12,090
Underidentification test	52.92	52.92	52.92
p-value	0	0	0
Weak identification test	451.8	451.8	451.8
IV: Straight line 114nodes			
IOHA_bilat	0.0023***	0.0007***	0.0003***
	(0.0006)	(0.0001)	(0.0001)
Underidentification test	54.64	54.64	54.64
p-value	0	0	0
Weak identification test	396.4	396.4	396.4
IV: Ming and Qing routes			
IOHA_bilat	0.0022***	0.0008***	0.0003***
	(0.0005)	(0.0002)	(0.0001)
Underidentification test	50.85	50.85	50.85
p-value	0	0	0
Weak identification test	357.1	357.1	357.1
IV: LCP time-invariant			
IOHA_bilat	0.0024***	0.0008***	0.0003***
	(0.0006)	(0.0002)	(0.0001)
Underidentification test	54.37	54.37	54.37
p-value	0	0	0
Weak identification test	373.4	373.4	373.4

Note: Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

4.7.2 Related Industries and Coagglomeration

Industries within the same sector are likely to co-locate to enhance their bargaining power, benefit from knowledge spillovers, and access a shared labour pool. Dai et al. (2021) indicate that firms occupying different positions within the value chain, but utilising shared inputs, tend to be co-located in China. The research on clusters by Ruan & Zhang (2009) also in-

dicates the co-location of industries that produce the different stages of final products in a cashmere sweater cluster in Puyuan, China. This research displays the highest pairwise co-agglomeration for 3-digit industry, and finds that industries within the same 2-digit sectors are more likely to coagglomerate.

Industries within the same sectors are more likely to be similar industries and use similar inputs. Similar industries may produce substitute products, such as convenience food manufacturing and canning (both in the food manufacturing sector). It is highly possible for them to buy the same inputs and thus coagglomerate. Thus this study constructs a dummy variable that captures whether 3-digit industry pairs are in the same sector, that is, if their first 2-digit codes are the same, the dummy called 'same sector' is one, otherwise, it is zero. Due to this dummy being time-invariant, 2SLS is used to estimate the coefficients.

The effects of highway access on the coagglomeration of industry pairs whether related or not are expected to be different. To investigate the effects of highway access on the coagglomeration of pairs within or outside the same two-digit industry sector, the interaction of 'highway access' and 'same sector' is employed. The 2SLS estimator is employed with the instrument LCP IV, constructed using NEN nodes, to mitigate the endogeneity problem associated with the explanatory variable, highway access. Table 4.17 shows the second stage of 2SLS estimation. The coefficients of 'same sector' are positive and significant at 1% levels at three geographic levels. This means that industries within the same sector coagglomerate more than other industry pairs. The coefficients of the interaction term (highway*same_sector) are all positive and statistically significant at three geographic levels. This means that for related industries, the effects of highway access on their coagglomeration are larger.

Table 4.17: Related industries, highway access and coagglomeration

VARIABLES	(1) coagglomeration province	(2) coagglomeration city	(3) coagglomeration county
highway_access	0.0660***	0.0083***	0.0009
	(0.011)	(0.002)	(0.001)
highway*same_sector	0.1461***	0.0456***	0.0215***
	(0.040)	(0.010)	(0.004)
same_sector	0.0067***	0.0017***	0.0006***
	(0.001)	(0.000)	(0.000)
knowledge_spillover	0.0006***	0.0002***	0.0001***
	(0.000)	(0.000)	(0.000)
input_ouput	0.0276***	0.0040**	0.0008
	(0.008)	(0.002)	(0.001)
labour_pooling	0.0037***	0.0007***	0.0001*
	(0.001)	(0.000)	(0.000)
agriculture	0.0674***	0.0117***	0.0032***
	(0.006)	(0.001)	(0.000)
petroleum	0.0647***	0.0125***	0.0050***
	(0.016)	(0.003)	(0.001)
coal	0.3228***	0.0502***	0.0152***
	(0.067)	(0.010)	(0.003)
cons	-0.0143***	-0.0024***	-0.0007***
	(0.002)	(0.000)	(0.000)
Observations	127,686	127,686	127,686
R-squared	0.038	0.050	0.030
Underidentification test	243.7	243.8	243.9
p-value	0.000	0.000	0.000
Weak identification test	4106	4106	4106

Note: The second stage of 2SLS is presented in this table. The instrument for highway access is the LCP-MST IV constructed with NEN target nodes. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

4.7.3 State-Owned Enterprises and Coagglomeration

Lu & Tao (2009) prove that local protectionism (measured by the share of SOEs) hinders within-industry agglomeration. For coagglomeration, this study investigates whether the effects of highways are influenced when industry pairs are comprised of SOEs. This research constructs a dummy variable named ‘SOE industry’. For two industries that form an industry

pair, 'SOE industry' equals one if either one of them has an above average share of state-owned enterprises; 'SOE industry' equals zero, if neither of them has a share of SOEs higher than the average share. This study regresses the level of coagglomeration on the interaction term of highway access and 'SOE share' to investigate heterogeneity.

Table 4.18: FE-2SLS results for SOE industries

VARIABLES	(1) coagglomeration province	(2) coagglomeration city	(3) coagglomeration county
highway_access	0.0699***	0.0080***	-0.0003
	(0.008)	(0.002)	(0.001)
highway*SOE	-0.0604***	-0.0053***	-0.0015*
	(0.007)	(0.002)	(0.001)
SOE	0.0003	0.0001**	0.0001***
	(0.000)	(0.000)	(0.000)
knowledge_spillover	0.0001	0.0000***	0.0000***
	(0.000)	(0.000)	(0.000)
input_ouput	-0.0158*	-0.0035	-0.0022**
	(0.008)	(0.003)	(0.001)
labour_pooling	0.0024	0.0003	-0.0002
	(0.003)	(0.001)	(0.000)
agriculture	0.0367***	0.0078***	0.0036***
	(0.007)	(0.001)	(0.000)
petroleum	0.1314***	0.0184***	0.0052***
	(0.018)	(0.003)	(0.001)
coal	0.1409***	0.0179***	0.0033***
	(0.021)	(0.003)	(0.001)
Observations	127,686	127,686	127,686
R-squared	0.013	0.005	0.002
Underidentification test	2454	2454	2454
p-value	0.000	0.000	0.000
Weak identification test	5067	5067	5067

Note: The second stage results of FE-2SLS are presented in this table. The instrument for highway access is the LCP-MST IV constructed with NEN target nodes. Standard errors clustered at the industry pair level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

In Table 4.18, the coefficients of the interaction between SOE and highway access are negative at the 1% significance level at province and city levels, and at the 10% significance level at the county level. When 'SOE industry' equals one, the effects of highway access on coagglomeration at province, city at county levels are 0.0095 (i.e. 0.0699-0.0604), 0.0027, and

-0.0018 respectively. The reasons for the much smaller effects of highways on coagglomeration for industries with a high share of SOEs are probably that most SOEs are established before highways are constructed, and hence their locations are not affected by highways. In the dataset, the average age of SOEs over the sample period is 18 years, compared to an average age of 11 years for non-SOEs. Additionally, firms that are not state-owned enterprises tend to co-locate due to the influence of highways and are more likely to move away from locations where SOEs are situated.

4.8 Conclusion

This chapter has investigated the effects of highway access on the level of pairwise coagglomeration. The level of coagglomeration is measured by the EG pairwise coagglomeration index at three geographic levels (county, city and province). The research focus, highway access of a pair of industries, captures how far the industry pair are from highways. Highway GIS routes are used to generate the weighted distance from firms to highway networks to form the highway access variable. The weights are the share of employment of firms in the industry. The baseline results generated by the fixed effects model indicate that highway access has a positive relationship with pairwise coagglomeration at all geographic levels.

In order to examine whether there is a causal effect, this study adopts three types of time-variant instruments to address the endogenous problem for the highway access variable, including the historical routes IV, LCP-MST IV and Euclidean-MST IV. The historical routes IV consist of Ming dynasty routes, Qing Dynasty routes and their combination. LCP-MST IV and Euclidean-MST IV are formed by two categories of target nodes (nodes in NTHS and NEN plans) to align with the development of highways. The empirical results indicate that highway access has positive effects on coagglomeration at province and city levels, while their relationship is insignificant at the county level.

The control variables include natural advantages and three Marshallian mechanisms. Regarding natural advantages, agriculture, petroleum and coal resources positively affect coagglomeration. The proxies for Marshallian externalities have different performances on coagglomeration. The estimated coefficients of knowledge spillovers are positive and statistically significant on pairwise coagglomeration. The proxy for input-output linkages positively affects coagglomeration at the province level, while the results for the proxy for labour market pooling are insignificant.

This study exploits the heterogeneous features for the effects of highways on coagglomeration, which consist of bilateral input-output adjusted Highway Access, industry pairs that are related, and industries comprised of a larger amount of SOEs. The impact of bilateral IOHA on industrial coagglomeration is positive and statistically significant. For industry pairs that are related and in the same classification categories, their levels of coagglomeration rely more on highway access. Regarding industries with a larger share of SOEs, the effect of highway access on coagglomeration is smaller.

This study highlights several policy implications concerning industrial coagglomeration. Coagglomeration fosters idea sharing, labour pooling, and input sharing, thereby enhancing productivity. The findings indicate that highway access significantly promotes coagglomeration at both provincial and city levels. For developing countries aiming to boost industrial productivity through coagglomeration, investing in transportation infrastructure, such as highway networks, appears to be a viable strategy. Governments should consider supporting the expansion of highway networks and extending routes to less developed regions to facilitate regional industrial growth.

However, it is crucial to consider that the study period (1998 to 2007) reflects a time when China's highway infrastructure was still in its developmental stages, and issues related to overbuilding were less evident. During this period, the observed positive effects of highway expansion were aligned with growing demand and substantial economic contributions. As highway networks approach saturation, the marginal benefits of additional infrastructure may diminish, potentially leading to a point where further expansion may not justify the associated costs.

Second, natural advantages, such as agricultural products and mining resources, have a strong power that facilitates industrial coagglomeration. This study calls for policies that adapt to local conditions for the development of industries, especially the use of sustainable resources and keeping industries successful over the long term. Third, different features of industries change the effects of highways on coagglomeration. For instance, for industries with a larger number of SOEs, highway access has a smaller effect on coagglomeration, while for industries that are related, the effect of highway access on coagglomeration is larger. This implies that local governments can identify the different features of the industries in their administrative regions and provide policies that work best for them.

4.9 Appendices for Chapter 4

Appendix 4.A FE-2SLS Second Stage Results

Table 4.A-1: FE-2SLS results for groups with different input-output linkages

VARIABLES	(1) province EG_coag2	(2) province EG_coag2	(3) province EG_coag2	(4) city EG_coag4	(5) city EG_coag4	(6) city EG_coag4	(7) city EG_coag6	(8) city EG_coag6	(9) city EG_coag6
highway_access	0.0707***	0.0369**	-0.0257	0.0230***	0.0103**	-0.0183***	0.0033	-0.0031*	-0.0067*
	(0.025)	(0.015)	(0.024)	(0.007)	(0.005)	(0.006)	(0.003)	(0.002)	(0.004)
knowledge_spillover	0.0001	0.0002***	-0.0000	0.0000*	0.0000	-0.0000	0.0000***	0.0000*	0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
input_output_link	-0.0127***	0.0522**	-0.4153	-0.0025***	0.0121**	0.1630	-0.0013***	0.0004	-0.0360
	(0.002)	(0.024)	(0.483)	(0.001)	(0.006)	(0.103)	(0.000)	(0.003)	(0.065)
labour_pooling	-0.0039*	0.0082**	-0.0304**	0.0002	0.0015	0.0008	-0.0000	-0.0003	-0.0002
	(0.002)	(0.004)	(0.012)	(0.001)	(0.001)	(0.003)	(0.000)	(0.001)	(0.002)
agriculture	0.0575***	0.0851***	-0.0082	0.0100***	0.0181***	0.0039***	0.0030***	0.0047***	0.0030***
	(0.007)	(0.009)	(0.006)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
petroleum	0.1253***	0.0405*	0.3296***	0.0333***	0.0120**	0.0483***	0.0090***	-0.0003	0.0149***
	(0.023)	(0.021)	(0.066)	(0.005)	(0.005)	(0.011)	(0.002)	(0.002)	(0.005)
coal	0.1072***	0.0757***	0.1762***	0.0298***	0.0190***	0.0120**	0.0061***	0.0024	-0.0019
	(0.028)	(0.020)	(0.032)	(0.005)	(0.004)	(0.005)	(0.002)	(0.002)	(0.004)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	31,978	63,821	31,887	31,978	63,821	31,887	31,978	63,821	31,887
Number of industry pairs	4,697	9,177	5,042	4,697	9,177	5,042	4,697	9,177	5,042
Underidentification test	1421	3516	1201	1421	3516	1201	1421	3516	1201
Underidentification p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	1342	3409	1511	1342	3409	1511	1342	3409	1511

Note: The second stage of FE-2SLS estimation results are reported. The LCP-MST IVs constructed with NTHS target nodes are used for highway access. The clustered standard errors are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 4.A-2: 2SLS results for industry pairs in the same sector with upstream-downstream ties

Dependent variables: EG coagglomeration index at different geographic levels									
	province	province	province	city	city	city	county	county	county
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Both upstream	Both downstream	upstream&downstream	Both upstream	Both downstream	upstream&downstream	Both upstream	Both downstream	upstream&downstream
highway_access	0.0952***	0.2908***	0.4589***	0.0225***	0.1033***	0.1230***	0.0091***	0.0529***	0.0486***
	(0.026)	(0.043)	(0.058)	(0.007)	(0.011)	(0.015)	(0.003)	(0.005)	(0.006)
knowledge_spillover	0.0015***	-0.0002	0.0015*	0.0004***	0.0002***	0.0001	0.0003***	0.0001***	0.0001
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
input_output_link	0.0047	-0.0449***	-0.0097	0.0016	-0.0032*	0.0028	0.0004	-0.0009	0.0018
	(0.005)	(0.008)	(0.014)	(0.001)	(0.002)	(0.004)	(0.000)	(0.001)	(0.002)
labour_pooling	-0.0112***	0.0138**	0.0033	-0.0059***	-0.0011	-0.0035*	-0.0020***	-0.0003	-0.0002
	(0.004)	(0.006)	(0.005)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
agriculture	0.0113	0.0302***	0.1161***	0.0102***	0.0149***	0.0275***	0.0052***	0.0063***	0.0104***
	(0.010)	(0.011)	(0.015)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
petroleum	0.0308***	-3.1697	0.1799	0.0031**	-1.5349***	-0.3917*	0.0024***	-0.5232**	0.0008
	(0.006)	(2.071)	(0.746)	(0.001)	(0.579)	(0.221)	(0.001)	(0.228)	(0.067)
coal	0.0533*	0.1313	0.3732**	0.0107**	0.4513***	0.2352***	0.0042**	0.1622***	0.0784***
	(0.028)	(0.522)	(0.153)	(0.005)	(0.150)	(0.042)	(0.002)	(0.052)	(0.014)
Constant	0.0045	-0.0272***	-0.0606***	0.0030**	-0.0101***	-0.0112***	-0.0003	-0.0068***	-0.0068***
	(0.005)	(0.010)	(0.010)	(0.001)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,152	1,373	774	2,152	1,373	774	2,152	1,373	774
R-squared	0.046	0.047	0.175	0.069	0.144	0.209	0.093	0.182	0.185
Underidentification test	467.8	309.5	212	467.8	309.5	212	467.8	309.5	212
Underidentification p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	1068	1282	1028	1068	1282	1028	1068	1282	1028

Note: The second stage of 2SLS estimation results are reported. The LCP-MST IVs constructed with NTHS target nodes are used for highway access. The clustered standard errors are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Chapter 5

The Effects of Highway Access on Firm Productivity

5.1 Introduction

Economic growth in China over the last two decades has experienced a sharp increase, evidenced by the surge in firm productivity. The increase in firm productivity can be partly attributed to the upgrades of transport infrastructure. China has carried out massive highway construction since the 1990s, connecting up cities with large populations. This infrastructural revolution has boosted Chinese firms' productivity by cutting transportation time and costs, making it easier to transfer goods and people. Transport infrastructure has been investigated and there is evidence of an increase in regional output (Ozbay et al. 2007, Jiwattanakulpaisarn et al. 2012, Na et al. 2013, Yu et al. 2013, Tong et al. 2013, Faber 2014, Baum-Snow et al. 2017), and firms' labour productivity and TFP (Martín-Barroso et al. 2015, Holl 2016, Wan & Zhang 2018, Gibbons et al. 2019).

The upgrading of highways is especially beneficial for the productivity of manufacturing firms. Holl (2016) finds that highways positively and directly affect Spanish manufacturing firms' productivity. Efficient transportation facilitates the seamless movement of products between different regions. If there is a highly efficient transportation system where goods are seamlessly transported to their downstream manufacturing firms, the improved efficiency will significantly increase firm productivity.

However, in addition to the fact that highways affect productivity, it is crucial to understand how, and that question motivates this chapter. It identifies four potential channels: within-industry agglomeration, coagglomeration, export and innovation. The direct effect of highways on firm-level productivity can result from higher transportation efficiency of products. Upgrading transportation infrastructure can also increase firm productivity through several mechanisms. Agglomeration is an important channel as it may benefit firm productivity through input sharing, knowledge spillovers and labour market pooling. The improvement in transport infrastructure can facilitate agglomeration, thereby increasing firm-level productivity.

Additionally, industrial concentration and coagglomeration are both worth separately investigating as channels. These two concepts are different but when investigating the channel of agglomeration previous literature uses ambiguous proxies, mostly using density (Holl 2016, Wan & Zhang 2018). On the other hand, other mechanisms have not yet been examined. Highways can also affect export and knowledge spillovers, which in turn improve firm-level productivity. This chapter aims to analyse these novel channels to deeply investigate how highways affect firm-level productivity.

This chapter's contributions can be summarised in three aspects. First, it sheds light on the effects of highway access on Chinese firms' productivity and how it influences productivity. The mechanisms are within-industry agglomeration, cross-industry agglomeration, export, and innovation. These mechanisms reveal the underlying effects of highways on firm productivity and are new channels that have not been explored by the existing literature. Though Holl (2016), Wan & Zhang (2018) investigate the channel of agglomeration. However, their agglomeration measure is the local density or regional share of sales, which are aggregate measures and do not identify the difference between intra-industry and inter-industry agglomeration. This chapter provides new perspectives by using firm-level agglomeration and separately investigates intra-industry and inter-industry agglomeration, which mitigates the criticism of omitted variable bias.

Second, micro-level datasets are used for variable construction, which reveals new information about the variable, such as firm-level agglomeration. Firm-level agglomeration uses the location information for each firm. Additionally, GIS highway data that depict the explicit highway locations are used to capture the distance from each firm to the highway network. The NBS firm-level data and customs data are used to examine the different mechanisms and heterogeneous characteristics to make the results more convincing. These firm-level variables capture the factors that are omitted when using aggregate variables. Third, this chapter investigates the causal effect of highway access. To mitigate the potential endogeneity issue, this chapter utilizes time-variant instrumental variables for highway access, fixed-effects specifications, and accounts for factors such as firms' characteristics, region, industry, and year fixed effects.

This chapter finds that highway access positively affects firm-level TFP with a 1% change in highway access related to a 0.048% increase in TFP using the LP-ACF method. Four channels, including within-industry agglomeration, coagglomeration, export and innovation are proven to be valid. With IV estimation, the 1% increase in highway access is associated with a 0.268 to 0.462% change in within-industry agglomeration, and a 0.4% to 1.0% change in coagglomeration of different radii from 5km to 50km. The positive effects are robust after adopting new entrants as the independent variable, dropping the targeted cities in the sample, and using employment to construct the firm-level agglomeration indexes. Regarding the export channel, the IV estimation results indicate that highways stimulate exports and in turn increase firm productivity with the learning-by-exporting effects in China. The results are consistent using transport time to port through the highway network. The innovation channel is less significant. The result is significant using the new product ratio as the proxy, especially for firms with higher labour productivity and those under foreign ownership. However, the results are insignificant when using patents to capture innovation.

The structure of the chapter is as follows. Section 2 reviews both theoretical and empirical literature on productivity and transport infrastructure. Section 3 provides a description of the data, the key variable and the model specification. The empirical results with instrumental variables are presented in Section 4. Sections 5, 6, 7 and 8 investigate the mechanisms of within-industry agglomeration, coagglomeration, export and innovation, respectively. Finally, the conclusion and limitations of this study are discussed in Section 9.

5.2 Literature Review

5.2.1 Theories of Productivity

Productivity Conceptualized

De Loecker & Syverson (2021) explain the conceptualizations of productivity in three related ways. All productivity metrics measure the amount of output produced from a given set of inputs. The first conceptualization of productivity is a Hicks-neural (also called factor-neural) shifter of the production function. Regarding a general form of production function $Q = \theta F(\cdot)$. Q denotes output; $F(\cdot)$ is a function of observable inputs. θ is productivity. A higher θ means the production function isoquant shifts down and left, i.e., fewer inputs are required to produce the same amount of output.

The second interpretation of productivity is from the empirical aspect, i.e., productivity is a ratio of output with respect to inputs (De Loecker & Syverson 2021). This is related to the above conceptualization of the production function shifter, $\theta = Q/F(\cdot)$. If there is a combination of the observable inputs, productivity is called total factor productivity (also called multi-factor productivity). Additionally, when output is divided by single inputs such as labour or capital, these single-factor productivity measures are named labour productivity or capital productivity. The third conceptualization of productivity is a producer's cost curve shifter. A higher level of productivity makes production costs shift down.

Review of Micro TFP Estimation Methods

This section introduces different micro-level TFP estimation methods, including the OP, LP, ACF and the Wooldridge estimation with GMM setup. The OP method (Olley & Pakes 1992) can diminish the problem of simultaneity bias and selectivity and attrition bias with the proxy variable of investment. Instead of using investment as the proxy variable, the LP method (Levinsohn & Petrin 2003) uses intermediate inputs to estimate productivity. The Ackerberg–Caves–Frazer (ACF) correction is used to correct the first step in the OP or LP method. In the OP or the LP methods, the labour variable needs to be independent of the proxy variable;

otherwise, the coefficient of labour cannot be identified in the first stage of those methods. Ackerberg et al. (2015) change the timing of labour input choices and avoid that problem. The Wooldridge estimation (Wooldridge 2009) uses the system GMM method with IV of lags to obtain the LP estimator and avoid the problem pointed out by Ackerberg et al. (2006).

OP method. The OP method is designed to address simultaneity bias, selectivity, and attrition bias in estimating micro-level Total Factor Productivity (TFP). In a typical application using the Cobb-Douglas production function, the error term includes the logarithm of TFP. Simple linear regression methods to estimate TFP often suffer from simultaneity bias. This occurs because managers adjust inputs instantaneously based on current efficiency, leading to correlations between the error term (representing TFP) and regressors. For instance, labour inputs might be adjusted based on perceived productivity, resulting in an overestimation of labour's elasticity and an underestimation of capital's elasticity.

To overcome simultaneity bias, Olley & Pakes (1992) propose a semi-parametric estimator. Their approach assumes that firms make investment decisions based on current productivity conditions, using current investment as a proxy for unobservable productivity shocks. The capital accumulation equation implies that investment and capital decisions are made at different times, and that these decisions are orthogonal to labour inputs, which are decided contemporaneously. If a firm anticipates higher future productivity, it increases its investment. The relationship between investment and productivity is used to generate proxies for unobserved productivity. This method builds on earlier work by Marschak & Andrews (1944).

Moreover, Olley & Pakes (1992) also address selection bias, which arises when firms exit or do not survive, resulting in missing data. They use a survival probability model to estimate firm entry and exit, correcting for the potential bias due to non-random sample selection.

LP method. Instead of using investment as a proxy variable, Levinsohn & Petrin (2003) use intermediate inputs to estimate productivity. The OP method assumes a positive relationship between investment and productivity, but this is not always true. Given that intermediate inputs often have less missing data compared to investment, Levinsohn & Petrin (2003) replace investment with intermediate inputs in their production equation. This approach addresses data limitations and provides a more robust estimation when investment data is missing or unreliable.

