Effect of High-Speed Rail on City Tourism Revenue in China: A Perspective on Spatial Connectivity[*]

Lingyun Xiao

Thesis Advisor: Professor Jón Steinsson

December 1st, 2019

Abstract

This paper investigates the effect of High-Speed-Rail (HSR) network on the tourism revenue in 193 prefecture-level cities in China. This study defines a spatial measure of connectivity to take into account of several crucial factors of a network that previous studies often failed to address: the total number of cities connected to each city, the adjacency of any two cities in the network, and the relative centrality of each city. By using a panel data from 2007 to 2017, the empirical result from the multi-state Difference-in-Difference method suggests that an increase in spatial connectivity leads to a significant positive growth in tourism revenue, while addressing vital issues of endogeneity and spatial heterogeneity. Also, the result shows that a temporary effect of connectivity on tourism revenue exists, but lower-income and less-populated cities benefit more from the HSR expansion than the more developed ones.

Keywords: High-Speed Rail, Connectivity, Tourism, Transportation

[*] Special thanks to my thesis advisor Professor Jón Steinsson for his great help. I would also like to thank Professor Yuriy Gorodnichenko, Professor James L. Powell and Professor David Romer for their advice on this project. All errors are my own.

1 Introduction

Not surprisingly, following the booming construction of High-Speed Rail (HSR) network in China in the past few years, an abundance of literature has examined the societal benefits and economic impacts the network has induced. By the end of 2017, the HSR network has reached 15,000 miles, and China Railway corporation forecasts that an additional 2,000 miles of new railroads will be under construction in each coming year. The longest HSR network ever built in the world received vast media coverage and increasing interest from the academia. It might seem natural that HSR reduces the travel time by a significant amount, so tourists would be more likely to take the train and visit a city that is otherwise too time-consuming. In fact, Wang has shown that as more cities joining the network enlarges the tourist market, reduces the travel cost and redistributes local economies[14].

However, what makes the topic more nuanced is the influence from a) the quality (whether a city is connected or not); b) the quantity (how many cities are connected to each city); c) the centrality (the position of a city in a network) and d) neighborhood effect. From the research up to date, a plethora of studies have only focused on a), while only a scant of studies also considered the others. Also, Several major challenges to the analysis need to be carefully addressed. Given the non-randomness of the chosen cities connected to HSR, the endogeneity issue complicates the research findings. Moreover, different cities can experience different effect on tourism based on its size or past development, so spatial heterogeneity is also a relevant factor to consider. Finally, the construction of HSR could also have a time lagged effect. My paper aims to define a new measure of connectivity and incorporate various methods to address those issues simultaneously.

This study begins with first introducing the spatial measure of connectivity that addresses: a) the total number of cities connected to a given city; b) the shortest distance between two cities via HSR network; c) neighborhood effect that diminishes with distance and d) spatial heterogeneity. Then, I use fixed effects regression on the panel data to find the relationship between the spatial connectivity and the tourism revenue, control for several economic variables, and get the main result. Next, I address the endogeneity issue by creating an instrumental variable and run the regression again. Moreover, I add a lagged connectivity measure to investigate whether the construction in one year has a temporary effect on the tourism revenue in the year after. My main conclusion agrees with the literature that higher HSR network connectivity does have a positive impact on city-level tourism revenue, and the construction of HSR has a stronger temporary effect on tourism for small/medium cities than larger ones.

2 Literature Review

2.1 HSR system in China

Originated in Japan in 1962, HSR has become a fast, seamless and popular way of transportation. The first Chinese HSR was first implemented in 2008, and have expanded to reach more than 150 cities by 2017. A vivid representation of the major network in China is by Four Horizontals and Four Verticals, while the whole network is demonstrated by Eight Horizontals and Eight Verticals including branch lines. It is not a surprise that various research studies have been