The ACF correction. Ackerberg et al. (2006) critique both the OP and LP methods, highlighting issues with collinearity from the first stage of the estimation process. Both methods assume firms can make instantaneous adjustments to inputs in response to productivity shocks. They point out that labour demand needs to be independent of the productivity term to estimate coefficients accurately. To address this, the Ackerberg–Caves–Frazer correction method modifies the timing of input decisions, assuming labour is chosen before productivity is observed, while capital is decided at an earlier time. This adjustment leads to a more accurate specification of the production function, whether using intermediate inputs as in the LP method or investment as in the OP method.

The Wooldridge estimation. The Generalized Method of Moments (GMM), as proposed by Blundell & Bond (1998), provides a solution to simultaneity problems by incorporating instrumental variables. Wooldridge (2009) recommends using GMM instead of the two-step procedures used in OP and LP methods. GMM resolves identification issues noted by Ackerberg et al. (2015) in the first stages of OP and LP methods. Compared to these methods, GMM offers robust standard errors that account for serial correlation and heteroskedasticity, thanks to cross-equation correlation and optimal weighting matrices.

In the Wooldridge estimation framework, the production function includes output, free variables (typically labour and intermediate inputs), and state variables (typically capital). The model strengthens the assumption from OP and LP methods by including lagged values of explanatory variables, allowing for serial dependence in the error term and improving the identification of parameters. By using the GMM framework, Wooldridge's method effectively estimates parameters and resolves the issues related to labour input correlations with proxy and state variables that complicate parameter identification in earlier methods.

5.2.2 Theories about Transport Costs and Productivity

This section explores the theories about transport costs and productivity. Venables (2007) builds a model indicating that reduced transport costs can increase agglomeration, and in turn increase productivity through the mechanism of agglomeration. However, the focus of Venables (2007) is on the relationship between urban transport improvement within cities.

Redding & Turner (2015) provide a review of transport costs and economic activities. Nevertheless, they do not provide the theoretical framework that reveals the relationship between transport cost and productivity. They find higher productivity and reduced transport costs both play a role in increasing wages, city population and welfare.

Commuting Cost and Productivity

Venables (2007) develops a model examining the connection between productivity and urban transport improvements, focusing specifically on commuting costs for urban workers. Their model builds upon the urban economics framework established by Alonso (1964), as reviewed by Fujita & Thisse (2002). The central premise is that productivity tends to be higher in cities compared to rural areas due to the trade-off workers face between housing size and access to employment in city centers. Jobs are concentrated in central business districts, and Venables (2007) investigates how commuting costs impact urban workers traveling to their jobs.

In the model, productivity at a given location is influenced by the number of jobs and their proximity. Essentially, productivity increases when there are more jobs nearby and these jobs are closer to the location under consideration.

A critical equilibrium condition in the model determines city size by balancing the costs and benefits associated with urban living. This condition ensures that workers are indifferent between living in a city of a certain population or in the countryside. Commuting costs, along with other factors such as wages and taxes, are factored into this equilibrium. Specifically, the model shows that as commuting costs decrease, city size (population) tends to increase.

The relationship between commuting costs and city size is also linked to productivity. The model indicates that transportation improvements directly affect commuting costs, and this has implications for urban agglomeration and productivity. Reduced commuting expenses can lead to increased city size, which in turn enhances productivity through greater agglomeration benefits. Essentially, lower transportation costs facilitate more concentrated urban development, which boosts productivity as more workers cluster in cities, benefiting from agglomeration externalities.

The Roles of Transport Costs and Productivity in Economic Activities

Redding & Turner (2015) provide a comprehensive review of both theoretical and empirical research on the relationship between transport costs and economic activities. Their theoretical framework explores how reductions in transportation costs influence the spatial distribution of land rent, wages, population, and trade across multiple regions. While their analysis does not directly examine the relationship between productivity and transportation costs, they find that increased productivity and lower transportation costs are associated with higher wages, larger city populations, and improved welfare.

In their model, they illustrate that wages increase in response to higher firm productivity and better market access. Specifically, as transportation costs decrease, firms benefit from expanded market access, which in turn boosts their wages. Improved intra-city transport infrastructure enhances commuting capabilities, increases the effective labour supply, and further augments market access. This increased market access, coupled with greater productivity, contributes to higher wages.

Moreover, Redding & Turner (2015) find that rising productivity, reduced transport costs, and advancements in commuting technology can lead to higher land prices and population growth. They also conclude that lower transportation and commuting costs positively impact welfare by boosting labour supply and, consequently, overall welfare.

Through the Mechanism of Agglomeration and Trade

The empirical research of Holl (2016) and Faber (2014) uses agglomeration as the mechanism for the effects of transport infrastructure. Holl (2016) indicates that transport infrastructure can attract firms and economic activities and hence increase local density (agglomeration) supported by empirical evidence that transport infrastructure attracts firms and increases local employment (Holl 2004, Duranton & Turner 2012). Holl (2016) then refers to agglomeration theory by (Marshall 1890) to explain the benefits of agglomeration on firm productivity through input sharing, knowledge spillovers and labour market pooling. The improvement in transport infrastructure facilitates the transportation of goods and the movement of labour and information, and thereby increases firm-level productivity.

Faber (2014) uses trade integration (related to the concept of agglomeration) as the mechanism for the effect of highways. They cite the core-periphery model by Krugman (1991) to support the view that transport costs can increase trade integration, and increase spatial concentration in the core region. Thus, output growth is affected by the trade integration process and leads to uneven developments between core and peripheral regions.

Regarding the mechanism of trade, the gravity model can be used to reflect the effects of transport costs on trade. The basic idea is that lower transport costs lead to more trade flow between two regions. With respect to the relationship between trade and productivity, Martín-Barroso et al. (2015) estimate the effects of accessibility to labour and commodities (related to transport costs) on firm-level productivity. They use firms' international trade activities (exports and imports) as the control variables for productivity.

In summary, though there are no theories that thoroughly explore the relationship between transport costs and productivity, transport costs can affect productivity through agglomeration and trade mechanism, mainly supported by the theories by Krugman (1991) and Marshall (1890).

5.2.3 Empirical Research on Infrastructure and Productivity

Effects of Transport Infrastructure on Productivity

More research has investigated the effects of transport infrastructure on productivity using transport infrastructure data and micro-level data. Holl (2016) investigates the influence of highways on firm-level TFP. Gibbons et al. (2019) estimate the effect of the new road network in the UK on productivity. Baum-Snow et al. (2017) focus on the effects of railways and highways on Chinese cities' population. Using the GIS transport database Atack et al. (2010) research the relationship between railways and economic development in the US. Wan & Zhang (2018) investigate the effect of infrastructure on productivity in China, and find that both direct and indirect effects of infrastructure through agglomeration on productivity are positive and significant.

The macro literature uses geographic levels such as country level, state level and province level, and the transport infrastructure is usually measured by stock, density or the investment in transport infrastructure in these geographic levels. Their results for the output elasticity of transport infrastructure vary widely. This study investigates productivity on a micro-level, which will be different from the macro studies and the literature review mainly focuses on micro studies about transport infrastructure and productivity.

de Vor & de Groot (2010) investigate the relationship between employment growth of industries, agglomeration externalities and industrial sites' accessibility (highways and ports). They measure the distance from industrial sites to highway exits as highway access. They find that industrial sites near highways and ports grow fast.

Holl (2012) investigates the impact of market potential, as determined by transport costs, on firm-level total factor productivity (TFP). Using data from Spanish manufacturing firms between 1991 and 2005, and after filtering out multi-plant firms and those that changed location or industry, the study includes a final sample of 2,470 firms. Holl (2012) first estimates the firm-level TFP and then examines how market potential influences this TFP.

The market potential is calculated by incorporating transport costs, which takes into account the size of the market in other cities adjusted by the distance from each firm. Specifically, the market potential for a region is the sum of its own market size and the market sizes of other cities, weighted by their distance. The distances are based on the shortest travel time along the main road network in Spain, and the data includes the opening times of new highways.

To address the endogeneity of market potential—stemming from both population size and transport infrastructure—Holl (2012) uses instruments. They employ historical population data from 1900, adjusted by geodesic distance, and use 1760 postal routes in Spain as instruments for highways. Their findings indicate that a higher market potential significantly enhances firm-level TFP.

In a follow-up study, Holl (2016) examines the effect of highway access on firm-level TFP using annual financial data from Spanish and Portuguese firms from 1997 to 2007. This analysis focuses on single-plant firms that did not change locations. The distance from firms to their nearest highways is calculated using the Spanish road network data. The study demonstrates that improved highway access positively affects firm-level TFP, supporting the importance of transport infrastructure in enhancing productivity.

The instruments are historical routes: 1760 Spanish postal routes and Roman roads. Holl (2016) uses the time-varying historical routes following Hornung (2015) by using the historical routes that fall within a 10km buffer along highways. The control variables include industry fixed effects, year fixed effects, region fixed effects, firm characteristics, geographic characteristics, historical population and historical population growth. The results indicate that improved highway access positively affects firm-level TFP.

Holl (2016) then examines the mechanism, local density, through which highways affect productivity. Since agglomeration has positive effects on productivity and improved highway access is expected to increase agglomeration, Holl controls for the agglomeration effect to examine the direct effect of highways on firm-level productivity. Holl uses local density as the mechanism and due to it having an endogeneity issue, two instruments are adopted.

The first instrument for local density is the 1900 market potential calculated as the sum of the population of each municipality and population of other municipalities weighted by distance. The second instrument for local density is underground water, which determines human settlements.

The results of Holl (2016) show that even after controlling for agglomeration effects, the effect of highways on productivity is significant. However, the impact of the highway variable reduced slightly after controlling for local density, which implies that the highway variable in the earlier estimations has picked up a small proportion of the agglomeration effect. However, after adding the firm fixed effect, Holl finds that the effect of highways on productivity is beyond the effect of local density, which means highways improve productivity directly without the local density mechanism.

Holl (2016) also investigates firm relocation and finds that firms are attracted by highways and relocate into the vicinity of highways. In general, firms with larger size, older age and higher productivity are more like to relocate near highways but they find heterogeneity in the types of firms that are attracted to highways in city and rural areas.

Gibbons et al. (2019) investigate the impact of new roads on both employment and firm-level labour productivity. Their study encompasses two main analyses: an aggregate employment analysis and a micro-level analysis. For the aggregate analysis, they use data from approximately 10,300 electoral wards across the UK. For the micro-level analysis, they utilize establishment-level employment data from the UK Secure Data Service, covering the period from 1997 to 2008. Labour productivity is measured through per-worker values such as turnover, output, value added, and labour costs. The road network data used includes major roads from 1998 to 2007, specifically trunk and principal roads as well as highways.

To address endogeneity issues, Gibbons et al. (2019) focus on changes in road accessibility in regions near road project sites. They construct an accessibility index which is based on the inverse journey times along the roads from one region to others, adjusted by regional characteristics from a base period. This index essentially measures market access in line with trade theory. The study finds that improvements in road accessibility can significantly affect both employment and labour productivity at the firm level.

Gibbons et al. (2019) estimate the effects of road access improvements on ward employment and number of firms. They find that improved roads increase aggregate employment. For firm-level estimation of road access improvements on firm-level employment, they find that the effects of new roads are close to zero and statistically insignificant. They explain that improved roads do not affect employment in existing firms, but increase local employment through new entries and their associated employment, i.e., accessibility attracts new firms that benefit most from new roads and yield new local employment. They then regress accessibility improvements on labour productivity, and find that accessibility improvement positively affects labour productivity.

Graham (2007) examines the relationship between transport investment, agglomeration and productivity. Graham measures agglomeration of each ward in the UK with ward-level employment data. The agglomeration measure equation is similar to the approach to measuring market access in the trade literature, as Graham incorporates distance between ward pairs to the measure equation. Graham assumes that transport investment will change the agglomeration value as investment will affect the relative proximity of economic activity available to each ward. Graham then builds a model to include agglomeration effects in the production function. Graham estimates the translog production function by regressing firms' outputs on the ward-level agglomeration measure using firm-level data with accounting information and locations for UK companies from 1995 to 2002. The results show that the agglomeration effect is important and can be large, particularly for the service sector. Regarding the effects of transport cost on agglomeration economies, Graham (2007) use the previous analysis by the UK Department for Transport as evidence. The UK Department for Transport finds that a London rail scheme investment positively increases agglomeration benefits.

Martín-Barroso et al. (2015) examine how accessibility to workers and commodities affects firm-level productivity in Spain. Their study uses firm-level data from Spain spanning from 1999 to 2009. They focus on both types of accessibility because they consider them crucial factors influencing production costs. Accessibility to workers pertains to the labour supply, acknowledging that labour markets are segmented and worker mobility is restricted by spatial boundaries. In their analysis, accessibility to workers is quantified by measuring how easily a firm can attract potential employees from its region and surrounding areas, taking into account the characteristics of both the firm and the regions involved.

Additionally, Martín-Barroso et al. (2015) assess accessibility to commodities, which is determined by the ease with which firms can obtain the necessary goods. This measure considers factors such as the quantity of goods available and the similarity between the firm's needs and the commodities produced in other regions.

The impedance functions used for measuring accessibility to workers and commodities are derived from commuting and shipping data, respectively. Commuting data is categorized into various time intervals, ranging from less than 10 minutes to more than 90 minutes, using information from Google Maps for urban and intercity road networks. For commodities, the data on shipping distances are divided into several ranges.

Their baseline estimation function explores the relationship between accessibility and productivity, with firm-level TFP serving as the dependent variable. The results indicate that accessibility positively impacts firm TFP. However, accessibility to workers is found to be less influential compared to accessibility to commodities. This lower impact of labour matching is attributed to the production of medium-low technological goods being more prevalent in Spain.

Research in China

Baum-Snow et al. (2017) examine the effects of railways and highways in China on the population of cities. They use data about lights at night to identify the central business districts in China, as light at night and GDP growth are correlated. They use GIS data for highways, railways and ring roads in China and city-level economic activity data for 1990 and 2010. They regress the change in population, employment or output on transport infrastructure. They find that the activity of the service sector is decentralized by radial highways, industrial activity is decentralized by radial trains, and both are decentralized by ring roads.

Faber (2014) investigates the impact of highway construction in China on the economic activities of peripheral counties. The study utilizes county-level economic data sourced from the 1990 Chinese census dataset, as well as the 1990, 1997, and 2006 Provincial Statistical Yearbooks. Highway GIS data covering the period from 1998 to 2007 is obtained from the ACASIAN database. The focus of the study is on how the National Trunk Highway System, which aims to connect major cities, also links numerous smaller peripheral areas between these major cities.

To estimate the effects, Faber (2014) compares the economic outcomes of counties connected by highways between 1992 and 2003 with those that are not. They do this by analyzing changes in economic outcomes from 1997 to 2006 for counties that were connected to highways within a 10km corridor of the NTHS by 2003. The study employs a regression model to evaluate the impact of highway connectivity on county-level economic changes, incorporating fixed effects for provinces and control variables to account for other factors influencing economic outcomes.

The results reveal that counties connected by highways experience significant changes in economic activity compared to non-connected counties, highlighting the substantial influence of infrastructure improvements on regional economic development. Two instruments are constructed including the least cost path spanning tree network and Euclidean spanning tree network. The major difference between the two instruments is that the former uses land cover and elevation information to capture the least cost between target cities, and the latter uses straight lines to connect target cities. They find that the non-targeted peripheral counties experience a reduction in GDP growth after connecting to highways.

Faber (2014) further investigates mechanisms including the trade integration channel and local decentralization channel. The inter-regional trade-based channel emphasizes that a reduction in inter-regional trade costs between targeted metropolitan regions and non-targeted peripheral regions decreases output growth in peripheral regions. This is supported by the core-periphery model of Krugman (1991), that is, under the increasing return to scale, reduced transport cost increases agglomeration in the core region. Due to the home market effect, the development of core and peripheral regions is determined by ex-ante asymmetric market size, and spatial concentration will appear in a larger market.

The other channel is local decentralization in the connected peripheral regions, i.e., the connection to highways induces decentralization from highway-connected peripheral counties to non-connected peripheral counties near them. This is supported by Baum-Snow et al. (2017), who find that metropolises in China have experienced industrial decentralization over recent decades, with population growth at a higher rate than their urban peripheries but production growth in metropolises is at a lower rate. Faber (2014) tests these two channels and the results support the inter-regional trade channel, i.e., reduced transport cost makes the core and peripheral regions more asymmetric, with agglomeration appearing in large cities and a lower output growth rate happening in highway-connected peripheral counties.

Wan & Zhang (2018) investigate the effect of infrastructure on Chinese firms' productivity through direct and indirect ways over the period 2002-2007. They examine the agglomeration channel measured by the provincial share of sales. They use the method of Levinsohn & Petrin (2003) to estimate TFP. Their research uses data from Chinese province statistical yearbooks to construct an infrastructure consisting of roads, telecommunication servers, and cables. Their findings suggest that telecommunication and roads contribute to agglomeration. They also find that the indirect effect of infrastructure on productivity through agglomeration is small and most infrastructure effects are raised from the direct effect.

5.3 Data and Methodology

5.3.1 Data of Baseline Model

The main micro-level dataset used in this chapter is the Annual Survey of Industrial Firms, which is collected by the NBS based on the annual survey and reports submitted by firms. The firm-level data in the manufacturing industries from 1998 to 2007 is employed in this chapter. Additionally, this research adopts the industry concordances list by Brandt et al. (2012) to make the industry code consistent over the sample period. The region code in China changes over the period 1998-2007 and thus, this study also matches the region code over the sample period.

This chapter follows the method of Brandt et al. (2012) to match the firms in the ASIF from 1998 to 2007. Firm IDs are the major source of matching. Additionally, firm name, name of the legal person, phone, industry code, region code, beginning year, and main products are used to match firms over time. After the step of matching firms, an unbalanced panel is formed.

The GIS highway routes are obtained from ACASIAN Dataset. The annual GIS highway routes from 1998-2007 and the coordinates of firms are used in this chapter to calculate the distance from firms to the nearest highways. The coordinates of firms are generated using the address information in the ASIF dataset with Gaode application programming interface.

TFP measures

This chapter uses different methods to construct the TFP, including the OLS, FE, OP, OP-ACF, LP, LP-ACF, and Wooldridge estimation. This study uses the following model to estimate input coefficients:

$$lnV_{it} = \beta_0 + \beta_k lnK_{it} + \beta_l lnL_{it} + \varepsilon_{it} \tag{5.1}$$

where lnV_{it} is the logarithm of the value added of firm i at time t. Compared with the total output, the value added does not take the intermediate inputs into account, which reflects the production efficiency of firms. lnK_{it} and lnL_{it} are the logarithms of the capital and labour inputs respectively. The employment is used to construct the labour inputs. The real capital stock is used to construct capital inputs following Brandt et al. (2012). These variables are all deflated to obtain the real values.

In the OP method, investment is used as the proxy for productivity observed by firms but not by econometrics. This research constructs the investment as $I_t = K_t - K_{t-1} * (1 - depreciation\ rate)$, where the depreciation rate is 9%. The OP method assumes that productivity monotonically increases with productivity but many firms have zero investment, and there is a lot of missing data on investment in the dataset, which means the productivity of these firms cannot be identified. The LP method uses intermediate inputs as the proxy variable, with which the productivity of more firms can be estimated.

After the coefficients of capital and labour inputs are estimated ($\hat{\beta}_k$ and $\hat{\beta}_l$), the logarithm of TFP is calculated as follows:

$$lnTFP_{it} = lnV_{it} - \hat{\beta}_k lnK_{it} - \hat{\beta}_l lnL_{it} \tag{5.2}$$

5.3.2 Model Specification

The baseline model

The baseline model to investigate the effects of the distance to highways on TFP is as follows:

$$lnTFP_{it} = \alpha + \beta_1 lnhighway\ access_{it} + \beta_2 X_{it} + \delta_j + \gamma_r + \sigma_t + \varepsilon_{it} \tag{5.3}$$

where $lnTFP_{it}$ is the logarithm of TFP of firm i at time t. $lnhighway\ access_{it}$ is the negative logarithm of the distance from a firm to the nearest highway. The control variables X_{it} include firm-specific characteristics, including the age of firms, the state-ownership dummy and firm size. The industry-fixed effect δ_j, region-fixed effect γ_r and year fixed effect σ_t are also included.

Mechanism Analysis Models

This study explores the mechanism of within-industry agglomeration, coagglomeration for industry pairs, trade and innovation, through which highway improvement can affect firm productivity. A summary of the mechanism analysis methods is in the Appendix. The interaction term is usually used in the heterogeneity analysis rather than the channel analysis. For the mechanism analysis method, mediation analysis should not be adopted since it is hard to address the endogeneity problem of the mediator and explanatory variable.

Mediation analysis has been criticized because of the endogeneity problem. The prerequisite for testing the mediation effect is that it is easy to identify the causal relationship between the dependent variable, mechanisms and independent variable. However, in addition to the causal relationship between X and Y, finding the causal relationship between the mediation effect and the dependent variable is difficult. This is why mediation analysis has been rarely

seen in the empirical research literature in economics. Even without considering the endogeneity bias in testing the mediation effect, most studies that conduct such tests often find that a significant portion is a direct effect, that is the coefficient of X does not change much or becomes insignificant after adding a mechanism in the regression. This is determined by the complexity of socioeconomic phenomena, and it is something that could be expected in the first place. It would be better to introduce several mechanisms, which have theoretically straightforward causal relationships with Y. These mechanisms logically affect Y, to the extent that it is unnecessary to use formal causal inference methods.

Thus this study focuses on regressing mechanisms on X to explore the effect of X on M. Additionally, for the direct and indirect effects the mechanism is added to the baseline model as a supportive method. The mechanism analysis method is

$$Y_{it} = \alpha_0 + \beta_1 highway\ access_{it} + \tau_1 Controls_{it} + \delta_i + \gamma_r + \sigma_t + \varepsilon_{it} \tag{5.4}$$

$$M_{it} = \sigma_0 + \beta_2 highway\ access_{it} + \tau_2 Controls_{it} + \delta_i + \sigma_t + \varepsilon_{it} \tag{5.5}$$

Equation 5.4 is the baseline equation. Equation 5.5 regresses mechanisms on X. This method has been used widely, for instance, Levchenko et al. (2009) investigate the impact of financial liberalization (X) on economic growth (Y) through the mechanisms of greater entry (increased number of establishments), more employment, capital accumulation and TFP growth. They first regress economic growth (Y) on the financial liberalization dummy variable (X). They then regress these channels (M) on the financial liberalization dummy variable (X), by replacing the dependent variable of output growth and keeping the same controls.

Regarding the within-industry agglomeration and coagglomeration channels in this study, we regress these two channels on X. Additionally, we also add the mechanism to the baseline model to compare the changes in the coefficient of highway variable as a supportive method.

$$Y_{it} = \alpha_0 + \beta_3 highway\ access_{it} + \beta_4 M_{it} + \tau_3 Controls_{it} + \delta_i + \gamma_r + \sigma_t + \varepsilon_{it} \tag{5.6}$$

5.4 Summary Statistics and Baseline Results

5.4.1 Summary Statistics of Key Variables

Since the estimators of different industries vary with different characteristics, the coefficients are estimated by industries with a 2-digit classification. Table 5.1 illustrates the correlation between TFP generated by different methods. Most methods obtain highly correlated productivity with the majority around 0.8. The productivity generated by OP and OP-ACF is less correlated with that of other methods, which is because there are fewer observations of these two methods (using investment as the proxy variable) than with other methods.

Table 5.1: The correlation between TFP estimated with different methods

	OLS	FE	op	opacf	lp	lpacf	WRDG
OLS	1						
FE	0.9795	1					
op	0.6411	0.5611	1				
opacf	0.8712	0.8344	0.6656	1			
lp	0.8735	0.9156	0.4704	0.7602	1		
lpacf	0.8466	0.8309	0.494	0.8611	0.8045	1	
WRDG	0.8803	0.9292	0.4689	0.7864	0.9916	0.8034	1

Note: OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework.

The summary statistics of TFP estimated by OLS, FE, OP, OP-ACF, LP, LP-ACF, and Wooldridge estimation are shown in Table 5.2. The TFPs generated by OP and OP-ACF have much fewer observations due to the missing or negative investment data. The results of the LP and the Wooldridge GMM estimation are similar with higher productivity.

Table 5.2: The summary statistics of TFP estimated by different methods

Variable	Obs	Mean	Std. dev.	Min	Max
OLS	1,913,103	3.7292	1.0957	-8.4757	11.5499
FE	1,913,103	4.8790	1.1145	-6.5227	12.9953
op	465,534	1.2363	1.5735	-10.2183	8.9006
opacf	465,534	3.4241	1.1711	-6.3720	9.4526
lp	1,913,103	5.6733	1.2164	-6.3269	13.7359
lpacf	1,913,103	3.9842	1.2551	-8.8530	12.2402
WRDG	1,913,103	5.8329	1.2057	-5.7265	13.8841

Note: OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework.

Table 5.3: The estimated coefficient of capital and labour inputs

	OLS	FE	op	opacf	lp	lpacf	WRDG
Beta_lnK	0.2857	0.1837	0.6296	0.3622	0.2767	0.2862	0.2684
Beta_lnL	0.5206	0.4629	0.4334	0.4598	0.1301	0.4666	0.1105
N	1,913,103	1,913,103	465,534	465,534	1,913,103	1,913,103	1,913,103

Note: β_K is the estimated coefficient of ln(capital).β_L is the coefficient of ln(labour). N is the number of observations. OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework.

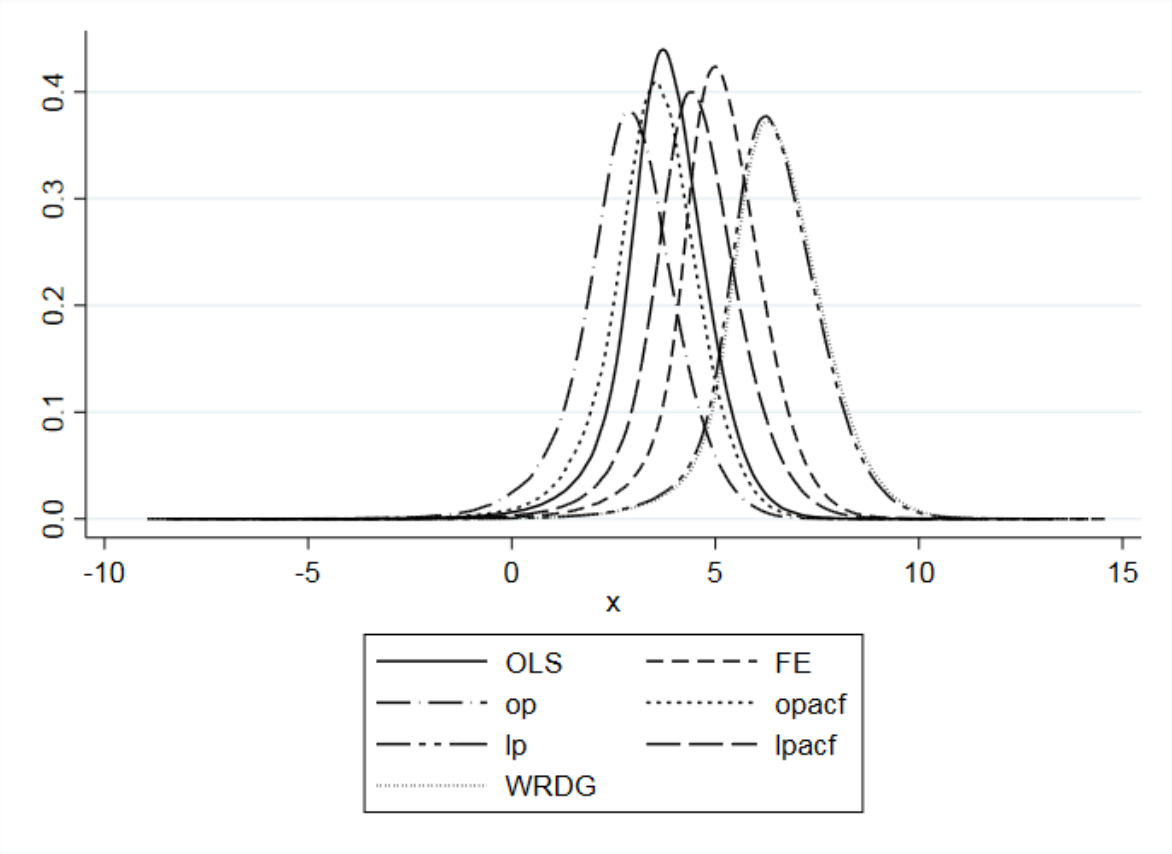


Figure 5.1: The kernel density estimations of TFP

Note: The kernel density estimations of TFP by different methods. OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework.

The correlation matrix of all variables in the baseline model is shown in Table 5.4. Besides the TFP generated by the LP and LPACF methods, there are no variables correlated. More information can be obtained from Table 5.3. The simultaneity bias means that firm managers can observe some information about productivity and adjust inputs. In this case, the error term is correlated to the input variables. If the manager observes higher productivity and uses more labour inputs, then the estimated labour inputs are overestimated. In Table 5.3,

the coefficients of capital inputs of OP-ACF, LP, LP-ACF and the Wooldridge estimation methods are lower than those of OLS and FE. The estimated coefficients of capital with the LP and Wooldridge estimation methods are similar and smaller than other methods, which explains why the TFPs generated by these two methods are higher than others. Additionally, the kernel density estimations of TFP generated by different methods are shown in Figure 5.1. The distributions of the Wooldridge estimation and LP are very similar and have the highest means compared with other methods.

Table 5.4: Correlation coefficients of all variables

	lp	lpacf	ln(highway_access)	age	state	size
lp	1					
lpacf	0.7963	1				
ln(highway_access)	0.1200	0.0785	1			
age	-0.0401	-0.1516	-0.0394	1		
state	-0.1890	-0.2654	-0.0889	0.3298	1	
size	0.2287	-0.0018	0.0268	0.1441	0.1483	1

Table 5.5 provides the summary statistics of variables used in the baseline model. The control variables include the age of firms, the state-ownership dummy and the size of firms.

Table 5.5: Summary statistics of key variables

Variable	Obs	Mean	Std. dev.	Min	Max
lp	1,911,661	5.6690	1.1614	2.1449	8.4561
lpacf	1,911,661	4.0113	1.1853	0.4557	6.9509
ln(highway_access)	1,911,720	-1.7851	1.4647	-5.1677	2.3595
age	1,834,686	1.8709	0.9604	0	3.8712
state	1,909,721	0.1291	0.3353	0	1
size	1,911,720	8.2894	1.6499	4.1287	12.6446

5.4.2 The Baseline Results

The OLS estimation results for the baseline model are shown in Table 5.6, where Columns (1) to (7) with the dependent variable ln(TFP) calculated by OLS, FE, OP, OP-ACF, LP, LP-ACF, and the Wooldridge GMM estimation. The estimated coefficients of ln(highway access) are all positive. This indicates that ln(TFP) and ln(highway access) are positively associated, which means that firms close to highways are related to higher firm TFP.

Table 5.6: OLS results

	OLS estimation,dependent variable: ln(TFP)						
	OLS	FE	OP	OPACF	LP	LPACF	WRDG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(highway_access)	0.0251***	0.0251***	0.0193***	0.0200***	0.0260***	0.0234***	0.0256***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
age	-0.0659***	-0.0597***	-0.0378***	-0.0499***	-0.0087***	-0.0544***	-0.0055***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
state	-0.6029***	-0.5928***	-0.5258***	-0.4885***	-0.5890***	-0.5961***	-0.5789***
	(0.003)	(0.003)	(0.006)	(0.005)	(0.003)	(0.003)	(0.003)
size	0.0267***	0.1474***	-0.2068***	0.0210***	0.1914***	0.0441***	0.2082***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)
cons	3.8596***	4.0011***	4.7949***	4.4450***	4.8666***	4.1811***	4.6737***
	(0.012)	(0.012)	(0.025)	(0.023)	(0.013)	(0.012)	(0.013)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1832689	1832689	454417	454417	1832689	1832689	1832689
r2_a	0.194	0.220	0.608	0.365	0.273	0.364	0.257

Note: OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

The FE estimation results are illustrated in Table 5.7. The estimated coefficients of ln(highway access) are all positive but less statistically significant overall. For the control variable, older firms are associated with more productivity. Firms that are state-owned have a negative relationship with firm productivity. The reduction in the coefficient of the key variable from OLS to Fixed Effects arises because FE controls for time-invariant unobserved factors that OLS cannot. OLS estimates include both within- and between-group variations, potentially inflating the coefficient if unobserved variables correlated with the key variable are omitted. FE, by focusing solely on within-group variation, removes the influence of these unobserved factors, resulting in a smaller and more precise estimate of the key variable's impact. In this study, the significant drop in the coefficient for the key variable suggests that while OLS might reflect both the self-selection of more productive firms near highways and the long-term effects of

initial highway placement, FE focuses on variations within firms over time. This reduction likely indicates that the initial OLS estimates captured both these effects, whereas FE isolates the impact of changes in highway access over time, filtering out the influence of past highway placements.

Table 5.7: FE results

	FE estimation,dependent variable: ln(TFP)						
	OLS	FE	OP	OPACF	LP	LPACF	WRDG
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(highway_access)	0.0022**	0.0021*	0.0027	0.0024	0.0019*	0.0024**	0.0019*
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
age	0.0131***	0.0172***	0.0195***	0.0107***	0.0386***	0.0175***	0.0401***
	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
state	-0.0780***	-0.0750***	-0.0575***	-0.0614***	-0.0487***	-0.0707***	-0.0464***
	(0.006)	(0.006)	(0.012)	(0.011)	(0.006)	(0.006)	(0.006)
size	-0.0395***	0.0367***	-0.1599***	-0.0318***	0.0298***	-0.0313***	0.0386***
	(0.001)	(0.001)	(0.004)	(0.003)	(0.001)	(0.001)	(0.001)
cons	4.5227***	5.0243***	5.2069***	5.8626***	5.9529***	4.3913***	6.1366***
	(0.137)	(0.128)	(0.218)	(0.200)	(0.135)	(0.157)	(0.159)
Year_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry_FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1832689	1832689	454417	454417	1832689	1832689	1832689
r2_a	0.070	0.083	0.235	0.139	0.101	0.118	0.094

Note: OPACF (or LPACP) means the OP method (or LP method) with ACF correction. WRDG means the Wooldridge estimation, which obtains the LP estimators with the system GMM framework. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 5.8: IV estimation results

	OLS		2SLS		FE-2SLS	
Dependent variable: ln(TFP)	LP	LPACF	LP	LPACF	LP	LPACF
	(1)	(2)	(3)	(4)	(5)	(6)
			Second stage: Dependent variable ln(TFP)			
ln(highway access)	0.0260***	0.0234***	0.0529***	0.0480***	0.0163***	0.0183***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
age	-0.0087***	-0.0544***	-0.0085***	-0.0542***	0.0388***	0.0177***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
state	-0.5890***	-0.5961***	-0.5878***	-0.5950***	-0.0487***	-0.0707***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)
size	0.1914***	0.0441***	0.1901***	0.0430***	0.0297***	-0.0314***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,832,689	1,832,689	1,832,689	1,832,689	1,669,767	1,669,767
R-squared	0.273	0.364	0.272	0.364	0.100	0.118
Underidentification test			285229	285229	33503	33503
p-value			0.000	0.000	0.000	0.000
Weak identification test			417194	417194	57225	57225
			First stage: Dependent variable ln(highway access)			
lnLCP			0.4365***	0.4365***	0.4142***	0.4142***
			(0.0012)	(0.0012)	(0.0024)	(0.0024)
age			-0.0194***	-0.0194***	-0.0010	-0.0010
			(0.0016)	(0.0016)	(0.0020)	(0.0020)
state			-0.0582***	-0.0582***	0.0028	0.0028
			(0.0050)	(0.0050)	(0.0066)	(0.0066)
size			0.0303***	0.0303***	0.0030**	0.0030**
			(0.0010)	(0.0010)	(0.0013)	(0.0013)
Year FE			Yes	Yes	Yes	Yes
Region FE			Yes	Yes	Yes	Yes
Industry FE			Yes	Yes	Yes	Yes
r2_a			0.357	0.357	0.224	0.224

Note: The instrument for ln(highway access) in 2SLS and FE-2SLS is the lnLCP IV. LPACP means the LP method with ACF correction. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 5.8 indicates the estimation results using OLS, 2SLS and FE-2SLS. Columns (1) and (2) are with OLS estimation with the dependent variable lnTFP calculated by the LP and LP-ACP methods; Columns (3) and (4) use 2SLS estimation with the ln(Least Cost Path) as the instrumental variable for ln(highway access); Columns (5) and (6) use FE-2SLS estimation with LCP IV. The estimated coefficients with LCP IVs are all statistically significant at the 1 % level. The coefficient for highway access increases with 2SLS compared to OLS because

2SLS corrects for endogeneity bias. OLS may have underestimated the effect due to omitted variables or measurement error, while 2SLS uses instrumental variables to provide a more accurate measure of the causal impact. The estimated coefficients between the LP and LPACF are similar in each estimation method, which indicates that the results consistently vary from different TFP estimation methods. In addition to the TFP estimated by the LP and LP-ACF, the TFPs estimated by other methods have consistent results. The FE-2SLS results for LPACF indicate a 1% increase in highway access yields a 0.018% increase in firm-level productivity.

5.5 Within-industry Agglomeration Channel

5.5.1 Hypothesis Development

This section investigates the firm-level agglomeration channel. The previous two chapters have shown that highway access increases within-industry agglomeration at the province, city, and county levels. Highways are expected to affect firm productivity through agglomeration, with firm-level agglomeration measures used to examine heterogeneity features, as the agglomeration levels of firms within the same industry can vary significantly.

Empirical research by Holl (2016) and Faber (2014) utilizes agglomeration as a mechanism for the effects of transport infrastructure. Faber (2014) particularly highlights trade integration (related to the concept of agglomeration) as a pathway through which highways influence productivity. This is supported by the core-periphery model of Krugman (1991), which suggests that lower transport costs can increase trade integration and spatial concentration in core regions, thereby affecting output growth and contributing to uneven development between core and peripheral areas.

A substantial body of research further supports the positive impact of agglomeration on productivity at both regional and firm levels. Studies such as Ciccone & Hall (1996) and Ciccone (2002) demonstrate that increased employment density, a common measure of agglomeration, can significantly boost labour productivity in various contexts, including the United States and several European countries. These studies generally align with the findings that agglomeration fosters productivity through mechanisms like externalities, intermediate goods variety, and labour market matching (Andersson et al. 2007).

In the context of China, Au & Henderson (2006) and Ke (2010) provide evidence that agglomeration effects enhance labour productivity, particularly in smaller cities where agglomeration economies are most pronounced. However, as city size grows, these benefits taper off, suggesting a non-linear relationship between agglomeration and productivity. Moreover, Lin et al. (2011) and Hu et al. (2015) find that industrial agglomeration can positively impact both labour productivity and total factor productivity, but these effects can vary based on factors such as industry type and regional characteristics.

Overall, these findings indicate that agglomeration generally stimulates productivity by enhancing intra-industry and inter-industry linkages, improving labour market efficiency, and fostering economies of scale and scope (Ciccone & Hall 1996, Au & Henderson 2006, Hu et al. 2015). It is therefore hypothesized that highway access positively affects the agglomeration of firms within the same industry, which in turn increases firm-level productivity.

5.5.2 Firm-level within-industry Agglomeration and Summary Statistics

Measure of firm-level within-industry agglomeration

This study uses within-industry agglomeration as the mechanism, since agglomeration has positive effects on productivity, and improved highway access is expected to increase agglomeration. Wang et al. (2022) create firm-level agglomeration with a set of radii and rings for employment for each firm. To construct the within-industry agglomeration measure, this study counts the number of firms in the same industry within 5-50km radii of each firm. The sum of employment is also used to create the measure as robustness checks. With the address

information in the NBS dataset, geocoding is applied to obtain firms' latitudes and longitudes. Then we split the sample into different 3-digit industry classification categories. Then the number of firms within the same industry in the circle with the different radii of each firm is captured using ArcGIS. This study takes the logarithm of the number of firms to create the within-industry agglomeration variable. For example, the within-industry agglomeration with a $d$km radius is:

$$agg[d]km_{ijt} = ln(N_{jt}) \tag{5.7}$$

where $agg[d]km_{ijt}$ is the logarithm of the number of firms of industry j that are located within the d km buffer of firm i at time t. $d$kms include 5, 10, 20, 30,40, and 50kms. N_{jt} represents the number of firms that are in the same industry as firm i.

Figure 5.2 illustrates an example of how the firm-level agglomeration measure is constructed. The centre of Figure 5.2 is a firm that belongs to the industry classification of grain grinding in Beijing in 2000. The points represent the location of firms that are in the same industry. The radius of the circle is 20km. The number of same-industry firms is counted to construct the measure of firm-level agglomeration. In addition to using the number of firms, as a robustness check this chapter also adopts the sum of employment in the circle to create the firm-level agglomeration measure.

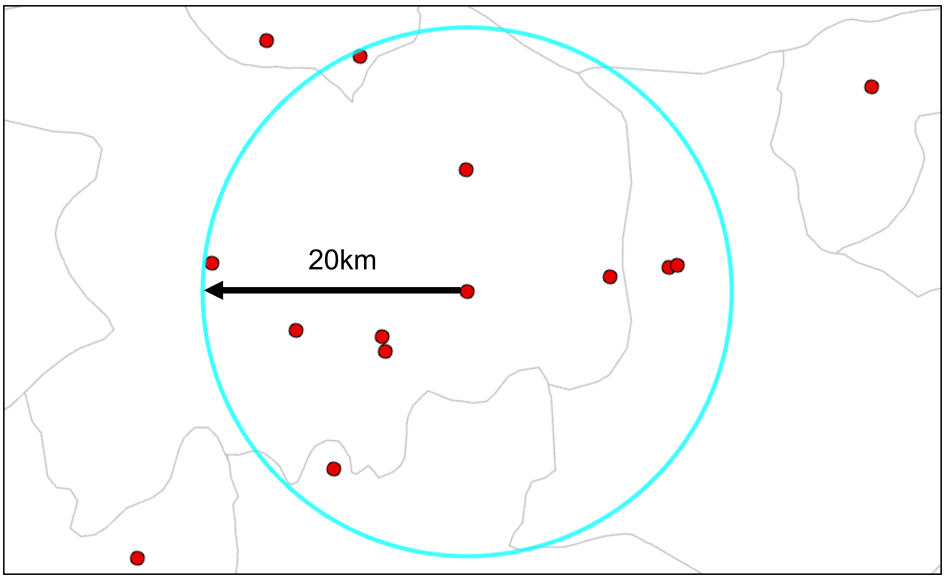


Figure 5.2: Radius of firms

Note: This is an example of firms in Beijing in 2000. The points represent the location of firms that belong to the industry classification of grain grinding. The radius of the circle is 20km. The number of firms in the radius is counted to construct the measure of firm-level agglomeration. The lines on the map are the county-level boundaries in China.

There are three possibilities for firm location: new firm, relocated firms, and the threshold in the NBS dataset. Then changes in the number of neighbouring firms are selected to construct the agglomeration variable, $\Delta agg[d]km_{ijt}$, to represent the agglomeration of new entrants. To construct this change in the neighbouring firm index, this chapter first selects firms that have data in the benchmark year. 1998 is used as the benchmark, which means that the sample size is greatly reduced. Then it only uses firms whose counties have not changed over the sample period and assumes that these firms have not moved. The change in the number of firms is the number of new entrants minus exit firms. The change in the logarithm of the number of firms in the neighbourhood of firm i between subsequent years and the benchmark year is the $\Delta agg[d]km_{ijt}$:

$$\Delta agg[d]km_{ijt_\sigma} = ln(N_{jt_\sigma}) - ln(N_{jt_0}) \tag{5.8}$$

where $\Delta agg[d]km_{ijt}$ is the change in the logarithm of the number of firms of industry j that are located within the d km buffer of firm i since t_0. t_σ is from year 1999 to 2007. t_0 in the sample is the year 1998. In this chapter, the buffers d km include 5, 10, 20, 30, 40 and 50kms. This index $\Delta agg[d]km_{ijt}$ significantly mitigates the endogeneity issue, as the change in the number of neighbouring firms is mostly likely to be affected by the highways. It reduces the reverse causality problem as highway connections can attract new firms, but the increase of neighbouring firms to firm i is not powerful enough to let the government invest in new highways around this firm, especially when most highway construction plans are pre-determined by the government.

Comparison of Firm-level and Industrial agglomeration indexes

The EG agglomeration index used in the last two chapters is the industry-level index, which uses the data of employment for each industry in given regions (province, city and county levels) in the whole country to calculate the level of agglomeration of each industry. Compared with the industrial agglomeration index, the firm-level agglomeration index has the advantage that it uses firm-level data and also avoids the Modifiable Areal Unit Problem (MAUP), that is, the province, city and county boundaries given in advance make the EG index lose location information in the process. One of the main criticisms of MAUP is that its results are inconsistent due to the use of different regional units. Thus, we examine EG

indexes at the county, city, and province levels. However, due to the boundaries between these regional units, the situation that firms may be close to each other but located in different regions cannot be investigated. The firm-level agglomeration index in this chapter uses different radii to avoid the discontinuous issue in geographical space and the MAUP.

Additionally, the firm-level agglomeration index takes into account heterogeneity between firms. The firms' location information also contains the regional effect. In the same industry, using industry-level data loses the information about firms' characteristics. For instance, two firms share the same industry-level agglomeration index, while a firm can be located in a place where there are no other firms in the neighbourhood, and another firm in the same industry can be located somewhere very spatially concentrated. Thus, it is more precise to use the firm-level agglomeration index in the firm-level estimation.

However, this firm-level agglomeration index has the disadvantage that it is highly affected by the number of firms in the same industry but less by spatial distribution. If firms are evenly located in the country but there are a lot of them, industry-level agglomeration should be small, but when we aggregate the firm-level agglomeration index to the industry-level index, this is not the case. Nevertheless, when considering the agglomeration effects on firm productivity, it is reasonable to consider that a firm's agglomeration level is determined by its neighbouring firms rather than by industry distribution over the whole country. Thus, using this firm-level agglomeration measure in this chapter is an appropriate choice.

Summary Statistics of Key Variables

This section provides the summary statistics for the firm-level within-industry agglomeration index. Table 5.9 shows the average values of $agg5km$ to $agg50km$ per year from 1998 to 2007. The value of firm-level within-industry agglomeration increases when the radius grows from 5km to 50km. This is intuitive as the larger the radii, the higher the number of firms counted within the radii. Figure 5.3 clearly illustrates the trend of firm-level within-industry agglomeration. Firm-level within-industry agglomeration for each radius overall increased from 1998 to 2007, except for 2004, which is the census year, thus the NBS data in 2004 are more explicit and contain more firms.

Table 5.9: Average values of firm-level within-industry agglomeration

year	agg5km	agg10km	agg20km	agg30km	agg40km	agg50km
1998	0.9464	1.3095	1.7698	2.1095	2.3858	2.6231
1999	0.9284	1.3044	1.7835	2.1304	2.4097	2.6466
2000	0.9504	1.3339	1.8240	2.1794	2.4639	2.7013
2001	1.0392	1.4733	2.0166	2.3917	2.6847	2.9233
2002	1.0726	1.5270	2.0898	2.4720	2.7702	3.0115
2003	1.0584	1.5434	2.1440	2.5570	2.8732	3.1290
2004	1.3464	1.9387	2.6294	3.0734	3.4015	3.6647
2005	1.2658	1.8295	2.5041	2.9457	3.2738	3.5383
2006	1.3038	1.8735	2.5560	3.0020	3.3333	3.6027
2007	1.3640	1.9465	2.6458	3.1000	3.4362	3.7086
Total	1.1822	1.6903	2.3074	2.7210	3.0353	3.2917

Note: *agg*[*d*]*km* is the logarithm of the number of same-industry firms that are located within the *d*km buffer. *d*kms include 5, 10, 20, 30,40, and 50kms.

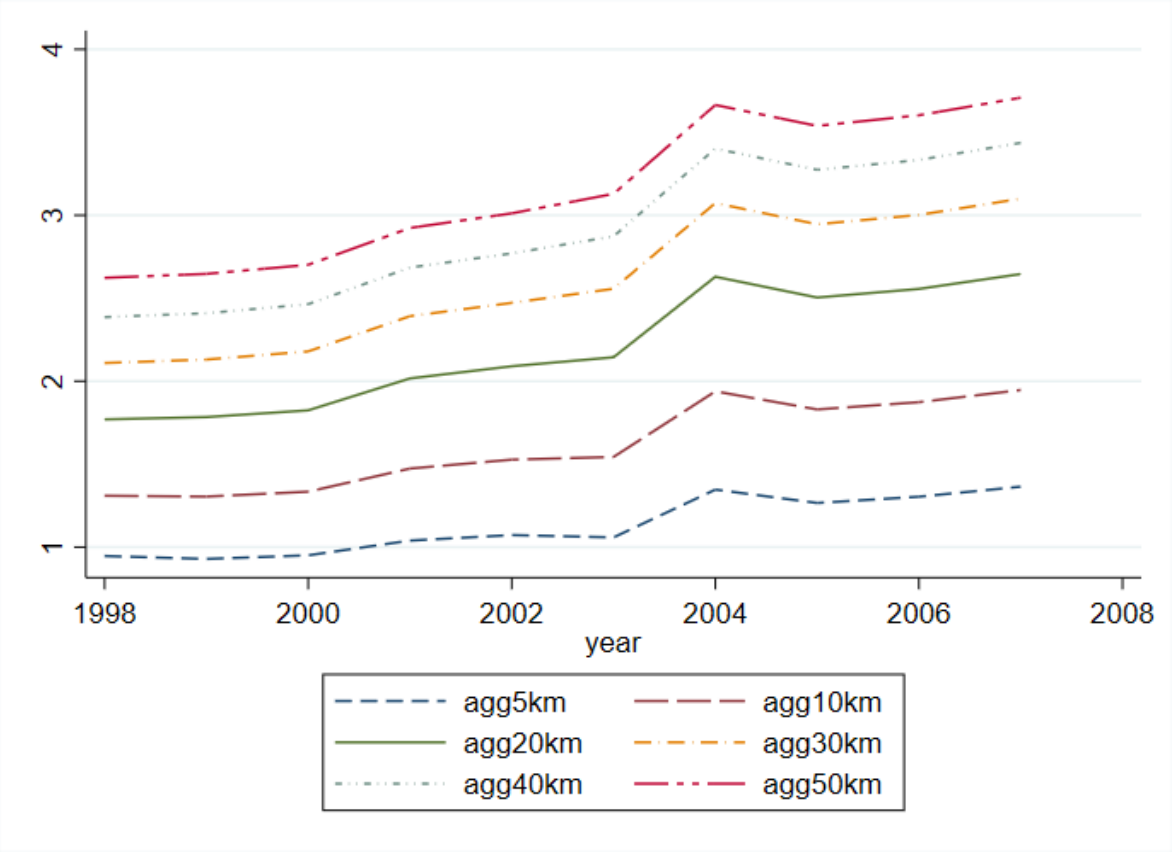


Figure 5.3: The trend of firm-level within-industry agglomeration

Note: *agg*[*d*]*km* is the logarithm of the number of same-industry firms that are located within the *d*km buffer. *d*kms include 5, 10, 20, 30,40, and 50kms. The average of *agg*[*d*]*km* per year and their trend is displayed in this figure. The value of firm-level within-industry agglomeration increases when the radius grows from 5km to 50km.

Table 5.10 provides the summary statistics of variables used in the following regressions. The following estimations use the sample that only contains firms that have not moved to other counties during the sample period. For firms that have moved to other counties, their productivity might be suddenly changed because of omitted variables that are not controlled. Thus this study excludes firms that have moved to other counties to mitigate the endogeneity problem.

Table 5.10: Summary statistics of key variables

Variable	Obs	Mean	Std. dev.	Min	Max
lpacf	1,713,138	4.0207	1.1795	0.4557	6.9509
ln(highway access)	1,713,185	-1.8129	1.4709	-5.1677	2.3595
lndistance	1,713,185	1.8129	1.4709	-2.3595	5.1677
age	1,641,349	1.8441	0.9613	0	3.8712
state	1,711,471	0.1200	0.3250	0	1
size	1,713,185	8.2416	1.6269	4.1287	12.6446
intermediate ratio	1,663,149	0.6791	0.1525	0.1899	0.9723
labour productivity	1,664,852	3.9011	1.2293	0.3701	6.7766
new product ratio	1,676,700	0.0308	0.1396	0	0.9816
agg5km	1,713,185	1.1822	1.2030	0	4.6821
agg10km	1,713,185	1.6903	1.3935	0	5.2204
agg20km	1,713,185	2.3074	1.5586	0	5.8319
agg30km	1,713,185	2.7210	1.6235	0	6.1944
agg40km	1,713,185	3.0353	1.6465	0	6.4409
agg50km	1,713,185	3.2917	1.6527	0	6.7007
$\Delta agg5km$	217,945	-0.0058	0.4190	-2.0149	1.7918
$\Delta agg10km$	217,945	-0.0010	0.4434	-2.0794	1.9459
$\Delta agg20km$	217,945	0.0077	0.4257	-1.8718	1.9459
$\Delta agg30km$	217,945	0.0112	0.4008	-1.7047	1.8718
$\Delta agg40km$	217,945	0.0143	0.3807	-1.6094	1.7918
$\Delta agg50km$	217,945	0.0173	0.3601	-1.4351	1.7358

5.5.3 Results for within-industry Agglomeration Channel

This section presents the regression results using agg_{ijt} as the dependent variable. Table 5.11 shows the FE estimations of the relationship between highway access and within-industry agglomeration of firms in radii of 5km, 10km, 20km, 30km, 40km and 50km in Columns 1 to 6 respectively. The estimated coefficients of ln(highway access) are all positive and statistically significant. This indicates that highway access is positively associated with within-industry

agglomeration of firms. Regarding the control variables, intermediate ratio, labour productivity and new product ratio are proxies for input sharing, labour market pooling and knowledge spillovers. The results also imply that these three Marshallian externalities are all positively associated with within-industry agglomeration of firms.

Table 5.11: The effects of highway access on within-industry agglomeration

	agg5km (1)	agg10km (2)	agg20km (3)	agg30km (4)	agg40km (5)	agg50km (6)
ln(highway access)	0.0748***	0.0962***	0.1069***	0.0994***	0.0888***	0.0783***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
intermediate ratio	0.0151**	0.0249***	0.0289***	0.0288***	0.0294***	0.0269***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
labour productivity	0.0023**	0.0037***	0.0048***	0.0049***	0.0050***	0.0043***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
new product ratio	0.0447***	0.0499***	0.0473***	0.0394***	0.0333***	0.0312***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)
age	0.0359***	0.0361***	0.0325***	0.0307***	0.0288***	0.0281***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
state	0.0541***	0.0660***	0.0772***	0.0738***	0.0707***	0.0709***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
size	0.0211***	0.0250***	0.0262***	0.0252***	0.0245***	0.0227***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
N	1559713	1559713	1559713	1559713	1559713	1559713
$r2_a$	0.069	0.121	0.191	0.243	0.285	0.323

Note: The dependent variable is within-industry agglomeration with different radii from 5km to 50km. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

In order to evaluate the causal relationship between highways and within-industry agglomeration, this study uses the LCP IV with NEN target nodes for highway access. Table 5.12 displays the second stage of FE-2SLS results with the dependent variables of within-industry agglomeration with different radii. The results support the hypothesis that highways increase within-industry agglomeration of firms. For a radius of 10km, the 1% change in highway access is associated with a 0.4616% change in within-industry agglomeration. The elasticity of within-industry agglomeration of a 10km radius with respect to highway access is 0.4616.

Table 5.12: The effects on within-industry agglomeration with IV

VARIABLES	(1) agg5km	(2) agg10km	(3) agg20km	(4) agg30km	(5) agg40km	(6) agg50km
Second stage of FE-2SLS						
Dependent variable: within-industry agglomeration with different radius						
ln(highway access)	0.4242***	0.4616***	0.4258***	0.3723***	0.3180***	0.2683***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
intermediate ratio	0.0191***	0.0292***	0.0327***	0.0320***	0.0320***	0.0291***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
labour productivity	0.0025**	0.0039***	0.0050***	0.0050***	0.0051***	0.0044***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
new product ratio	0.0437***	0.0489***	0.0464***	0.0387***	0.0326***	0.0307***
	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)
age	0.0391***	0.0395***	0.0354***	0.0332***	0.0309***	0.0298***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
state	0.0534***	0.0652***	0.0765***	0.0733***	0.0702***	0.0705***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
size	0.0190***	0.0228***	0.0243***	0.0236***	0.0232***	0.0215***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562
Number of panel id	362,152	362,152	362,152	362,152	362,152	362,152
First stage of FE-2SLS						
Dependent variable: ln(highway access)						
LCP NEN IV	0.4374***	0.4374***	0.4374***	0.4374***	0.4374***	0.4374***
	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)
intermediate ratio	-0.0215**	-0.0215**	-0.0215**	-0.0215**	-0.0215**	-0.0215**
	(0.0084)	(0.0084)	(0.0084)	(0.0084)	(0.0084)	(0.0084)
labour productivity	-0.0037***	-0.0037***	-0.0037***	-0.0037***	-0.0037***	-0.0037***
	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
new product ratio	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030
	(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0076)
age	-0.0004	-0.0004	-0.0004	-0.0004	-0.0004	-0.0004
	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0021)
state	0.0058	0.0058	0.0058	0.0058	0.0058	0.0058
	(0.0074)	(0.0074)	(0.0074)	(0.0074)	(0.0074)	(0.0074)
size	0.0026*	0.0026*	0.0026*	0.0026*	0.0026*	0.0026*
	(0.0014)	(0.0014)	(0.0014)	(0.0014)	(0.0014)	(0.0014)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
N	1559713	1559713	1559713	1559713	1559713	1559713
r2_a	0.238	0.238	0.238	0.238	0.238	0.238
Underidentification test	13480	13480	13480	13480	13480	13480
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	26401	26401	26401	26401	26401	26401

Note: The dependent variable is within-industry agglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV with NEN nodes. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 5.13 provides FE-2SLS estimation results without and with within-industry agglomeration variables respectively. The coefficients of the highway variable and the within-industry agglomeration variables are all positive. Additionally, the coefficient of ln(highway access) variable reduces by about 1/3, which means that within-industry agglomeration captures the effect of highway access on TFP. Within-industry agglomeration is an important channel for the effect of highway access on TFP.

Table 5.13: The effects on highway access on lnTFP

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Second stage of FE-2SLS Dependent variable:lnTFP (LPacf)							
ln(highway access)	0.0189***	0.0125***	0.0108***	0.0101***	0.0105***	0.0112***	0.0123***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
agg5km		0.0152***					
		(0.001)					
agg10km			0.0178***				
			(0.001)				
agg20km				0.0209***			
				(0.001)			
agg30km					0.0229***		
					(0.001)		
agg40km						0.0243***	
						(0.002)	
agg50km							0.0248***
							(0.002)
age	0.0206***	0.0200***	0.0199***	0.0199***	0.0198***	0.0198***	0.0199***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
state	-0.0756***	-0.0764***	-0.0768***	-0.0772***	-0.0773***	-0.0773***	-0.0773***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
size	-0.0322***	-0.0325***	-0.0326***	-0.0327***	-0.0327***	-0.0327***	-0.0327***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES
Observations	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395
Number of panel id	377,651	377,651	377,651	377,651	377,651	377,651	377,651
Underidentification test	29779	29579	29382	29420	29739	30007	30186
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	51985	50621	49584	48799	49243	49850	50403

Note: This table reports the second stage of the FE-2SLS estimator. The dependent variable is lnTFP, measured by the LP acf-correction method. The instrument for highway access in FE-2SLS is the LCP IV constructed with the NEN target nodes. Columns 2 to 7 controls for within-industry agglomeration with different radii from 5km to 50km.

5.5.4 New Entrants

This section uses the change in the number of firms since 1998 to construct the dependent variable. Highways can attract new entrants, which leads to agglomeration. The change in the number of firms from 1998 is used to capture new entrants and relocated firms.

Table 5.14 shows the FE estimations of the relationship between highway access and new entrants. The estimated coefficients of ln(highway access) are all positive and statistically significant. Owing to only firms that have data from 1998 being used to construct the change in agglomeration variable, the sample size is much reduced. The coefficients of the control variables, intermediate ratio, labour productivity and new product ratios are inconsistent with using all firms in the last section.

Table 5.14: The effects of highway access on new entrants

	$\Delta agg5km$ (1)	$\Delta agg10km$ (2)	$\Delta agg20km$ (3)	$\Delta agg30km$ (4)	$\Delta agg40km$ (5)	$\Delta agg50km$ (6)
ln(highway access)	0.1145***	0.1458***	0.1630***	0.1496***	0.1363***	0.1204***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
intermediate ratio	-0.0231	-0.0445**	-0.0298*	-0.0340**	-0.0368**	-0.0413***
	(0.017)	(0.017)	(0.016)	(0.015)	(0.015)	(0.014)
labour productivity	-0.0033	-0.0050*	-0.0046*	-0.0066***	-0.0080***	-0.0092***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
new product ratio	-0.0218	-0.0267	-0.0111	0.0033	0.0024	0.0022
	(0.018)	(0.019)	(0.018)	(0.017)	(0.016)	(0.015)
age	0.0127**	0.0179***	0.0239***	0.0227***	0.0218***	0.0207***
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
state	0.0091	0.0134	0.0122	0.0007	0.0026	-0.0023
	(0.010)	(0.011)	(0.011)	(0.010)	(0.010)	(0.009)
size	0.0021	0.0030	0.0017	0.0040	0.0042	0.0045
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
N	206474	206474	206474	206474	206474	206474
r2_a	.05526	.08461	.1234	.1458	.1645	.1791

Note: The dependent variable is Δwithin-industry agglomeration with different radii from 5km to 50km. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

In order to evaluate the causal relationship between highways and new entrants, this study uses the LCP IV with NEN target nodes for highway access. Table 5.15 displays the second stage of FE-2SLS results with the dependent variables of change in within-industry agglomeration with different radii. The positive and significant coefficients of ln(highway access) indicate that highways attract new firms or increase the number of relocated firms, and lead to agglomeration. For a radius of 10km, the 1% change in highway access is associated with a 0.4210% change in new entrants. The elasticity of new entrants in a 10km radius with respect to highway access is 0.4210.

Table 5.15: The effects on new entrants with IV

	$\Delta agg5km$ (1)	$\Delta agg10km$ (2)	$\Delta agg20km$ (3)	$\Delta agg30km$ (4)	$\Delta agg40km$ (5)	$\Delta agg50km$ (6)
Second stage of FE-2SLS						
Dependent variable: change in within-industry agglomeration						
ln(highway access)	0.3745***	0.4210***	0.3835***	0.3290***	0.2796***	0.2295***
	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)
intermediate ratio	-0.0286*	-0.0503***	-0.0345**	-0.0379**	-0.0398***	-0.0436***
	(0.017)	(0.018)	(0.017)	(0.015)	(0.015)	(0.014)
labour productivity	-0.0035	-0.0052*	-0.0048*	-0.0068***	-0.0081***	-0.0093***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
new product ratio	-0.0224	-0.0273	-0.0116	0.0028	0.0021	0.0019
	(0.019)	(0.019)	(0.018)	(0.017)	(0.015)	(0.014)
age	0.0160***	0.0214***	0.0267***	0.0250***	0.0236***	0.0221***
	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)
state	0.0182*	0.0230**	0.0200*	0.0070	0.0077	0.0016
	(0.011)	(0.011)	(0.011)	(0.010)	(0.009)	(0.009)
size	-0.0003	0.0005	-0.0004	0.0024	0.0029	0.0035
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	166,927	166,927	166,927	166,927	166,927	166,927
Number of panel id	68,518	68,518	68,518	68,518	68,518	68,518
Underidentification test	3781	3781	3781	3781	3781	3781
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	8292	8292	8292	8292	8292	8292

Note: The dependent variable is Δwithin-industry agglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV. Robust standard errors are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

5.5.5 Robustness Checks: Agglomeration Measure using Employment

In addition to using the number of firms that co-located to construct the within-industry agglomeration measure, the sum of employment is also used for robustness checks.

Table 5.16 displays the second stage of FE-2SLS results with the dependent variables of within-industry agglomeration calculated with employment. The results consistently the hypothesis support that highways increase within-industry agglomeration of firms. The estimated coefficient of ln(highway access) is slightly larger than when using the number of firms to calculate the agglomeration measure. For a radius of 10km, the 1% change in highway access is associated with a 0.5704% change in within-industry agglomeration.

Table 5.16: The effects on within-industry agglomeration (employment)

	agg5km (1)	agg10km (2)	agg20km (3)	agg30km (4)	agg40km (5)	agg50km (6)
Second stage of FE-2SLS						
Dependent variable: within-industry agglomeration (employment)						
ln(highway access)	0.5182***	0.5704***	0.5275***	0.4637***	0.3965***	0.3362***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
intermediate ratio	-0.5258***	-0.3998***	-0.2906***	-0.2303***	-0.1862***	-0.1592***
	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)	(0.007)
labour productivity	-0.1520***	-0.1189***	-0.0879***	-0.0708***	-0.0601***	-0.0530***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
new product ratio	0.0764***	0.0718***	0.0608***	0.0505***	0.0435***	0.0448***
	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
age	0.0814***	0.0752***	0.0612***	0.0554***	0.0516***	0.0478***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
state	0.0914***	0.1055***	0.1169***	0.1154***	0.1073***	0.1025***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
size	0.1154***	0.0983***	0.0785***	0.0658***	0.0573***	0.0490***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562
Number of panel id	362,152	362,152	362,152	362,152	362,152	362,152
First stage of FE-2SLS						
Dependent variable: ln(highway access)						
LCP NEN IV	0.4374***	0.4374***	0.4374***	0.4374***	0.4374***	0.4374***
	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)	(0.0027)
Control variables	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
N	1559713	1559713	1559713	1559713	1559713	1559713
r2_a	0.238	0.238	0.238	0.238	0.238	0.238
Underidentification test	13480	13480	13480	13480	13480	13480
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	26401	26401	26401	26401	26401	26401

Note: The dependent variable is within-industry agglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 5.17 displays the second stage of FE-2SLS results using employment to measure new entrants. The results are consistent compared with using the number of new entrants to measure agglomeration and with slightly larger estimated coefficients of ln(highway access). The elasticity of new entrants calculated by employment in a 10km radius with respect to highway access is 0.5319.

Table 5.17: The effects on new entrants (employment) with IV

	$\Delta agg5km$ (1)	$\Delta agg10km$ (2)	$\Delta agg20km$ (3)	$\Delta agg30km$ (4)	$\Delta agg40km$ (5)	$\Delta agg50km$ (6)
Second stage of FE-2SLS						
Dependent variable: change in within-industry agglomeration (employment)						
ln(highway access)	0.4898***	0.5319***	0.4896***	0.4199***	0.3540***	0.2996***
	(0.011)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)
intermediate ratio	-0.8029***	-0.7338***	-0.5926***	-0.5114***	-0.4521***	-0.4159***
	(0.028)	(0.028)	(0.025)	(0.024)	(0.022)	(0.021)
labour productivity	-0.2165***	-0.1887***	-0.1533***	-0.1321***	-0.1182***	-0.1069***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
new product ratio	0.0418	0.0267	0.0210	0.0309	0.0346*	0.0339*
	(0.027)	(0.027)	(0.025)	(0.023)	(0.021)	(0.019)
age	0.0552***	0.0536***	0.0589***	0.0544***	0.0484***	0.0465***
	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)
state	0.0235	0.0326**	0.0250*	0.0127	0.0186	0.0088
	(0.016)	(0.016)	(0.015)	(0.014)	(0.012)	(0.012)
size	0.0970***	0.0797***	0.0593***	0.0513***	0.0451***	0.0387***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	166,927	166,927	166,927	166,927	166,927	166,927
Number of panel id	68,518	68,518	68,518	68,518	68,518	68,518
Underidentification test	3781	3781	3781	3781	3781	3781
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	8292	8292	8292	8292	8292	8292

Note: This table reports the second stage of the FE-2SLS estimator. The dependent variable is Δwithin-industry agglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes.

5.5.6 Robustness Checks: without Targeted Cities

For firms in the targeted cities in the NTHS plan, the highways are connected to these cities. These targeted cities are provincial capitals and big cities. In addition to these cities, others are more likely to be randomly connected to highways. Thus, this section excludes firms in targeted cities to investigate the relationship between highways and firm-level productivity through the channel of within-industry agglomeration.

Table 5.18 shows the second stage of FE-2SLS results for firms that are not in the targeted cities in the NTHS plan. The estimated coefficients of the highway variable are all positive and statistically significant, which further supports the hypothesis that highways increase within-industry agglomeration of firms even in less developed regions. For the agglomeration of same-industry firms in a radius of 10km, the 1% change in highway access gives a 0.3466% change in within-industry agglomeration in non-targeted cities. The highway access elasticity for firms in non-targeted cities is smaller than that of the whole sample in Table 5.12.

Table 5.18: The effects on within-industry agglomeration without targeted cities

VARIABLES	(1) agg5km	(2) agg10km	(3) agg20km	(4) agg30km	(5) agg40km	(6) agg50km
Second stage of FE-2SLS						
Dependent variable: within-industry agglomeration with different radius						
ln(highway access)	0.3082***	0.3466***	0.3413***	0.3196***	0.2960***	0.2677***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
intermediate ratio	-0.0197*	-0.0170	-0.0218*	-0.0181	-0.0127	-0.0065
	(0.011)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
labour productivity	0.0006	0.0018	0.0032*	0.0035**	0.0043**	0.0045***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
new product ratio	0.0442***	0.0630***	0.0499***	0.0376***	0.0306***	0.0294***
	(0.011)	(0.012)	(0.012)	(0.011)	(0.011)	(0.010)
age	0.0278***	0.0308***	0.0333***	0.0332***	0.0333***	0.0332***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
state	0.0810***	0.0979***	0.1107***	0.1078***	0.1020***	0.1019***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
size	0.0223***	0.0256***	0.0266***	0.0265***	0.0261***	0.0241***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	478,631	478,631	478,631	478,631	478,631	478,631
Number of panel id	126,140	126,140	126,140	126,140	126,140	126,140
Underidentification test	14453	14453	14453	14453	14453	14453
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	35492	35492	35492	35492	35492	35492

Note: This table presents the second stage of the FE-2SLS estimator. The dependent variable is within-industry agglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV.

This section then further discusses how highways affect firm productivity using the results in Table 5.19. The coefficient of ln(highway access) reduces by around 1/3 to 1/2 after adding the within-industry agglomeration variable. This indicates that after excluding the firms in targeted cities, the results still support the view that highways increase firm productivity through the channel of within-industry agglomeration.

Table 5.19: The effects on highway access on lnTFP without targeted cities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Second stage of FE-2SLS Dependent variable:lnTFP (LPacf)							
ln(highway access)	0.0149***	0.0097***	0.0084**	0.0078**	0.0076**	0.0078**	0.0079**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
agg5km		0.0171***					
		(0.002)					
agg10km			0.0191***				
			(0.002)				
agg20km				0.0212***			
				(0.002)			
agg30km					0.0230***		
					(0.002)		
agg40km						0.0242***	
						(0.002)	
agg50km							0.0266***
							(0.003)
age	0.0173***	0.0168***	0.0167***	0.0166***	0.0165***	0.0165***	0.0164***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
state	-0.0914***	-0.0927***	-0.0932***	-0.0937***	-0.0938***	-0.0938***	-0.0940***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
size	-0.0352***	-0.0355***	-0.0356***	-0.0357***	-0.0358***	-0.0358***	-0.0358***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Year FE	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES
Observations	499,044	499,044	499,044	499,044	499,044	499,044	499,044
Number of panel id	130,698	130,698	130,698	130,698	130,698	130,698	130,698
Underidentification test	14926	15087	15103	15188	15272	15298	15316
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	36380	34763	34364	34315	34505	34689	34972

Note: This table presents the second stage of the FE-2SLS estimator. The dependent variable is lnTFP, measured by the LP acf-correction method. The instrument for highway access in FE-2SLS is the LCP IV. Columns 2 to 7 control for the within-industry agglomeration at different radii.

5.6 Coagglomeration Channel

5.6.1 Hypothesis Development

Highways can affect firm total factor productivity through multiple mechanisms, including both within-industry agglomeration and coagglomeration. While within-industry agglomeration refers to the spatial concentration of firms within the same industry, coagglomeration pertains to the spatial concentration of different industries within the same area. This section explores how coagglomeration enhances productivity at the firm level.

Empirical studies provide evidence of the positive effects of coagglomeration on productivity. For instance, Barrios et al. (2006) examine plant-level data from Ireland and employ the coagglomeration measure derived from Ellison & Glaeser (1997). Their study finds significant coagglomeration of domestic and foreign enterprises, with spillover effects from multinational plants boosting employment and productivity for domestic firms, but only in industries where these plants are co-located.

Similarly, Glaeser et al. (1992) investigate urbanisation and specialisation in US cities and find that urbanisation economies, rather than specialisation, drive employment growth, suggesting that cross-industry knowledge spillovers contribute to this growth. Supporting this view, Feldman & Audretsch (1999) highlight that diversification economies significantly impact innovation in US cities, indicating the importance of a varied industrial environment. Moreover, Duranton & Puga (2001) develop a dynamic model demonstrating that firms tend to relocate to diversified cities, further emphasising the value of coagglomeration.

In the context of Japan, Nakamura (1985) finds that light industries benefit more from urbanisation economies, while heavy industries favour localisation economies, highlighting that the impact of coagglomeration may vary depending on the industry type. In China, He & Pan (2010) analyse data on manufacturing industries and show that specialisation economies initially promote city growth but may eventually impede it, whereas diversification economies have a consistently positive effect on growth after a certain threshold. This nonlinear relationship between agglomeration economies and productivity is also observed by Au & Henderson (2006).

Overall, these findings suggest that coagglomeration can stimulate firm-level productivity by facilitating knowledge spillovers and economic diversification. Therefore, it is hypothesised that highway access enhances the coagglomeration of firms across different industries, which in turn improves firm-level productivity.

5.6.2 Firm-level Coagglomeration and Summary Statistics

Measure of firm-level coagglomeration

The method of constructing the firm-level coagglomeration variable is similar to that of the firm-level within-industry agglomeration variable. This study uses the number of firms of all other industries in the circle centred on each firm to construct the firm-level coagglomeration variable. This study takes the logarithm of the number of firms in all other industries. For example, the coagglomeration with $d$km radius is:

$$coagg[d]km_{ijt} = ln(N_{kt}), k \notin j \tag{5.9}$$

where $coagg[d]km_{ijt}$ is the logarithm of the coagglomeration of firm i of industry k ($k \notin j$) at time t. The buffers $d$km include 5, 10, 20, 30, 40 and 50kms. The coagglomeration of firm i, $ln(N_{kt})$, is the number of firms in industries $\notin j$ in a circle with a given radius of firm i.

To capture new entrants for coagglomeration, the variable, $\Delta coagg[d]km_{ijt_\sigma}$, is constructed. 1998 is used as the benchmark and firms whose counties have not changed over the sample period are selected. $\Delta coagg[d]km_{ijt_\sigma}$, actually captures the change in the number of firms that is the number of new entrants minus exit firms. The change in the logarithm of the number of firms of different industries in the neighbourhood of firm i between subsequent years and the benchmark year is the $\Delta agg[d]km_{ijt}$:

$$\Delta coagg[d]km_{ijt_\sigma} = ln(N_{kt_\sigma}) - ln(N_{kt_0}), k \notin j \tag{5.10}$$

where $\Delta coagg[d]km_{ijt}$ is the change in the logarithm of the number of firms of industry k ($k \notin j$) that are located within the $d$km buffer of firm i since t_0. t_σ is from 1999 to 2007. t_0 in the sample is 1998. The buffers $d$km include 5, 10, 20, 30, 40 and 50kms.

Summary Statistics of Key Variables

This section provides the summary statistics for the firm-level coagglomeration index. Table 5.20 displays the average values of $coagg[d]km$ per year from 1998 to 2007. The value of firm-level coagglomeration increases when the radius grows from 5km to 50km. Compared with within-industry agglomeration, the value of the coagglomeration index is larger, which is because the firms in all other industries are counted to construct the firm-level coagglomeration index. Figure 5.4 illustrates the trend of firm-level industry coagglomeration. Owing to 2004 being the census year, there is a bump in the coagglomeration index as well. The overall trend is still upward, which means that more firms are located in the neighbourhood.

Table 5.20: Average values of firm-level coagglomeration

year	coagg5km	coagg10km	coagg20km	coagg30km	coagg40km	coagg50km
1998	3.7873	4.5751	5.4207	5.9587	6.3636	6.6907
1999	3.7265	4.5340	5.4096	5.9593	6.3716	6.7008
2000	3.7040	4.5318	5.4246	5.9868	6.4055	6.7377
2001	3.7807	4.6639	5.6107	6.1924	6.6169	6.9451
2002	3.7807	4.6917	5.6599	6.2538	6.6837	7.0167
2003	3.6005	4.5621	5.5876	6.2275	6.6896	7.0481
2004	3.9870	5.0446	6.1320	6.7836	7.2457	7.6023
2005	3.8545	4.8918	5.9737	6.6298	7.0961	7.4571
2006	3.8884	4.9292	6.0204	6.6797	7.1504	7.5161
2007	3.9730	5.0185	6.1174	6.7806	7.2531	7.6201

Note: $coagg[d]km$ is the logarithm of the number of cross-industry firms that are located within the $d$km buffer. $d$kms include 5, 10, 20, 30,40, and 50kms.

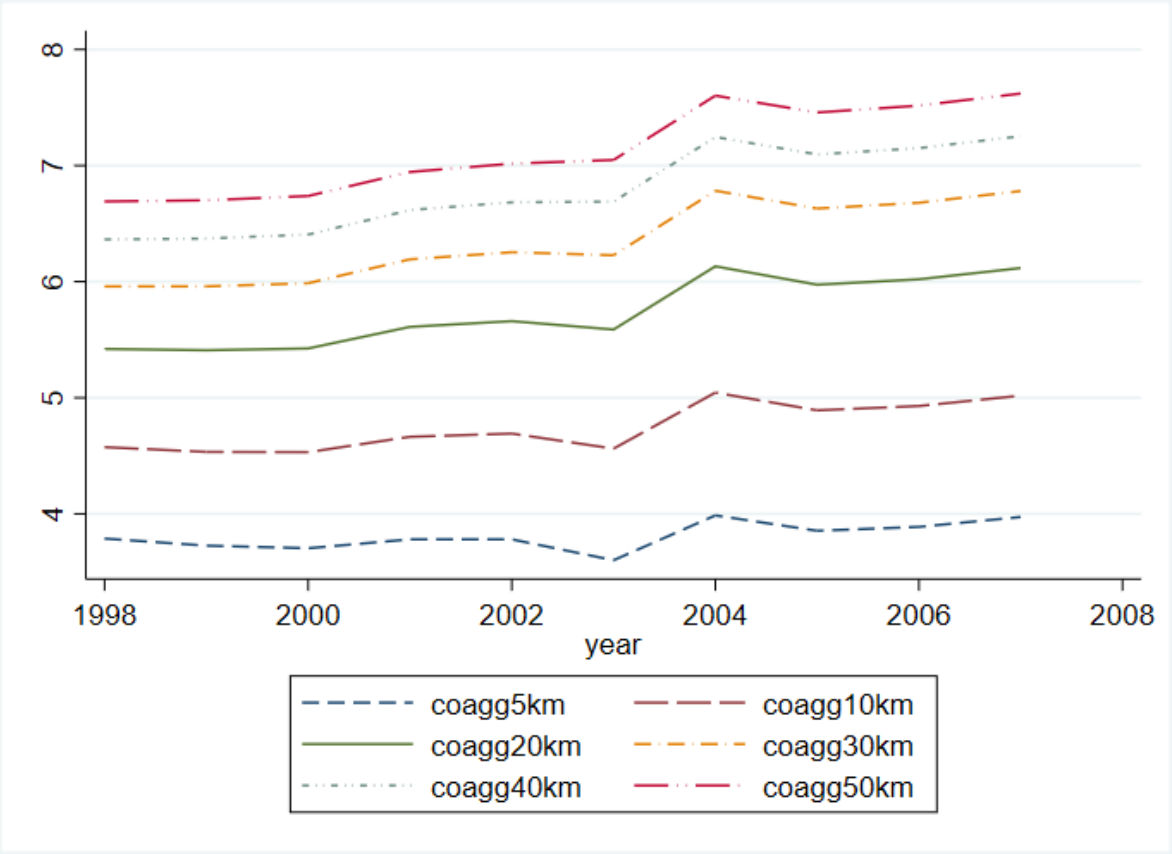


Figure 5.4: The trend of firm-level coagglomeration

Note: $coagg[d]km$ is the logarithm of the number of cross-industry firms that are located within the $d$km buffer. $d$kms include 5, 10, 20, 30,40, and 50kms. The average of $coagg[d]km$ per year and their trend are displayed in this figure. The value of firm-level within-industry agglomeration increases when the radius grows from 5km to 50km.

Table 5.21 illustrates the summary statistics of variables used in the following regressions. The following estimations use the sample only containing firms that have not moved to other counties during the sample period in order to mitigate the endogeneity issue. The $\Delta coagg[d]km$ has a much smaller sample size because only firms that exist in the 1998 NBS dataset are used for counting new entrants.

Table 5.21: Summary statistics of key variables

Variable	Obs	Mean	Std. dev.	Min	Max
lpacf	1,713,138	4.0207	1.1795	0.4557	6.9509
ln(highway access)	1,713,185	-1.8129	1.4709	-5.1677	2.3595
lndistance	1,713,185	1.8129	1.4709	-2.3595	5.1677
age	1,641,349	1.8441	0.9613	0	3.8712
state	1,711,471	0.1200	0.3250	0	1
size	1,713,185	8.2416	1.6269	4.12874	12.6446
intermediate ratio	1,663,149	0.6791	0.1525	0.18986	0.9723
labour productivity	1,664,852	3.9011	1.2293	0.37009	6.7766
new product ratio	1,676,700	0.0308	0.1396	0	0.9816
coagg5km	1,713,185	3.8382	1.6258	0	7.3727
coagg10km	1,713,185	4.8060	1.6902	0	7.7698
coagg20km	1,713,185	5.8266	1.6801	1.3863	8.5333
coagg30km	1,713,185	6.4508	1.6453	2.0794	9.1104
coagg40km	1,713,185	6.9011	1.5988	2.5649	9.4455
coagg50km	1,713,185	7.2520	1.5482	2.9957	9.6516
$\Delta coagg5km$	217,945	-0.0281	0.6934	-3.6337	3.0425
$\Delta coagg10km$	217,945	-0.0183	0.6011	-3.1616	2.7597
$\Delta coagg20km$	217,945	-0.0059	0.4737	-2.5069	2.2892
$\Delta coagg30km$	217,945	0.0008	0.3893	-2.0571	1.9213
$\Delta coagg40km$	217,945	0.0065	0.3353	-1.7579	1.7114
$\Delta coagg50km$	217,945	0.0110	0.2942	-1.4733	1.5635

5.6.3 Results for Coagglomeration Channel

Table 5.22 shows the FE-2SLS estimation results for the relationship between highway access and coagglomeration. Firms that have not moved to other counties during the sample period are used as the sample in this section. The coefficients of ln(highway access) are all positive and statistically significant, which means that a 1% change in highway access is associated with a 0.4% to 1.0% change in coagglomeration of firms in a radius of 5km to 50km, holding other control variables fixed. This proves that the channel of coagglomeration is valid and that highways can affect productivity through coagglomeration.

Table 5.23 displays the results after adding coagglomeration in the FE-2SLS estimations. The coefficient of ln(highway access) variable is statistically significant but reduces by about 1/3 and more after adding coagglomeration captured by different radii. This further supports the view that coagglomeration is a strong channel.

Table 5.22: The effects of highway access on coagglomeration

	Δcoagg5 (1)	Δcoagg10 (2)	Δcoagg20 (3)	Δcoagg30 (4)	Δcoagg40 (5)	Δcoagg50 (6)
Second stage of FE-2SLS						
Dependent variable: coagglomeration with different radii						
ln(highway access)	1.0166***	0.9461***	0.7543***	0.6074***	0.4901***	0.4019***
	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)
intermediate ratio	0.0058	0.0137	0.0101	0.0065	0.0035	-0.0029
	(0.011)	(0.009)	(0.008)	(0.006)	(0.005)	(0.005)
labour productivity	0.0035**	0.0068***	0.0067***	0.0063***	0.0054***	0.0037***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
new product ratio	0.0557***	0.0528***	0.0501***	0.0439***	0.0395***	0.0393***
	(0.009)	(0.008)	(0.007)	(0.006)	(0.005)	(0.004)
age	0.0605***	0.0459***	0.0331***	0.0269***	0.0230***	0.0210***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
state	0.0678***	0.0659***	0.0636***	0.0573***	0.0541***	0.0524***
	(0.009)	(0.008)	(0.006)	(0.005)	(0.005)	(0.004)
size	0.0103***	0.0117***	0.0124***	0.0114***	0.0105***	0.0100***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562	1,408,562
Number of panel id	362,152	362,152	362,152	362,152	362,152	362,152
Underidentification test	13480	13480	13480	13480	13480	13480
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	26401	26401	26401	26401	26401	26401

Note: The dependent variable is coagglomeration with a radius of 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV. Standard errors clustered at the firm level are displayed in parentheses. *,** and *** mean the coefficient is significant at the 10%, 5% and 1% levels respectively.

Table 5.23: Coagglomeration channel

Second stage of FE-2SLS Dependent variable:lnTFP (LP-ACF)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(highway access)	0.0189***	0.0109***	0.0061**	0.0048	0.0057*	0.0066**	0.0088***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
coagg5km		0.0079***					
		(0.001)					
coagg10km			0.0136***				
			(0.001)				
coagg20km				0.0188***			
				(0.002)			
coagg30km					0.0219***		
					(0.002)		
coagg40km						0.0252***	
						(0.002)	
coagg50km							0.0254***
							(0.002)
age	0.0206***	0.0201***	0.0200***	0.0200***	0.0200***	0.0200***	0.0201***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
state	-0.0756***	-0.0762***	-0.0765***	-0.0768***	-0.0769***	-0.0770***	-0.0769***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
size	-0.0322***	-0.0323***	-0.0323***	-0.0324***	-0.0324***	-0.0325***	-0.0324***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Year FE	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES	YES
Observations	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395	1,477,395
Number of panel id	377,651	377,651	377,651	377,651	377,651	377,651	377,651
Underidentification test	29779	27232	25836	25338	26033	27014	27947
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	51985	43824	40106	38250	39392	41319	43227

Note: The dependent variable is lnTFP, measured by the LP ACF-correction method. The instrument for highway access in FE-2SLS in Columns 3 and 4 is the LCP IV.

5.6.4 New Entrants

This section uses new entrants to capture coagglomeration since 1998. In order to evaluate the causal relationship between highways and new entrants, this study uses the LCP IV with NEN target nodes for highway access. Table 5.24 displays the second stage of FE-2SLS results with the dependent variables of change in coagglomeration. The number of firms' coagglomeration in the radii is the benchmark, the difference between the benchmark and the number of firms afterwards is used to construct firm-level coagglomeration. In Table 5.24, the positive and significant coefficients of ln(highway access) indicate that highways attract new firms or increase relocated firms, which stimulates coagglomeration. For a radius of 10km, the elasticity of new entrants with respect to highway access is 0.8136, while the larger the radius, the smaller the coefficients of ln(highway access).

Table 5.24: The effects on new entrants with IV

	Δcoagg5 (1)	Δcoagg10 (2)	Δcoagg20 (3)	Δcoagg30 (4)	Δcoagg40 (5)	Δcoagg50 (6)
Second stage of FE-2SLS						
Dependent variable: change in coagglomeration						
ln(highway access)	0.8914***	0.8136***	0.6182***	0.4735***	0.3729***	0.2953***
	(0.015)	(0.013)	(0.009)	(0.008)	(0.006)	(0.005)
intermediate ratio	-0.0811***	-0.0930***	-0.0864***	-0.0783***	-0.0788***	-0.0835***
	(0.029)	(0.025)	(0.019)	(0.015)	(0.013)	(0.011)
labour productivity	-0.0131***	-0.0133***	-0.0127***	-0.0125***	-0.0138***	-0.0148***
	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
new product ratio	-0.0165	-0.0214	-0.0109	-0.0009	0.0036	0.0035
	(0.031)	(0.027)	(0.021)	(0.017)	(0.014)	(0.012)
age	0.0509***	0.0424***	0.0281***	0.0222***	0.0183***	0.0156***
	(0.010)	(0.008)	(0.006)	(0.005)	(0.004)	(0.004)
state	0.0429**	0.0403**	0.0288**	0.0204**	0.0194**	0.0153**
	(0.019)	(0.017)	(0.013)	(0.010)	(0.009)	(0.007)
size	-0.0119**	-0.0122**	-0.0105***	-0.0065**	-0.0059**	-0.0050**
	(0.006)	(0.005)	(0.004)	(0.003)	(0.002)	(0.002)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	166,927	166,927	166,927	166,927	166,927	166,927
Number of panel id	68,518	68,518	68,518	68,518	68,518	68,518
Underidentification test	3781	3781	3781	3781	3781	3781
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	8292	8292	8292	8292	8292	8292

Note: This table presents the second stage of the FE-2SLS estimator. The dependent variable is Δ coagglomeration with different radii from 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes.

5.6.5 Robustness Checks: without Targeted Cities

As explained above, access to highways for firms in the targeted cities is not a random process. This section excludes firms in targeted cities to investigate the relationship between highways and firm-level productivity through the channel of coagglomeration.

Table 5.25 shows the second stage of FE-2SLS results for firms that are not in the targeted cities in the NTHS plan. The estimated coefficients of the highway variable are all positive and statistically significant, which indicates that highways increase coagglomeration of firms even in non-targeted cities. Compared with Table 5.24, the highway elasticity is slightly smaller, which indicates that for firms in non-targeted cities, the effects of highway access on new entrants are only slightly smaller.

Table 5.25: The effects of highway access on coagglomeration without targeted cities

	Δcoagg5 (1)	Δcoagg10 (2)	Δcoagg20 (3)	Δcoagg30 (4)	Δcoagg40 (5)	Δcoagg50 (6)
Second stage of FE-2SLS						
Dependent variable: change in coagglomeration						
ln(highway access)	0.7009***	0.6492***	0.5122***	0.4039***	0.3341***	0.2793***
	(0.016)	(0.014)	(0.011)	(0.009)	(0.008)	(0.007)
intermediate ratio	-0.1600***	-0.1675***	-0.1497***	-0.1633***	-0.1592***	-0.1629***
	(0.044)	(0.039)	(0.030)	(0.025)	(0.022)	(0.020)
labour productivity	-0.0321***	-0.0293***	-0.0257***	-0.0281***	-0.0302***	-0.0319***
	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)
new product ratio	-0.0491	-0.0210	0.0045	0.0203	0.0208	0.0255
	(0.062)	(0.056)	(0.044)	(0.037)	(0.033)	(0.029)
age	0.0332**	0.0224*	0.0057	0.0072	0.0078	0.0054
	(0.014)	(0.012)	(0.009)	(0.008)	(0.007)	(0.006)
state	0.0372	0.0396*	0.0164	0.0045	0.0053	0.0040
	(0.027)	(0.024)	(0.019)	(0.016)	(0.013)	(0.012)
size	0.0008	-0.0026	-0.0021	-0.0009	-0.0020	-0.0005
	(0.009)	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)
Year FE	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES
Industry FE	YES	YES	YES	YES	YES	YES
Observations	61,361	61,361	61,361	61,361	61,361	61,361
Number of panel id	25,080	25,080	25,080	25,080	25,080	25,080
Underidentification test	1810	1810	1810	1810	1810	1810
p-value	0.000	0.000	0.000	0.000	0.000	0.000
Weak identification test	5604	5604	5604	5604	5604	5604

Note: This table presents the second stage of the FE-2SLS estimator. The sample is the firms in non-targeted cities in the NTHS plan. The dependent variable is coagglomeration with radii of 5km to 50km. The instrument for highway access in FE-2SLS is the LCP IV.

5.7 Export Channel

5.7.1 Hypothesis Development

Regarding the mechanism of export, the reduction in transport costs may lead to export, and the learning-by-exporting effect in turn increases firm productivity. We investigate the causal relationship between highway and export, but the relationship between export and productivity according to the literature is more complex, and there are selection effects and learning-by-exporting effects. With the selection effect, highways may attract more productive firms to export. Since this chapter investigates the effects of highways on productivity through the export channel, learning-by-exporting is more emphasized in this mechanism analysis. This section summarizes the literature that supports the learning-by-exporting effect and some literature about the positive relationship between transport infrastructure and export. Thus the hypothesis is that with the learning-by-exporting effect for Chinese firms better highway access leads to more exporting activities, which then increases firm productivity.

Export and Productivity. The relationship between exports and productivity has been extensively examined from both theoretical and empirical viewpoints. A foundational theoretical study by Melitz (2003) presents a dynamic model of international trade and export decision-making involving heterogeneous firms. This model investigates firms' decisions related to market entry, exit, exporting, and foreign direct investment (FDI) through two main effects: the self-selection effect and the learning-by-exporting effect. The self-selection effect posits that firms with higher productivity are more likely to enter export markets, while less productive firms either stay in the domestic market or exit entirely if they are the least productive. Melitz (2003) also argues that trade induces intra-industry competition and reallocates resources towards more productive firms, thus enhancing aggregate industry productivity. This process, termed the learning-by-exporting effect, improves industry productivity as more efficient firms expand.

Empirical research on export-related productivity examines whether firms are more productive before or after entering export markets. Some studies argue that only the most productive firms can engage in international competition (Clerides et al. 1998, Bernard & Jensen 1999, Delgado et al. 2002, Isgut 2001), while others suggest that exporting can enhance productivity (Van Biesebroeck 2005, Greenaway et al. 2008).

Research supporting self-selection effects shows that more productive firms are more likely to export. Bernard et al. (1995) find that US exporters, who are more productive and larger, dominate manufacturing activity. Similarly, Clerides et al. (1998) find that more productive firms become exporters, with no significant cost impacts from previous exporting activities. Greenaway & Kneller (2004) and Greenaway & Yu (2004) provide UK data showing that more productive firms enter export markets and suggest focusing on firm development rather than subsidies. Alvarez & López (2005) find evidence of both self-selection and learning-by-exporting in Chile. Aw et al. (2000) observe that Taiwanese firms exhibit both self-selection and learning-by-exporting, while South Korean firms show weaker effects, possibly due to high entry costs and significant government subsidies. They note that knowledge spillovers from exporters might also obscure learning effects.

Evidence for learning-by-exporting is evident, particularly in less developed countries. Wagner (2007) find mixed results across 34 countries, which may reflect varying development stages and methodologies. For instance, Van Biesebroeck (2005) highlights learning-by-exporting in sub-Saharan Africa, and De Loecker (2007) provides strong evidence of productivity gains for Slovenian firms starting to export. Loecker (2013) suggests that the impact is not uniform and Blalock & Gertler (2004) find that plants appear to have about a 2% to 5% increase in productivity following the beginning of exporting. In Colombia, Isgut (2001) finds that exporters experience faster productivity growth over five years compared to non-exporters. Baldwin & Gu (2003) report faster productivity growth for Canadian exporters, with stronger effects for younger and domestic firms.

With respect to the learning-by-exporting effect in China, most literature supports this effect in Chinese export firms. Kraay (1999) finds significant performance improvements for established Chinese exporters, though effects for new entrants are less clear, potentially due to preferential access to foreign exchange. Van Biesebroeck (2014) finds improved performance

for Chinese SMEs post-export entry and explains that exporting provides a way for firms to expand their sales without the need to extend additional trade credit. Lin (2015) shows that a 1% increase in exporting boosts total factor productivity by 0.04%. Yang & Mallick (2010) confirm evidence for export premium, self-selection, and learning-by-exporting, with additional productivity gains particularly evident for new entrants in their second year.

Transport and Export. Empirical research consistently highlights that improved transport infrastructure reduces travel costs and boosts export activity. Liu et al. (2023) examine the impact of highway access on firm-level exports using data from 2000 to 2006, including firm-level information, export data, and GIS highway data. They measure highway access by the proximity of firms to highways and the length of surrounding highways. Their analysis, based on Melitz (2003) and Chaney (2008), shows that better highway access increases firm exports, particularly for less productive firms.Guo & Yang (2019) investigate the effect of transport infrastructure on exports using a stochastic frontier gravity model and data from China and abroad (2006-2012). They find that improved transport accessibility, measured by distances from cities to export ports, enhances trade efficiency by lowering transport costs.

Coşar & Demir (2016) assess the impact of highway upgrades on trade flows from Turkish provinces to international gateways. Their study, spanning 2003 to 2012, finds that increased road capacity significantly boosts trade flows and reduces transport costs by up to 70% when single carriageways are upgraded to highways. Martincus & Blyde (2013) use the 2010 Chilean earthquake as a natural experiment to study the effect of transport infrastructure disruption on exports. Their analysis of firm-level data from 2008 to 2011 shows that damaged transport infrastructure markedly decreases exports.

In summary, the above literature shows that better transport infrastructure, particularly highways, improves export performance by lowering travel costs and enhancing trade efficiency. Studies have found that better highway access increases exports, especially for less productive firms (Liu et al. 2023). Moreover, the learning-by-exporting effect is significant in various contexts, including Chinese firms, where increased exporting activity has been linked to higher productivity (Van Biesebroeck 2014, Lin 2015). Based on this evidence, the hypothesis is that for Chinese firms, improved highway access leads to increased exporting activities, which subsequently enhances firm productivity.

5.7.2 Summary Statistics of Export Channel

This section investigates the effect of highway access on firm TFP through the channel of export. The above literature review shows that the learning-by-exporting effect in China is predominant (Kraay 1999, Van Biesebroeck 2014, Dai & Yu 2013, Lin 2015, Yang & Mallick 2010). With the evidence from the literature on learning-by-exporting in China, this section further investigates whether highway access facilitates exports in China.

This study uses matched customs data and the NBS firm-level data from 2000 to 2006 to investigate the mechanism of trade. The transaction-level export value of firms from the Chinese General Administration of Customs is collected and added up to give the annual value of each exporter. The Chinese Customs dataset collects transaction-level exports and imports. It explicitly contains the categories of products, the values and unit price, trade regime, port information, the destination, shipment method, firm information, etc. Products are classified according to the eight-digit Harmonized System. The trade regime information makes the processing or ordinary trade visible in order to study this heterogeneous feature.

The Customs dataset provides monthly import and export information for firms. The monthly data are aggregated into annual import and export values for each firm. Following Yu (2015) and Ding, Jiang & Sun (2016), the trade intermediaries are removed from the sample. The customs dataset and the ASIF dataset can be merged mainly with the firm name, postcode, phone number and firm address are used when the firm name does not match. The currency unit of export and import value in the custom data is US dollars, while the currency unit of intermediate inputs and output in the ASIF dataset is Chinese yuan. This study uses the annual central parity rate USD/CNY as a concordance.

The merged data statistics are shown in Table 5.26. The sample period is from 2000 to 2006. The merge ratio is calculated as the number of exporting firms in the merged NBS and customs dataset divided by that in the NBS dataset, which is around 60%. The percentage of firms that are matched has an overall increasing trend from 2000 to 2006. Compared with the

number of exporters in the NBS dataset, the Customs dataset contains a much greater number of exporters. The Customs dataset has the advantage that it collects the transaction-level data of all sizes of firms, while the NBS dataset only collects information on firms whose sales are above a certain threshold.

Table 5.26: Merged data from Customs and NBS datasets

year	Custom Obervations	NBS Obervations	NBS Export	Merged percent	Merged (export)
2000	65210	135099	35764	51%	18223
2001	69827	144325	39084	54%	21161
2002	83456	155178	43492	58%	25269
2003	100630	172754	49585	62%	30896
2004	123996	232473	73201	58%	42360
2005	158111	242129	72752	65%	47061
2006	162644	269734	76721	67%	51302
Total	763874	1351692	390599	60%	236272

Notes: The customs dataset and the NBS dataset can be merged mainly with firm name, supplemented by postcode, phone number and firm address. The Customs dataset provides monthly import and export information for firms. The monthly data are aggregated into annual import and export values for each firm.

Processing exporters are required to sell all their products to the international market (export is the total revenue) and have duty exemption import. Processing trade means assembling imported inputs into final goods for resale in export markets. This section assigns firms that only deal with processing trade as processing exporters, and firms that are only involved in ordinary trade as ordinary exporters.

The trends for the share of processing and ordinary exporters are illustrated in Figure 5.5. The share of firms that are engaged in both processing and ordinary is not shown in this figure. The share of processing exporters decreased gradually from 2000 to 2006 in China, while the share of processing exporters increased by one-third over the sample period.

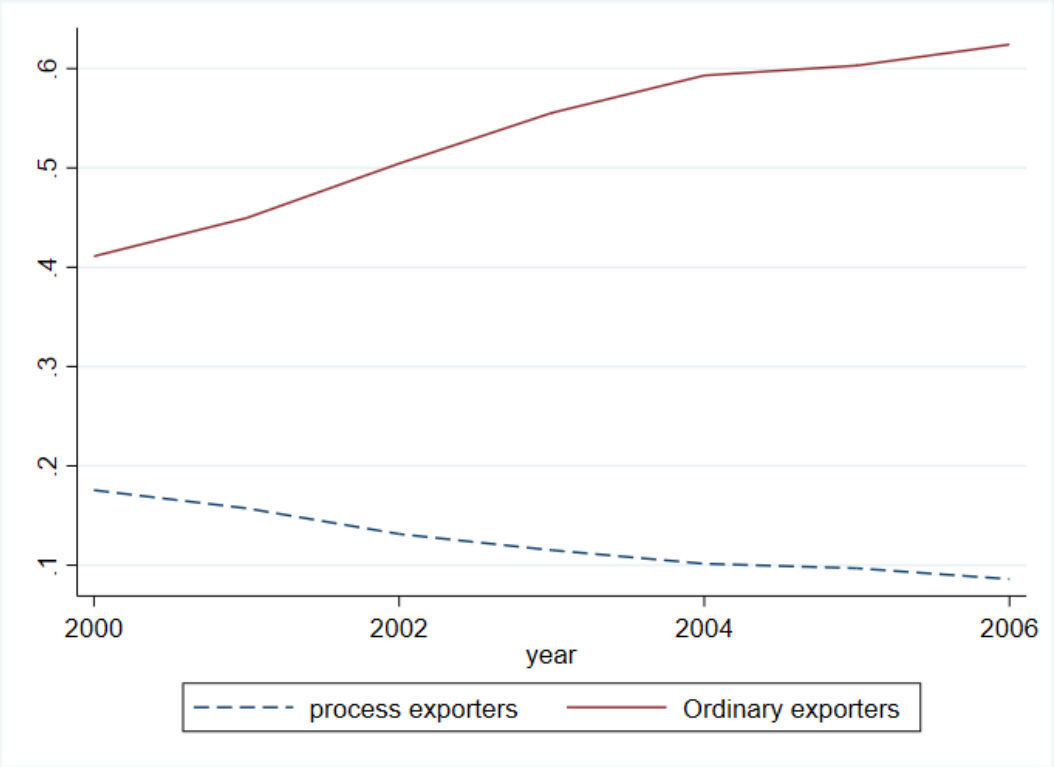


Figure 5.5: The trend of processing and ordinary exporters

Note: Firms that only deal with processing trade are regarded as processing exporters in this figure, and firms that are only involved in ordinary trade are regarded as ordinary exporters, but state the rule can be relaxed.

5.7.3 Results for Export Channel

Tables 5.27 and 5.28 display the results when regressing export value on highway access. Processing exporters import materials free of tariffs and sell their output abroad. Using firm-level data from 2000 to 2005 Dai et al. (2016) find that processing exporters in China are associated with low productivity. They indicate that processing exporters are less productive than non-processing exporters and nonexporters. Processing exporters benefit from low fixed costs for exporting processed goods and beneficial industrial policies. Additionally, processing firms are more likely to be located near ports in coastal areas, and may take less advantage of highway access compared with other firms. Thus, we split the samples into processing exporters and non-processing firms to investigate the heterogeneous effect of highways on exports.

Table 5.28 displays the second-stage results for the FE-2SLS estimation on the effect of highway access on export. The results indicate that highway access increases export activities for exporters, which is consistent with the results of Liu et al. (2023) who also find that highway density positively promotes exports with the IV of the average gradients of the surface.

Table 5.27: Export channel

	(1)	(2)	(3)	(4)
Dependent variable:ln(export+1)	All	Processing Only	Ordinary Only	Both
ln(highway access)	0.0218***	0.0174	0.0100	0.0041
	(0.005)	(0.011)	(0.008)	(0.006)
age	0.1051***	0.2780***	0.0899***	0.1267***
	(0.011)	(0.031)	(0.015)	(0.014)
SOE	-0.1297***	-0.0892	-0.0963**	-0.0756**
	(0.031)	(0.099)	(0.042)	(0.037)
size	0.2478***	0.1629***	0.2061***	0.1948***
	(0.007)	(0.016)	(0.010)	(0.009)
Year FE	Yes	Yes	Yes	Yes
N	231343	37209	122415	71719
r2_a	.1118	.03082	.1543	.1319

Note: The dependent variable is ln(export+1). Column 2 presents the results for processing firms. Column 3 displays results for firms that are only involved in ordinary trade. Column 4 reports results for firms that participate both in processing and ordinary trade.

Table 5.28: Export channel with IV

	(1)	(2)	(3)	(4)
Second stage of FE-2SLS Dependent variable:ln(export+1)	All	Processing Only	Ordinary Only	Both
ln(highway access)	0.0409*	0.4476*	0.0186	0.0544*
	(0.024)	(0.257)	(0.030)	(0.029)
age	0.1051***	0.2614***	0.0900***	0.1258***
	(0.010)	(0.030)	(0.013)	(0.012)
SOE	-0.1303***	-0.0476	-0.0969**	-0.0752**
	(0.028)	(0.096)	(0.040)	(0.035)
size	0.2476***	0.1679***	0.2060***	0.1948***
	(0.006)	(0.014)	(0.009)	(0.008)
Year FE	Yes	Yes	Yes	Yes
Observations	206,377	30,609	104,704	60,457
Underidentification test	1260	10.33	845.1	309.6
p-value	0.000	0.000	0.000	0.000
Weak identification test	1513	10.31	1168	358.7

Note: The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes. Year fixed effect is included. The dependent variable is ln(export+1). For the panel data analysis from 2000 to 2006. Column 2 presents the results for processing firms. Column 3 displays results for firms that are only involved in ordinary trade. Column 4 reports results for firms that participate both in processing and ordinary trade.

Regarding the results for processing exporters, the coefficient of the highway variable is positive but owning to the sample size, the LCP IV is weaker. The elasticity of exports for the whole merged sample with respect to highway access is 0.0409. Additionally, for firms that participate both in processing and ordinary trade, the elasticity of exports with respect to highway access is 0.0544. Liu et al. (2023) find that for less productive exporters, the benefits of highway access on exports are greater. They explain that more highway access results in lower markups for less productive exporters, whereas markups rise for more productive exporters. Since processing exporters are less productive than non-processing exporters (Dai et al. 2016) and less productive exporters are associated with stronger highway access effects, this explains why the effects of highways on processing exporters are stronger in this study.

Table 5.29 presents the effect of highway access on exports for firms in different areas and firms with different ownership. The effects of highway access on exports are significant for foreign firms. When splitting the sample into coastal and inland areas, the coefficients of the highway variable are insignificant.

Table 5.29: Export channel with different ownership and areas with IV

	Ownership			Area	
	Foreign	Private	SOE	Coastal	Inland
	(1)	(2)	(3)	(4)	(5)
Second stage of FE-2SLS					
Dependent variable:ln(export+1)					
ln(highway access)	0.2498**	0.0209	-0.0993	0.0511	0.0251
	(0.100)	(0.046)	(0.067)	(0.032)	(0.033)
age	0.2977***	0.0997***	0.0722	0.1234***	0.0404
	(0.023)	(0.019)	(0.046)	(0.010)	(0.025)
SOE				-0.1367***	-0.0719
				(0.034)	(0.054)
size	0.1973***	0.2535***	0.1471***	0.2528***	0.1991***
	(0.011)	(0.014)	(0.047)	(0.006)	(0.020)
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	50,507	47,008	10,252	183,994	22,380
Underidentification test	60.88	404.2	125.8	844.4	415.9
p-value	0.000	0.000	0.000	0.000	0.000
Weak identification test	63.22	572.6	192.9	940.4	679.6

Note: This table presents the second stage of the FE-2SLS estimator. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes. Year fixed effect is included. The dependent variable is ln(export+1). Columns 1-3 show the effect of highway access on exports for firms with different types of ownership, while columns 4 and 5 show this effect for firms in different areas.

5.7.4 Robustness Checks: Export channel with Transport Time

This section constructs the variable called access to port, which is the travel time from firms to port through the road network, where the road network is time-variant with time-variant highways and the primary and secondary roads in 2000 in China.

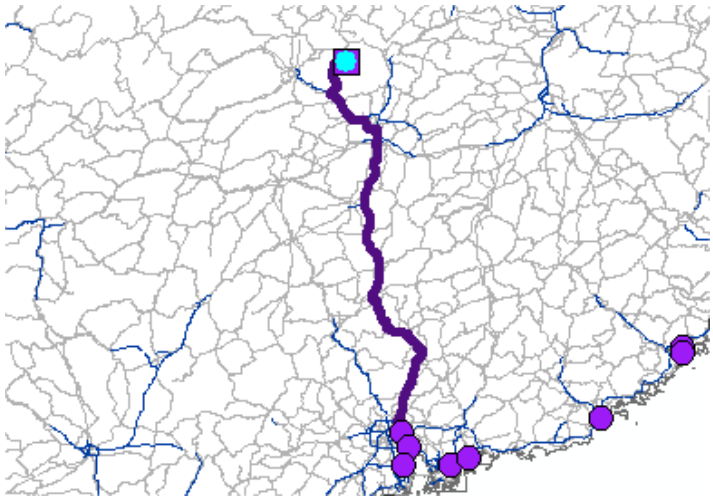


Figure 5.6: Firms' transport costs to port

Figure 5.6 provides an example of the routes from a firm to a port with the least transport cost. The blue lines are highways in 2002, and the grey lines are the primary and secondary roads in China in 2000. The speed limit for highways, primary roads and secondary roads are 120km/h, 80km/h, and 60km/h, respectively. In ArcGIS, a rank is set for choosing highways first and secondary roads last because the time cost for primary and secondary roads is higher due to traffic lights and junctions.

The regression equation with access to port as the independent variable is

$$lnexp_{it} = \sigma_0 + \beta_2 port\ access_{it} + \tau_2 Controls_{it} + \delta_i + \sigma_t + \varepsilon_{it} \tag{5.11}$$

where $port\ access_{it}$ is minus the logarithm of transport time, $port\ access_{it} = -ln(transport\ time)$. The $Controls_{it}$ include the firm age, size and SOE.

Table 5.30 displays the results when regressing the export value on port access. The coefficient of port access on export is positive and statistically significant at the 10% level in Column 1, and port access elasticity is 0.0818. For processing exporters as shown in Column 2, a 1% change in port access increases exports by 0.2550%. The results indicate that with the upgrade of highways, lower transport time increases exporting. The results are consistent with using highway access as the independent variable.

Table 5.30: Export channel with port access as independent variable

	(1)	(2)	(3)	(4)
	All	Processing Only	Ordinary Only	Both
Second stage of FE-2SLS Dependent variable:ln(export+1)				
port access	0.0818*	0.2550*	0.0523	0.0850*
	(0.047)	(0.137)	(0.083)	(0.044)
age	0.1043***	0.2844***	0.0891***	0.1247***
	(0.010)	(0.027)	(0.013)	(0.012)
SOE	-0.1301***	-0.0785	-0.0929**	-0.0757**
	(0.029)	(0.090)	(0.040)	(0.035)
size	0.2480***	0.1618***	0.2061***	0.1943***
	(0.006)	(0.013)	(0.009)	(0.008)
Year FE	Yes	Yes	Yes	Yes
Observations	205,672	30,562	104,330	60,189
Underidentification test	800.6	67.56	383.4	228.3
p-value	0.000	0.000	0.000	0.000
Weak identification test	896.9	75.21	417.2	263

Note: The instrument for port access in FE-2SLS is the LCP IV with NEN target nodes. Year fixed effect is included. The dependent variable is ln(export+1). Column 2 presents the results for processing firms. Column 3 displays results for firms that are only involved in ordinary trade. Column 4 reports results for firms that participate both in processing and ordinary trade.

Table 5.31 presents the effect of port access on exports for firms in different areas and firms with different ownership. The results are consistent with those using highway access as the independent variable. With more completed highways, the lower transport time to a nearby port increases the exports of foreign firms.

Table 5.31: Export channel with different ownership and areas

	Ownership			Area	
	Foreign (1)	Private (2)	SOE (3)	Coastal (4)	Non-costal (5)
Second stage of FE-2SLS					
Dependent variable:ln(exp+1)					
port access	0.2841**	0.0567	-0.2900	0.0728	0.3461
	(0.110)	(0.116)	(0.196)	(0.044)	(0.461)
age	0.3150***	0.0982***	0.0673	0.1227***	0.0413*
	(0.023)	(0.019)	(0.046)	(0.010)	(0.025)
SOE				-0.1376***	-0.0752
				(0.034)	(0.054)
size	0.1981***	0.2542***	0.1428***	0.2534***	0.1986***
	(0.011)	(0.014)	(0.048)	(0.006)	(0.020)
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	50,400	46,930	10,113	183,347	22,322
Underidentification test	118.9	184.3	86.30	740	148.6
p-value	0.000	0.000	0.000	0.000	0.000
Weak identification test	130.9	211.5	91.62	852.8	165.2

Note: The instrument for highway access in FE-2SLS is the LCP IV. Year fixed effect is included. The dependent variable is ln(export+1). Columns 1-3 show the effect of highway access on exports for firms with different types of ownership, while columns 4 and 5 show this effect for firms in different areas.

5.8 Innovation Channel

5.8.1 Hypothesis Development

Highways can also facilitate innovation with easier connections between firms for knowledge spillovers. Innovation is an important driver of firm productivity. Thus this channel is worth investigating and it is hypothesized that better highway access brings knowledge spillovers, which in turn stimulates firm productivity.

There is some evidence that highways can improve innovation. Agrawal et al. (2017) investigate the impact of inter-state highways on regional innovation using historical data related to planned highways, railways, and exploration routes as sources of exogenous variation. They select 268 Metropolitan Statistical Areas (MSAs) defined in 1993 by the U.S. Office of Man-

agement and Budget. Detailed U.S. patenting activity data are obtained from the U.S. Patent and Trademark Office (USPTO), which includes information on the affiliation and location of patenting inventors in a region. Using the address information of inventors, patents were assigned to their respective MSAs.

In terms of dependent variables, Agrawal et al. (2017) measure innovative activity using a citation-weighted count of U.S. patents. They also consider an unweighted patent count as an additional metric of innovation. For their main explanatory variable, they use the total number of kilometres of interstate highway in the region in 1983. This information is gathered from the Highway Performance and Monitoring System data. They then use the distances between cities/towns within an MSA to conduct their analysis of local knowledge flows. They find that a 10% increase in a region's stock of highways leads to an approximate 1.7% increase in regional patenting over a five-year period. This is significant, being comparable to an over 3% increase in regional corporate R&D investments.

Their research also uncovers a mechanism by which transportation infrastructure stimulates innovation by facilitating local knowledge flows and enhancing the probability of innovators accessing knowledge inputs from more distant local sources. This finding broadens the understanding beyond traditional views that infrastructure development promotes innovation predominantly through agglomeration economies.

Zeng et al. (2022) investigate the spillover effect of highway connection on firm-level productivity in China, arguing that highways overcome localization of knowledge by facilitating the flow of ideas, information, and talent across regions, consequently improving firm productivity. The paper also explores how innovation mediates the relationship between transport infrastructure and productivity, providing a new perspective in the discourse of economic geography.

Zeng et al. (2022) use the ASIF dataset from 1998 to 2007 in China. The explanatory variable, highway accessibility, is a dummy variable indicating whether the county where the firm is located is connected to highway systems in a given year. The innovation performance is measured by the new product output value of a firm. Their results show a positive correlation between connection to highways and firm productivity. This relationship is found to be mediated by innovation performance. Furthermore, the mediating effect is stronger in regions with higher levels of market liberalization and more developed intermediaries.

5.8.2 Results for Innovation Channel

Zeng et al. (2022) investigate the spillover effect of highway connections on firm-level productivity in China through the innovation channel, arguing that highways overcome localization of knowledge by facilitating the flow of ideas, information, and talent across regions, consequently improving firm productivity.

The effects of highways on innovation are heterogeneous. Ding, Sun & Jiang (2016) indicate that owing to competition, firms' performances are different. Firms that are close to the world technology frontier or far away from the frontier behave differently on innovation and growth when facing import competition. They use US labour productivity as the world technology frontier and compare it with Chinese labour productivity to capture the distance to the technology frontier. This study splits the sample into high and low TFP to check the heterogeneity. Firms with higher labour productivity are expected to be involved in more innovative activities, and highways may facilitate their innovation more.

Table 5.32 uses the patent as the proxy for innovation. The FE-2SLS estimation results using all patents, invention patents, utility patents, and design patents are shown in Columns (1) to (4), respectively. The coefficients for highway access on patents are insignificant, except for the design patents. However, design patents only cover small-scale innovations, which cannot by themselves properly represent innovation. Thus, the results indicate that highways do not promote innovation when using the patent as the proxy for innovation.

Table 5.33 indicates the FE-2SLS estimation results with the dependent variable called new product ratio, which is used to capture innovation. Columns 2 and 3 use the sub-samples of high labour productivity firms and low labour productivity firms respectively. The coefficients for highway access on new product share are positive and statistically significant, which indicates that highway increases innovation activities. Thus innovation is a channel for the effects of highways on firm TFP. This result is consistent with that of Zeng et al. (2022), who also use new products to measure innovation. However, this study also uses patents as the proxies for innovation, and the coefficients for highway access on patents are insignificant.

Table 5.33 also shows that for firms with higher labour productivity, the effects of highways are statistically significant, while the low labour productivity firms cannot stimulate innovation through better highway access. This indicates that firms with higher labour productivity can facilitate their innovation through highway access. Firms with higher labour productivity have more skilled workers, and highway access increases the knowledge spillovers for those firms with less transport time and easier communications.

Table 5.32: Innovation channel-patents

Dependent variable	patent			
	All	invention	utility	design
	(1)	(2)	(3)	(4)
Second stage of FE-2SLS				
ln(highway access)	0.0018	-0.0005	-0.0008	0.0008**
	(0.002)	(0.000)	(0.001)	(0.000)
age	-0.0050***	-0.0012***	-0.0019***	-0.0004**
	(0.001)	(0.000)	(0.000)	(0.000)
SOE	0.0030	0.0012	0.0036**	0.0003
	(0.004)	(0.001)	(0.002)	(0.001)
size	0.0186***	0.0033***	0.0065***	0.0027***
	(0.001)	(0.000)	(0.000)	(0.000)
Year FE	Yes	Yes	Yes	Yes
Observations	1,669,826	1,669,826	1,669,826	1,669,826
Underidentification test	31296	31296	31296	31296
p-value	0.000	0.000	0.000	0.000
Weak identification test	47604	47604	47604	47604

Note: The dependent variables are the innovation measurements constructed by patent data. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes. The results for innovation measured by all patents, invention patents, utility patents, and design patents are presented in Columns (1) to (4).

Table 5.33: Innovation channel-new product ratio

VARIABLES	(1) All	(2) high labour productivity	(3) low labour productivity
Second stage of FE-2SLS			
Dependent variable: new product ratio			
ln(highway access)	0.0011**	0.0022**	0.0003
	(0.001)	(0.001)	(0.001)
age	-0.0012***	-0.0015***	-0.0016***
	(0.000)	(0.000)	(0.000)
state	0.0032***	0.0046***	0.0018
	(0.001)	(0.001)	(0.002)
size	0.0041***	0.0042***	0.0043***
	(0.000)	(0.000)	(0.000)
Year FE	YES	YES	YES
Observations	1,459,977	699,218	616,433
Number of panel id	409,969	218,769	198,323
Underidentification test	18145	7030	7773
p-value	0.000	0.000	0.000
Weak identification test	26761	9708	12394

Note: This table presents the FE-2SLS estimation results with the dependent variable being the new product ratio, which serves as a measure of innovation. Columns 2 and 3 use sub-samples of firms with high and low labour productivity, respectively. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes.

Table 5.34 presents the IV estimation results for the heterogeneity based on ownership. Columns 1, 2, and 3 uses the interaction term between highways and foreign-owned, private-owned and state-owned dummies, respectively. The estimated coefficient for the interaction term between highway and foreign ownership is positive and statistically significant, while the other two interaction terms are insignificant. The results indicate that for firms with foreign ownership, highways facilitate innovation more. Since innovation intuitively improves firm productivity, this further indicates that through the innovation channel, highways affect the productivity of firms with foreign ownership.

Table 5.34: Innovation channel with different ownership

	(1)	(2)	(3)
Second stage of FE-2SLS Dependent variable: new product ratio			
ln(highway access)	0.0010*	0.0013**	0.0013**
	(0.001)	(0.001)	(0.001)
Interaction_highway_foreign	0.0034**		
	(0.002)		
foreign	0.0041**		
	(0.002)		
Interaction_highway_private		-0.0006	
		(0.001)	
private		-0.0018*	
		(0.001)	
Interaction_highway_state			-0.0008
			(0.001)
state			0.0017
			(0.002)
age	-0.0012***	-0.0011***	-0.0013***
	(0.000)	(0.000)	(0.000)
size	0.0042***	0.0042***	0.0041***
	(0.000)	(0.000)	(0.000)
Year FE	Yes	Yes	Yes
Observations	1,462,382	1,450,929	1,459,977
Underidentification test	4538	17984	17236
p-value	0.000	0.000	0.000
Weak identification test	2415	13270	12021

Note: This table presents the IV estimation results for heterogeneity based on ownership. Columns 1, 2, and 3 use the interaction term between highways and dummy variables for foreign-owned, private-owned, and state-owned firms, respectively. The instrument for highway access in FE-2SLS is the LCP IV with NEN target nodes.

5.9 Conclusion

This chapter has investigated the effects of highway access on firms' TFP. Different methods have been examined to construct TFP, including the OLS, FE, OP, OP-ACF, LP, LP-ACF, and Wooldridge estimation. The LP-ACF method with a modification to the timing of input choices in the LP methods has been used to capture firms' TFP when investigating the impacts of highway access. The baseline results generated by the fixed effects model indicate that highway access is positively associated with firm-level TFP. To mitigate the endogeneity issue, the LCP IV is used in this chapter. The IV estimation yields consistent results for the positive effects of highway access on firm productivity.

Understanding how highways affect productivity is becoming increasingly important, which inspires this chapter. Highways can have a direct impact on firm-level productivity by improving product transportation efficiency. Additionally, through several channels, highways can also boost firm productivity. This chapter investigates four channels, including within-industry agglomeration, coagglomeration, export and innovation. Within-industry agglomeration effects and coagglomeration effects are not the same and have different features and advantages. Thus it is reasonable to examine them respectively as the channels. Firm-level agglomeration is the logarithm of the number of firms in the neighbourhood within the radii including 5, 10, 20, 30,40, and 50kms. Compared with EG industry-level agglomeration, the firm-level agglomeration index takes into account heterogeneity between firms and avoids the discontinuous issue in geography.

The mechanism analysis results indicate that highway access affects firm-level productivity through the within-industry agglomeration and coagglomeration channels. The results are consistent after only using new entrants as the sample, using employment to calculate the firm-level agglomeration and dropping big cities in the sample. The changes in the number of neighbouring firms compared with that of the initial year are used to capture the agglomeration of new entrants. This new entrants index captures the fact that highways can attract new firms to locate near them. The 1% change in highway access is associated with a 0.268 to 0.462% change in within-industry agglomeration, and a 0.4% to 1.0% change in coagglomeration of different radii from 5km to 50km, holding other control variables fixed. Moreover, when ln(TFP) is the dependent variable, the estimated coefficient of highway variable reduces by about 1/3 after adding the within-industry agglomeration and coagglomeration channels as controls, which means that these two channels efficiently capture the effects of highways on firm productivity.

Regarding the export channel, the IV estimation results indicate that highways stimulate exports. The elasticity of exports with respect to highway access is 0.0409. Then with the learning-by-exporting effects, highways can increase firm productivity in China. The effects of highway access on exports are significant for foreign firms, and firms participating in pro-

cessing trade. This chapter also uses transport time to port through the highway network as the robustness checks for the export channel. The IV estimation results for port access are consistent. Additionally, the innovation channel is also valid, especially for firms with higher labour productivity and firms under foreign ownership.

The limitations of this study are as follows. First, the firm-level agglomeration index is limited by the fact that only information for medium to large firms is collected by the NBS dataset. Thus using the number of other firms or the sum of employment in the neighbourhood to construct the index is not very accurate. Second, firm-level agglomeration is determined by the number of firms in the given radii rather than the spatial distribution, and thus, it is not appropriate to aggregate this firm-level agglomeration index to the industry level. Third, this chapter uses literature to support some channels, such as the learning-by-exporting effect in China, as it is a heavy task to further empirically examine every chain of the channel and it is difficult to deal with the endogeneity issue of the relationship between export and productivity.

This chapter has some policy implications related to improving firm productivity. This study finds that highway access promotes firm-level productivity through the channel of within-industry agglomeration and coagglomeration. Industrial zones and special zones can be fully used with fast access to highways. In these zones where within-industry agglomeration or coagglomeration between upstream and downstream firms appear, firm productivity is likely to be increased. Additionally, different types of ownership yield different effects of highways on productivity, and for firms with higher labour productivity, the effects of highway access are larger. These findings indicate that the industrial transformation involved in producing higher quality and more technological products is important, and with these industries, the positive effects of transport infrastructure can be fully explored.

5.10 Appendices for Chapter 5

Appendix 5.A Review of Mechanism Analysis Methods

Five mechanism analysis methods are commonly used. This section provides examples and models for each mechanism analysis method, listed in no particular order.

Table 5.A-1: Five mechanism analysis methods

Step 1	Step 2	Step 3	Description	Literature
X→Y	X→M		Regress Y on X and then regress M on X	Levchenko et al., 2009 (JDE)
X→Y	X+M→Y		Compare the coefficient changes of X	Alesina et al., 2011 (AER) Persico et al., 2004 (JPE) Holl, 2016 (JUE) Becker & Woessmann, 2009 (QJE)
X→M	M→Y		Regress M on X and then regress Y on M	Squicciarini, 2020 (AER)
X→Y	X→M	X+M→Y	Mediation analysis Criticism: unconvincing results, hard to address the Endogeneity problem of Ms	Faereng et al., 2021 (JPE) Wang and Zhang, 2018 (CER)
X*M→Y			Use Interaction term more suited for heterogeneity analysis	

Method #1: $X \rightarrow Y, X \rightarrow M$

The mechanism analysis method takes the form of

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Controls_1 + \varepsilon_1 \tag{5.A-1}$$

$$M = \beta_0 + \beta_1 X + \beta_2 Controls_2 + \varepsilon_2 \tag{5.A-2}$$

Levchenko et al. (2009) employ the impact of financial liberalization (X) on economic growth (Y) through the mechanisms of greater entry (increased number of establishments), more employment, capital accumulation and TFP growth. They first regress economic growth (Y) on the financial liberalization dummy variable (X). They then regress these channels (M) on the financial liberalization dummy variable (X), by replacing the dependent variable of output growth and keeping the same controls.

Method #2: Compare the Coefficient Changes of X

The second mechanism analysis method uses these steps. The baseline equation regresses Y on X to investigate their relationship. In order to examine the impact of X on Y through mechanism M, some literature regresses M on X to confirm that X affects M. Whether this step is used or not is optional; for instance, Alesina & Zhuravskaya (2011) estimating the effect of X on M, Persico et al. (2004) and Holl (2016) use previous literature to support their relationship. Then they regress Y on X and M, and compare the estimated coefficient of X with that in the baseline estimation. If the estimated coefficient of X changes, the mechanism is approved.

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Controls + \varepsilon_1 \tag{5.A-3}$$

$$M = \beta_0 + \beta_1 X + \varepsilon_2 (optional, used\ by\ some\ papers) \tag{5.A-4}$$

$$Y = \sigma_0 + \sigma_1 X + \sigma_2 M + \sigma_3 Controls + +\varepsilon_3 \tag{5.A-5}$$

Alesina & Zhuravskaya (2011) examine the relationship between segregation and the quality of government. They find that more linguistic and ethnic segregation leads to a lower quality of government and trust is a crucial mechanism. They employ three channels, including generalized trust, secession threat and ethnic parties. They regress each channel on X to investigate how X affects each channel. They then regress Y on X and M to identify the changes in the coefficients of X from the baseline model. They find that after adding trust in equation 5.A-5, the coefficient of segmentation changes and becomes insignificant. The coefficient of trust is statistically significant. When adding the other two channels, the coefficient of segmentation does not change further.

Persico et al. (2004) also use this mechanism analysis method with equation 5.A-5 to investigate the effect of teen height on adult wage. They did not use equation 5.A-4 to test the relationship between M and X. They exclude the mechanisms related to the omitted resource variable such as native intelligence, health, etc. They then test the mechanisms related to human capital, including occupation choice, self-esteem, social activities, achievement tests and

years of completed schooling separately with equation 5.A-5. The mechanism is approved if the coefficient of X (teen height) changes after adding the mechanisms to the regression equation. They find that adding social activities changes the coefficient of X (teen height), while some mechanisms such as occupation choice and self-esteem are disproved.

Becker & Woessmann (2009) use a similar mechanism analysis method and further propose a three-stage model to address the endogeneity problem of the independent variable. They investigate the effect of Protestantism on economic outcomes through the mechanism of literacy. With equation 5.A-5, they find that after controlling for M (share of literates), the estimated coefficient of X (Protestantism) on Y (economic outcomes) changes and becomes statistically insignificant.

Apart from the above three steps to investigate the mechanism, they further use bounding analysis and find that Protestantism does not affect economic outcomes independent of the channel of literacy. They then use the 3SLS model with IV for Protestantism:

$$Y = \alpha_3 + \alpha_4 \hat{M} + \alpha_5 Controls_1 + \varepsilon_1 \tag{5.A-6}$$

$$M = \beta_3 + \beta_4 \hat{X} + \beta_5 Controls_2 + \varepsilon_2 \tag{5.A-7}$$

$$X = \sigma_4 + \sigma_5 IV + \sigma_6 Controls_3 + \varepsilon_3 \tag{5.A-8}$$

In the first stage, they regress X (share of Protestantism) on the IV (distance from Wittenberg). In the second stage, they regress M (share of literates) on the predicted share of Protestants (caused by the distance to Wittenberg). In the third stage, they use the variation in literacy related to Protestants to predict the economic outcome. This three-stage model is used to indicate that the part of Protestantism that is caused by the distance from Wittenberg (instrumental variable) has a positive effect on literacy, and the part of literacy due to Protestantism positively affects economic progressiveness.

Holl (2016) investigates the effects of highway access on firm-level TFP of Spanish and Portuguese firms from 1997 to 2007 through the mechanisms of local density. Holl controls for local density as equation 5.A-5. Holl uses two instruments for local density: the 1900 market potential and underground water. The results show that highways improve productivity directly without the local density mechanism after adding the firm fixed effect.

Wurgler (2000) investigates the effect of the financial market (X) on capital allocation efficiency (Y). They first regress capital allocation efficiency (Y) on the financial market variable (X) and find that the financial market plays a significant role in capital allocation efficiency. They further examine different mechanisms by which the financial market facilitates capital allocation efficiency, including stock price synchronicity, SOE and investor right index. They regress capital allocation efficiency (Y) on these channels (M) one by one and all together. They also regress capital allocation efficiency (Y) on channels (M) and the financial development (X) and find the individual mechanism effects.

Method #3: $X \rightarrow M, M \rightarrow Y$

Some literature first explores the relationship between X and Y without mechanisms. In order to examine the impact of X on Y through mechanism M, they regress M on X to confirm the effects of X on M. Then they regress Y on M and other control variables and examine how M affects Y.

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Controls_1 + \varepsilon_1 \tag{5.A-9}$$

$$M = \beta_0 + \beta_1 X + \varepsilon_2 \tag{5.A-10}$$

$$Y = \sigma_0 + \sigma_1 M + \sigma_2 Controls_2 + \varepsilon_3 \tag{5.A-11}$$

Squicciarini (2020) investigates the impact of religion on knowledge and economic development during the Second Industrial Revolution. The mechanism is Catholic primary school education. They explain that the proportion of Catholic schools is higher in districts with a high level of religiosity. A push for Catholic education in more religious locations is negatively associated with industrial employment 10 to 15 years later when schoolchildren enter the labour market.

They regress the mechanism (share and the growth share of Catholic schools) on X (share of refractory clergy), and find that the share of Catholic schools increases more in districts with higher religious devotion. They then regress Y (industrial employment share) on M (share of Catholic schools) in ten-year lags and find that districts with a higher share of Catholics have a lower industrial employment share ten years later when pupils grow up and enter the labour market.

Method #4: Mediation Analysis

Mediation analysis is a way to quantify the direct and indirect effects of the explanatory variable on the response variable. It is not very commonly used in research in economics as it is difficult to address the endogeneity problems of both explanatory variable and mechanism variable. The following regression equations show the three steps for mediation analysis

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 Controls + \varepsilon_1 \tag{5.A-12}$$

$$M = \beta_0 + \beta_1 X + \varepsilon_2 \tag{5.A-13}$$

$$Y = \sigma_0 + \sigma_1 X + \sigma_2 M + \sigma_3 Controls + +\varepsilon_3 \tag{5.A-14}$$

In equation 5.A-12, α_1 represents the total effects of the explanatory variable, which is equal to $\sigma_1 + \beta_1 \sigma_2$. The direct effect of X on Y is captured by σ_1 in equation (5.A-14). The indirect effect of X on Y arises from the effect of X on mediators, which is captured by $\beta_1 \sigma_2$. Thus, the direct effect share is σ_1/α_1; Indirect effect share is $\beta_1 \sigma_2/\alpha_1$ or $1 - direct\ effect\ share$.

Wan & Zhang (2018) use the mediation analysis to investigate the effect of infrastructure on productivity with the indirect effect through the agglomeration channel. Fagereng et al. (2021) research the impacts of parental wealth on child wealth through channels including children's education, income, financial literacy and direct transfers of wealth from parents. They use mediation analysis and find that the effects of parents' wealth on wealth transfer is statistically significant, on children's education it is statistically significant but small, on child financial literacy or income it is small and not statistically significant. Regarding the effect of parental wealth, they conclude that indirect effects explain about 37% of the causal effect. Transfer of wealth is the most important mediator (90% of indirect effects).

Criticism of Mediation analysis This mechanism analysis method has been criticized by some scholars in economic research. First, equation 5.A-14 controls for the mechanism and obtains the estimated coefficient of X, which implies regression equation 5.A-12 is not correctly set. Second, in order to investigate the impact of M on Y, the endogeneity problem of M needs to be addressed. Most studies use this mechanism analysis method without addressing the endogeneity of M because it is difficult for one paper to address the endogeneity problems of both X and M. This can be beyond the workload of one paper and producing two papers might be a more reasonable way to explore the causal effects of X on Y and M on Y. Third, if two or more mechanisms are analysed, it is even harder to address the endogeneity problems of every mechanism.

Method #5: $X * M \rightarrow Y$

It is not very common to use the interaction term of the independent variable and the mechanism to examine the effect of the mechanism. The interaction term method is more suited for heterogeneity analysis. Estimation by groups is also used more for heterogeneity analysis rather than mechanism analysis. These two methods can be explained as the impacts of the explanatory variable on the response variable being conditional on some factors.

Summary of Mechanism Analysis Methods

The mechanism analysis models are summarised as

#1 $X \rightarrow Y$ and $X \rightarrow M$ (means regressing Y on X and then regressing M on X) ;

#2 $X \rightarrow Y$ and $X + M \rightarrow Y$, either with $M \rightarrow Y$ or not, compare the changes in the coefficient of X;

#3 $X \rightarrow M$ and $M \rightarrow Y$;

#4 Mediation analysis;

#5 Regress Y on the interaction term of $X * M$.

Regarding the effect of the highway variable on firm-level productivity, the first way is preferred. Mediation analysis are not adopted since it is hard to address the endogeneity problem of the mediator and explanatory variable. The interaction term between the explanatory variable and another factor is usually used in heterogeneity analysis rather than mechanism analysis. Thus the first two common methods are appropriate for investigating channels for the effects of highways on productivity.

Chapter 6

Conclusion and Discussion

This study has investigated the effects of highway access on the level of within-industry agglomeration, pairwise coagglomeration and firm productivity. Highway GIS routes have been utilized to compute the weighted distance between industries and highway networks in order to establish the highway access variable. The Input-Output adjusted Highway Access metric is designed to account for variations in industry size and transportation volumes. The level of industrial agglomeration is measured by the Ellison and Glaeser index at three geographic levels (county, city and province). Different approaches are studied to create TFP, including the OLS, FE, OP, OP-ACF, LP, LP-ACF, and Wooldridge estimate. Three types of time-variant instruments including the historical routes IV, LCP IV and straight line IV are used to address the endogeneity problem for the highway access variable. The historical route IVs encompass the routes established during the Ming dynasty, the routes established during the Qing dynasty, and a combination of both.

The fixed effect results of the baseline model indicate that highway access has a positive relationship with within-industry agglomeration, pairwise coagglomeration at all geographic levels, and firm productivity. Using IV estimation, the empirical findings confirm that highway access has a beneficial impact on within-industry agglomeration at province, city and county levels. This study's heterogeneity analysis shows that improved highway access generally decreases within-industry agglomeration related to petroleum at the provincial and city levels but increases it at the county level, highlighting the nuanced effects of transportation costs on industry location. It also finds that enhanced highway access boosts downstream in-

dustry agglomeration by improving location flexibility and reducing costs, promoting within-industry agglomeration overall. The study finds that improved highway access, as measured by the input-output adjusted metric, significantly increases within-industry agglomeration by lowering the costs of accessing inputs and outputs from other industries.

However, regarding the coagglomeration results, this study finds that highway access increases coagglomeration at the province and city levels, but reduces it at the county level. At province or city levels, firms can benefit from coagglomeration externalities and select to co-locate together. Highways improve economic efficiency by reducing the requirement to relocate together. However, for the county level, the higher coagglomeration value needs two industries that are large in one county. However, as a county has fewer resources, with the expansion of highways, firms may choose to satisfy agglomerate with those in the same industry first in a county.

Heterogeneity across industries for the effects of highways on coagglomeration is also examined, which includes industry pairs with different input-output linkages, related industry pairs, and industries with a higher proportion of SOEs. The impacts of highway access on coagglomeration at province and city levels are more significant for industries that have stronger input-output links. The coagglomeration levels of industry pairs that are related and belong to the same classification categories are more dependent on highway access. Within the set of industry pairs that are related, the impact of highway access on their coagglomeration is more significant at the provincial and city levels when one industry is upstream and the other is downstream. Furthermore, industries that have a higher proportion of SOEs experience a reduced impact of highway access on coagglomeration.

In terms of the three Marshallian externalities, this study finds that they have positive effects on within-industry agglomeration at three geographical levels after using IV, while for coagglomeration, knowledge spillovers have a significantly positive effect on pairwise coagglomeration. Input-output linkages positively affect coagglomeration at the province level, while the results for labour market pooling are insignificant. Additionally, natural advantages are found in both fostering within-industry agglomeration and coagglomeration.

Regarding the effects on firm productivity, the IV estimation produces reliable findings indicating the beneficial impact of highway access on firm productivity. This thesis investigates four channels for the effects of highway access on productivity, including within-industry agglomeration (firm-level), coagglomeration (firm-level), export and innovation. Firm-level agglomeration is used in the estimation aligning with firm-level productivity and highway access. The firm-level agglomeration is calculated as the logarithm of the number of enterprises located within a neighbourhood with radii of 5, 10, 20, 30, 40, and 50 kilometres. The firm-level agglomeration index considers heterogeneity across firms and addresses the problem of discontinuity in geography, compared with EG industry-level agglomeration.

The analysis of the mechanism reveals that the productivity of firms is influenced by highway access through the channels of firm-level agglomeration. After limiting the sample to new entrants, calculating firm-level agglomeration based on employment, and removing large cities from the sample, the results remain consistent. The IV estimation results for the export channel show that highways encourage exports. The elasticity of exports with respect to highway access is 0.0409. Highway access has a big impact on exports for foreign-invested firms and processing trade firms. The results indicate highways can affect firm productivity through the export channel, while the innovation channel is less significant.

This study has some policy implications. The findings demonstrate that highways significantly enhance both intra-industry and inter-industry agglomeration, suggesting that investment in infrastructure is an effective strategy for promoting industrial concentration and boosting productivity, especially in developing countries. However, as highway networks mature, the marginal benefits of additional highways may diminish, while the costs of maintenance and repair increase. Therefore, while ongoing investment in transportation infrastructure remains important, developing countries should carefully balance the expansion of infrastructure with its economic returns. Industries with high levels of input sharing, particularly downstream sectors, benefit most from highway access due to their sensitivity to transportation costs and reliance on highways for commodity movement.

Additionally, the study demonstrates that having access to highways enhances the productivity of firms through the channel of within-industry agglomeration and coagglomeration. Optimal utilization of industrial zones and special zones can be achieved by efficient access to roads and a well-designed industrial framework. These zones can include both firms in the same industry or firms with upstream-downstream linkages. This strategy aims to enhance firm productivity.

Bibliography

Ackerberg, D. A., Caves, K. & Frazer, G. (2015), 'Identification properties of recent production function estimators', *Econometrica* **83**(6), 2411–2451.

Ackerberg, D., Caves, K. & Frazer, G. (2006), 'Structural identification of production functions'.

Agrawal, A., Galasso, A. & Oettl, A. (2017), 'Roads and innovation', *Review of Economics and Statistics* **99**(3), 417–434.

Aleksandrova, E., Behrens, K. & Kuznetsova, M. (2020), 'Manufacturing (co) agglomeration in a transition country: Evidence from russia', *Journal of Regional Science* **60**(1), 88–128.

Alesina, A. & Zhuravskaya, E. (2011), 'Segregation and the quality of government in a cross section of countries', *American Economic Review* **101**(5), 1872–1911.

Alonso, W. (2013), Location and land use, *in* 'Location and land use', Harvard university press.

Alvarez, R. & López, R. A. (2005), 'Exporting and performance: evidence from chilean plants', *Canadian Journal of Economics/Revue canadienne d'économique* **38**(4), 1384–1400.

Andersson, F., Burgess, S. & Lane, J. I. (2007), 'Cities, matching and the productivity gains of agglomeration', *Journal of Urban Economics* **61**(1), 112–128.

Antonietti, R. & Cainelli, G. (2011), 'The role of spatial agglomeration in a structural model of innovation, productivity and export: a firm-level analysis', *The Annals of Regional Science* **46**(3), 577–600.

Antràs, P., Chor, D., Fally, T. & Hillberry, R. (2012), 'Measuring the upstreamness of production and trade flows', *American Economic Review* **102**(3), 412–16.

Atack, J., Bateman, F., Haines, M. & Margo, R. A. (2010), 'Did railroads induce or follow economic growth?: Urbanization and population growth in the american midwest, 1850–1860', *Social Science History* **34**(2), 171–197.

Au, C.-C. & Henderson, J. V. (2006), 'Are chinese cities too small?', *The Review of Economic Studies* **73**(3), 549–576.

Aw, B. Y., Chung, S. & Roberts, M. J. (2000), 'Productivity and turnover in the export market: micro-level evidence from the republic of korea and taiwan (china)', *The World Bank Economic Review* **14**(1), 65–90.

Bai, C.-E., Du, Y., Tao, Z. & Tong, S. Y. (2004), 'Local protectionism and regional specialization: evidence from china's industries', *Journal of international economics* **63**(2), 397–417.

Balat, J., Casas, C. & Casas, C. (2018), 'Firm productivity and cities: the case of colombia', *Borradores de Economía; No. 1032* .

Baldwin, J. R. & Gu, W. (2003), 'Export-market participation and productivity performance in canadian manufacturing', *Canadian Journal of Economics/Revue canadienne d'économique* **36**(3), 634–657.

Baptista, R. & Swann, P. (1998), 'Do firms in clusters innovate more?', *Research policy* **27**(5), 525–540.

Barrell, R. & Pain, N. (1998), 'Real exchange rate, agglomerations, and irreversibilities: macroeconomic policy and fdi in emu', *Oxford Review of Economic Policy* **14**(3), 152–167.

Barrell, R. & Pain, N. (1999), 'Domestic institutions, agglomerations and foreign direct investment in europe', *European Economic Review* **43**(4-6), 925–934.

Barrios, S., Bertinelli, L. & Strobl, E. (2006), 'Coagglomeration and spillovers', *Regional Science and Urban Economics* **36**(4), 467–481.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A. & Zhang, Q. (2017), 'Roads, railroads, and decentralization of chinese cities', *Review of Economics and Statistics* **99**(3), 435–448.

Baum-Snow, N., Henderson, J. V., Turner, M., Brandt, L. & Zhang, Q. (2015), 'Transport infrastructure, urban growth and market access in china'.

Becker, S. O. & Woessmann, L. (2009), 'Was weber wrong? a human capital theory of protestant economic history', *The quarterly journal of economics* **124**(2), 531–596.

Belleflamme, P., Picard, P. & Thisse, J.-F. (2000), 'An economic theory of regional clusters', *Journal of Urban Economics* **48**(1), 158–184.

Berman, L. & Zhang, W. (2017), Readmem*ingcourier*2016.*txt*, *in*'*V*6*MingDynastyCourierRoutesandStations*', *HarvardDataverse*. **URL:***https://doi.org/10.7910/DVN/SB8ZTM/0RQAGG*

Bernard, A. B. & Jensen, J. B. (1999), 'Exceptional exporter performance: cause, effect, or both?', *Journal of international economics* **47**(1), 1–25.

Bernard, A. B., Jensen, J. B. & Lawrence, R. Z. (1995), 'Exporters, jobs, and wages in us manufacturing: 1976-1987', *Brookings papers on economic activity. Microeconomics* **1995**, 67–119.

Blalock, G. & Gertler, P. J. (2004), 'Learning from exporting revisited in a less developed setting', *Journal of development economics* **75**(2), 397–416.

Blundell, R. & Bond, S. (1998), 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of econometrics* **87**(1), 115–143.

Brandt, L., Van Biesebroeck, J. & Zhang, Y. (2012), 'Creative accounting or creative destruction? firm-level productivity growth in chinese manufacturing', *Journal of development economics* **97**(2), 339–351.

Chaney, T. (2008), 'Distorted gravity: the intensive and extensive margins of international trade', *American Economic Review* **98**(4), 1707–21.

Chen, T., Chen, X., Wang, C. & Xiang, X. (2018), 'Export behavior and firm innovation: New method and evidence', *Economics Letters* **170**, 76–78.

Chen, Z., Zhang, J. & Zi, Y. (2021), 'A cost-benefit analysis of r&d and patents: Firm-level evidence from china', *European Economic Review* **133**, 103633.

Ciccone, A. (2002), 'Agglomeration effects in europe', *European economic review* **46**(2), 213–227.

Ciccone, A. & Hall, R. E. (1996), 'Productivity and the density of economic activity', *The American Economic Review* **86**(1), 54–70.
URL: *http://www.jstor.org/stable/2118255*

Clerides, S. K., Lach, S. & Tybout, J. R. (1998), 'Is learning by exporting important? micro-dynamic evidence from colombia, mexico, and morocco', *The quarterly journal of economics* **113**(3), 903–947.

Connell, J., Kriz, A. & Thorpe, M. (2014), 'Industry clusters: an antidote for knowledge sharing and collaborative innovation?', *Journal of Knowledge Management* .

Coşar, A. K. & Demir, B. (2016), 'Domestic road infrastructure and international trade: Evidence from turkey', *Journal of Development Economics* **118**, 232–244.

Dai, M., Maitra, M. & Yu, M. (2016), 'Unexceptional exporter performance in china? the role of processing trade', *Journal of Development Economics* **121**, 177–189.

Dai, M. & Yu, M. (2013), 'Firm r&d, absorptive capacity and learning by exporting: Firm-level evidence from china', *The World Economy* **36**(9), 1131–1145.

Dai, R., Mookherjee, D., Quan, Y. & Zhang, X. (2021), 'Industrial clusters, networks and resilience to the covid-19 shock in china', *Journal of Economic Behavior & Organization* **183**, 433–455.

De Loecker, J. (2007), 'Do exports generate higher productivity? evidence from slovenia', *Journal of international economics* **73**(1), 69–98.

De Loecker, J. & Syverson, C. (2021), An industrial organization perspective on productivity, *in* 'Handbook of Industrial Organization', Vol. 4, Elsevier, pp. 141–223.

de Vor, F. & de Groot, H. L. (2010), 'Agglomeration externalities and localized employment growth: the performance of industrial sites in amsterdam', *The Annals of Regional Science* **44**(3), 409–431.

Delgado, M. A., Farinas, J. C. & Ruano, S. (2002), 'Firm productivity and export markets: a non-parametric approach', *Journal of international Economics* **57**(2), 397–422.

Ding, S., Jiang, W. & Sun, P. (2016), 'Import competition, dynamic resource allocation and productivity dispersion: micro-level evidence from china', *Oxford Economic Papers* **68**(4), 994–1015.

Ding, S., Sun, P. & Jiang, W. (2016), 'The effect of import competition on firm productivity and innovation: does the distance to technology frontier matter?', *Oxford Bulletin of Economics and Statistics* **78**(2), 197–227.

Ding, S., Sun, P. & Jiang, W. (2019), 'The effect of foreign entry regulation on downstream productivity: Microeconomic evidence from china', *The Scandinavian Journal of Economics* **121**(3), 925–959.

Diodato, D., Neffke, F. & O'Clery, N. (2018), 'Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time', *Journal of Urban Economics* **106**, 1–26.

Dixit, A. K. & Stiglitz, J. E. (1977), 'Monopolistic competition and optimum product diversity', *The American economic review* **67**(3), 297–308.

Dumais, G., Ellison, G. & Glaeser, E. (1997), Geographic concentration as a dynamic process, Working Paper 6270, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w6270*

Dumais, G., Ellison, G. & Glaeser, E. L. (2002), 'Geographic concentration as a dynamic process', *Review of economics and Statistics* **84**(2), 193–204.

Duranton, G. & Overman, H. G. (2005), 'Testing for localization using micro-geographic data', *The Review of Economic Studies* **72**(4), 1077–1106.

Duranton, G. & Overman, H. G. (2008), 'Exploring the detailed location patterns of uk manufacturing industries using microgeographic data', *Journal of Regional Science* **48**(1), 213–243.

Duranton, G. & Puga, D. (2001), 'Nursery cities: urban diversity, process innovation, and the life cycle of products', *American Economic Review* **91**(5), 1454–1477.

Duranton, G. & Turner, M. A. (2012), 'Urban growth and transportation', *Review of Economic Studies* **79**(4), 1407–1440.

Ellison, G. & Glaeser, E. L. (1997), 'Geographic concentration in us manufacturing industries: a dartboard approach', *Journal of political economy* **105**(5), 889–927.

Ellison, G. & Glaeser, E. L. (1999*a*), 'The geographic concentration of industry: does natural advantage explain agglomeration?', *American Economic Review* **89**(2), 311–316.

Ellison, G. & Glaeser, E. L. (1999*b*), 'The geographic concentration of industry: Does natural advantage explain agglomeration?', *American Economic Review* **89**(2), 311–316.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.89.2.311*

Ellison, G., Glaeser, E. L. & Kerr, W. R. (2010), 'What causes industry agglomeration? evidence from coagglomeration patterns', *American Economic Review* **100**(3), 1195–1213.

Faber, B. (2014), 'Trade integration, market size, and industrialization: evidence from china's national trunk highway system', *Review of Economic Studies* **81**(3), 1046–1070.

Fagereng, A., Mogstad, M. & Rønning, M. (2021), 'Why do wealthy parents have wealthy children?', *Journal of Political Economy* **129**(3), 703–756.

Faggio, G., Silva, O. & Strange, W. C. (2017), 'Heterogeneous agglomeration', *Review of Economics and Statistics* **99**(1), 80–94.

Falck, O., Fritsch, M. & Heblich, S. (2014), 'Is industry location persistent over time? evidence from coagglomeration patterns between new and incumbent firms in germany', *Review of Regional Research* **34**(1), 1–21.

Feldman, M. P. (1999), 'The new economics of innovation, spillovers and agglomeration: Areview of empirical studies', *Economics of innovation and new technology* **8**(1-2), 5–25.

Feldman, M. P. & Audretsch, D. B. (1999), 'Innovation in cities:: Science-based diversity, specialization and localized competition', *European economic review* **43**(2), 409–429.

Fujita, M. (1988), 'A monopolistic competition model of spatial agglomeration: Differentiated product approach', *Regional science and urban economics* **18**(1), 87–124.

Fujita, M. & Krugman, P. (2004), The new economic geography: Past, present and the future, *in* 'Fifty years of regional science', Springer, pp. 139–164.

Fujita, M. & Thisse, J.-F. (2002), *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*, Cambridge University Press.

Gabe, T. M. & Abel, J. R. (2016), 'Shared knowledge and the coagglomeration of occupations', *Regional Studies* **50**(8), 1360–1373.

Gallagher, R. M. (2013), 'Shipping costs, information costs, and the sources of industrial coagglomeration', *Journal of Regional Science* **53**(2), 304–331.

Garcia-López, M.-À., Holl, A. & Viladecans-Marsal, E. (2015), 'Suburbanization and highways in spain when the romans and the bourbons still shape its cities', *Journal of Urban Economics* **85**, 52–67.

Ge, Y. (2009), 'Globalization and industry agglomeration in china', *World development* **37**(3), 550–559.

Gibbons, S., Lyytikäinen, T., Overman, H. G. & Sanchis-Guarner, R. (2019), 'New road infrastructure: the effects on firms', *Journal of Urban Economics* **110**, 35–50.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A. & Shleifer, A. (1992), 'Growth in cities', *Journal of political economy* **100**(6), 1126–1152.

Graham, D. J. (2007), 'Agglomeration, productivity and transport investment', *Journal of transport economics and policy (JTEP)* **41**(3), 317–343.

Greenaway, D., Girma, S. & Kneller, R. (2008), Does exporting lead to better performance? a microeconometric analysis of matched firms, *in* 'Does Exporting Lead to Better Performance? A Microeconometric Analysis of Matched Firms: Greenaway, David| uGirma, Sourafel| uKneller, Richard', [Sl]: SSRN.

Greenaway, D. & Kneller, R. (2004), 'Exporting and productivity in the united kingdom', *Oxford Review of Economic Policy* **20**(3), 358–371.

Greenaway, D. & Yu, Z. (2004), 'Firm-level interactions between exporting and productivity: Industry-specific evidence', *Review of World Economics* **140**, 376–392.

Guimaraes, P., Figueiredo, O. & Woodward, D. (2000), 'Agglomeration and the location of foreign direct investment in portugal', *Journal of Urban Economics* **47**(1), 115–135.

Guo, L. & Yang, Z. (2019), 'Relationship between shipping accessibility and maritime transport demand: the case of mainland china', *Networks and Spatial Economics* **19**(1), 149–175.

He, C., Guo, Q. & Ye, X. (2016), 'Geographical agglomeration and co-agglomeration of exporters and nonexporters in china', *GeoJournal* **81**(6), 947–964.

He, C. & Pan, F. (2010), 'Economic transition, dynamic externalities and city-industry growth in china', *Urban Studies* **47**(1), 121–144.

He, C., Ye, X. & Wang, J. (2012), 'Industrial agglomeration and exporting in china: What is the link?', *Regional Science Policy & Practice* **4**(3), 317–333.

He, Z.-L., Tong, T. W., Zhang, Y. & He, W. (2018), 'A database linking chinese patents to china's census firms', *Scientific data* **5**(1), 1–16.

Head, K., Ries, J. & Swenson, D. (1995), 'Agglomeration benefits and location choice: Evidence from japanese manufacturing investments in the united states', *Journal of international economics* **38**(3-4), 223–247.

Helsley, R. W. & Strange, W. C. (2014), 'Coagglomeration, clusters, and the scale and composition of cities', *Journal of Political Economy* **122**(5), 1064–1093.
URL: *<Go to ISI>://WOS:000343240300004*

Henderson, J. V. (2003), 'Marshall's scale economies', *Journal of urban economics* **53**(1), 1–28.

Henderson, V., Kuncoro, A. & Turner, M. (1995), 'Industrial development in cities', *Journal of political economy* **103**(5), 1067–1090.

Holl, A. (2004), 'Transport infrastructure, agglomeration economies, and firm birth: empirical evidence from portugal', *Journal of Regional Science* **44**(4), 693–712.

Holl, A. (2012), 'Market potential and firm-level productivity in spain', *Journal of Economic Geography* **12**(6), 1191–1215.

Holl, A. (2016), 'Highways and productivity in manufacturing firms', *Journal of Urban Economics* **93**, 131–151.

Holmes, T. J. (1999), 'Localization of industry and vertical disintegration', *Review of Economics and Statistics* **81**(2), 314–325.

Holmes, T. J. & Stevens, J. J. (2002), 'Geographic concentration and establishment scale', *Review of Economics and Statistics* **84**(4), 682–690.

Hoover, E. M. (1937), *Location theory and the shoe and leather industries*, Harvard University Press, Cambridge.

Hornung, E. (2015), 'Railroads and growth in prussia', *Journal of the European Economic Association* **13**(4), 699–736.

Howard, E., Newman, C. & Tarp, F. (2016), 'Measuring industry coagglomeration and identifying the driving forces', *Journal of Economic Geography* **16**(5), 1055–1078.

Hsu, W.-T., Lu, Y., Luo, X. & Zhu, L. (2019), 'Does foreign direct investment lead to industrial agglomeration?'.

Hu, C., Xu, Z. & Yashiro, N. (2015), 'Agglomeration and productivity in china: Firm level evidence', *China Economic Review* **33**, 50–66.

Isgut, A. (2001), 'What's different about exporters? evidence from colombian manufacturing', *Journal of development studies* **37**(5), 57–82.

Ito, B., Xu, Z. & Yashiro, N. (2015), 'Does agglomeration promote internationalization of chinese firms?', *China Economic Review* **34**, 109–121.

Jacobs, W., Koster, H. R. & van Oort, F. (2014), 'Co-agglomeration of knowledge-intensive business services and multinational enterprises', *Journal of Economic Geography* **14**(2), 443–475.

Jiwattanakulpaisarn, P., Noland, R. B. & Graham, D. J. (2012), 'Marginal productivity of expanding highway capacity', *Journal of Transport Economics and Policy (JTEP)* **46**(3), 333–347.

Jofre-Monseny, J., Marín-López, R. & Viladecans-Marsal, E. (2011), 'The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms', *Journal of Urban Economics* **70**(2-3), 61–74.

Jofre-Monseny, J., Marín-López, R. & Viladecans-Marsal, E. (2014), 'The determinants of localization and urbanization economies: Evidence from the location of new firms in spain', *Journal of Regional Science* **54**(2), 313–337.

Ke, S. (2010), 'Agglomeration, productivity, and spatial spillovers across chinese cities', *The annals of regional science* **45**(1), 157–179.

Ke, S., He, M. & Yuan, C. (2014), 'Synergy and co-agglomeration of producer services and manufacturing: A panel data analysis of chinese cities', *Regional Studies* **48**(11), 1829–1841.

Kim, S. (1995), 'Expansion of markets and the geographic distribution of economic activities: the trends in us regional manufacturing structure, 1860–1987', *The Quarterly Journal of Economics* **110**(4), 881–908.

Kraay, A. (1999), 'Exports and economic performance: Evidence from a panel of chinese enterprises', *Revue d'Economie du Developpement* **1**(2), 183–207.

Krugman, P. (1991), 'Increasing returns and economic geography', *Journal of political economy* **99**(3), 483–499.

Krugman, P. (1998), 'What's new about the new economic geography?', *Oxford review of economic policy* **14**(2), 7–17.

Krugman, P. R. (1979), 'Increasing returns, monopolistic competition, and international trade', *Journal of international Economics* **9**(4), 469–479.

Lall, S., Deichmann, U. & Shalizi, Z. (1999), *Agglomeration economies and productivity in Indian industry*, The World Bank.

Lanaspa, L., Sanz-Gracia, F. & Vera-Cabello, M. (2016), 'The (strong) interdependence between intermediate producer services' attributes and manufacturing location', *Economic Modelling* **57**, 1–12.

Levchenko, A. A., Ranciere, R. & Thoenig, M. (2009), 'Growth and risk at the industry level: The real effects of financial liberalization', *Journal of Development Economics* **89**(2), 210–222.

Levinsohn, J. & Petrin, A. (2003), 'Estimating production functions using inputs to control for unobservables', *The review of economic studies* **70**(2), 317–341.

Li, D., Lu, Y. & Wu, M. (2012), 'Industrial agglomeration and firm size: Evidence from china', *Regional Science and Urban Economics* **42**(1-2), 135–143.

Li, T., Han, D., Feng, S. & Liang, L. (2019), 'Can industrial co-agglomeration between producer services and manufacturing reduce carbon intensity in china?', *Sustainability* **11**(15), 4024.

Lin, F. (2015), 'Learning by exporting effect in china revisited: An instrumental approach', *China Economic Review* **36**, 1–13.

Lin, H.-L., Li, H.-Y. & Yang, C.-H. (2011), 'Agglomeration and productivity: Firm-level evidence from china's textile industry', *China Economic Review* **22**(3), 313–329.

Liu, D., Sheng, L. & Yu, M. (2023), 'Highways and firms' exports: Evidence from china', *Review of International Economics* **31**(2), 413–443.

Loecker, J. D. (2013), 'Detecting learning by exporting', *American Economic Journal: Microeconomics* **5**(3), 1–21.

Lu, J. & Tao, Z. (2009), 'Trends and determinants of china's industrial agglomeration', *Journal of urban economics* **65**(2), 167–180.

Ma, Y., Tang, H. & Zhang, Y. (2014), 'Factor intensity, product switching, and productivity: Evidence from chinese exporters', *Journal of International Economics* **92**(2), 349–362.

Marschak, J. & Andrews, W. H. (1944), 'Random simultaneous equations and the theory of production', *Econometrica, Journal of the Econometric Society* pp. 143–205.

Marshall, A. (1890), *Principles of Economics, by Alfred Marshall*, Macmillan and Company.

Martín-Barroso, D., Núñez-Serrano, J. A. & Velázquez, F. J. (2015), 'The effect of accessibility on productivity in spanish manufacturing firms', *Journal of Regional Science* **55**(5), 708–735.

Martin, P., Mayer, T. & Mayneris, F. (2011), 'Spatial concentration and plant-level productivity in france', *Journal of urban Economics* **69**(2), 182–195.

Martincus, C. V. & Blyde, J. (2013), 'Shaky roads and trembling exports: Assessing the trade effects of domestic infrastructure using a natural experiment', *Journal of International Economics* **90**(1), 148–161.

Martincus, C. V., Carballo, J. & Cusolito, A. (2017), 'Roads, exports and employment: Evidence from a developing country', *Journal of Development Economics* **125**, 21–39.

Melitz, M. J. (2003), 'The impact of trade on intra-industry reallocations and aggregate industry productivity', *econometrica* **71**(6), 1695–1725.

Mukim, M. (2015), 'Coagglomeration of formal and informal industry: evidence from india', *Journal of Economic Geography* **15**(2), 329–351.

Na, K.-Y., Han, C. & Yoon, C.-H. (2013), 'Network effect of transportation infrastructure: a dynamic panel evidence', *The Annals of Regional Science* **50**, 265–274.

Nakamura, R. (1985), 'Agglomeration economies in urban manufacturing industries: a case of japanese cities', *Journal of Urban economics* **17**(1), 108–124.

Neffke, F., Henning, M., Boschma, R., Lundquist, K.-J. & Olander, L.-O. (2011), 'The dynamics of agglomeration externalities along the life cycle of industries', *Regional studies* **45**(1), 49–65.

Olley, S. & Pakes, A. (1992), 'The dynamics of productivity in the telecommunications equipment industry'.

Ozbay, K., Ozmen-Ertekin, D. & Berechman, J. (2007), 'Contribution of transportation investments to county output', *Transport Policy* **14**(4), 317–329.

O'Sullivan, A. & Strange, W. C. (2018), 'The emergence of coagglomeration', *Journal of Economic Geography* **18**(2), 293–317.

Persico, N., Postlewaite, A. & Silverman, D. (2004), 'The effect of adolescent experience on labor market outcomes: The case of height', *Journal of political Economy* **112**(5), 1019–1053.

Porter, M. E. (1990), 'The competitive advantage of nations', *Harvard business review* **68**(2), 73–93.

Porter, M. E. (2000), 'Location, competition, and economic development: Local clusters in a global economy', *Economic development quarterly* **14**(1), 15–34.

Porter, M. E. et al. (1998), *Clusters and the new economics of competition*, Vol. 76, Harvard Business Review Boston.

Redding, S. J. & Turner, M. A. (2015), 'Transportation costs and the spatial organization of economic activity', *Handbook of regional and urban economics* **5**, 1339–1398.

Romer, P. M. (1986), 'Increasing returns and long-run growth', *Journal of political economy* **94**(5), 1002–1037.

Roos, M. W. (2005), 'How important is geography for agglomeration?', *Journal of Economic Geography* **5**(5), 605–620.

Rosenthal, S. S. & Strange, W. C. (2001), 'The determinants of agglomeration', *Journal of urban economics* **50**(2), 191–229.

Ruan, J. & Zhang, X. (2009), 'Finance and cluster-based industrial development in china', *Economic Development and Cultural Change* **58**(1), 143–164.

Skinner, G. W., Yue, Z. & Henderson, M. (2008), 'ChinaW–Cities, County Seats and Yamen Units (1820 - 1893)'.
URL: *https://doi.org/10.7910/DVN/JCT5NE*

Song, Y., Lee, K., Anderson, W. P. & Lakshmanan, T. (2012), 'Industrial agglomeration and transport accessibility in metropolitan seoul', *Journal of Geographical Systems* **14**(3), 299–318.

Squicciarini, M. P. (2020), 'Devotion and development: religiosity, education, and economic progress in nineteenth-century france', *American Economic Review* **110**(11), 3454–3491.

Tokunaga, S. & Kageyama, M. (2008), 'Impacts of agglomeration and co-agglomeration effects on production in the japanese manufacturing industry: using flexible translog production function', *Studies in Regional Science* **38**(2), 331–337.

Tong, T., Yu, T.-H. E., Cho, S.-H., Jensen, K. & Ugarte, D. D. L. T. (2013), 'Evaluating the spatial spillover effects of transportation infrastructure on agricultural output across the united states', *Journal of Transport Geography* **30**, 47–55.

Van Biesebroeck, J. (2005), 'Exporting raises productivity in sub-saharan african manufacturing firms', *Journal of International economics* **67**(2), 373–391.

Van Biesebroeck, J. (2014), Productivity, exporting and financial constraints of chinese smes, Technical report, IDB Working Paper Series.

Venables, A. J. (2007), 'Evaluating urban transport improvements: cost–benefit analysis in the presence of agglomeration and income taxation', *Journal of Transport Economics and Policy (JTEP)* **41**(2), 173–188.

Wagner, J. (2007), 'Exports and productivity: A survey of the evidence from firm-level data', *World economy* **30**(1), 60–82.

Wan, G. & Zhang, Y. (2018), 'The direct and indirect effects of infrastructure on firm productivity: Evidence from chinese manufacturing', *China Economic Review* **49**, 143–153.

Wang, J., Sun, F., Lv, K. & Wang, L. (2022), 'Industrial agglomeration and firm energy intensity: How important is spatial proximity?', *Energy Economics* **112**, 106155.

Weber, A. (1909), 'Über den standort der industrien (on the location of industries)'.

Wen, M. (2004), 'Relocation and agglomeration of chinese industry', *Journal of development economics* **73**(1), 329–347.

Wooldridge, J. M. (2009), 'On estimating firm-level production functions using proxy variables to control for unobservables', *Economics letters* **104**(3), 112–114.

Wurgler, J. (2000), 'Financial markets and the allocation of capital', *Journal of financial economics* **58**(1-2), 187–214.

Yang, Y. & Mallick, S. (2010), 'Export premium, self-selection and learning-by-exporting: Evidence from chinese matched firms', *The World Economy* **33**(10), 1218–1240.

Yu, M. (2015), 'Processing trade, tariff reductions and firm productivity: Evidence from chinese firms', *The Economic Journal* **125**(585), 943–988.

Yu, N., De Jong, M., Storm, S. & Mi, J. (2013), 'Spatial spillover effects of transport infrastructure: evidence from chinese regions', *Journal of Transport Geography* **28**, 56–66.

Zeng, S., Liu, Z. & Sun, D. (2022), 'How do highways enable firm productivity? the role of innovation', *IEEE Transactions on Engineering Management* .

Zhang, H. (2015), 'How does agglomeration promote the product innovation of chinese firms?', *China Economic Review* **35**, 105–120.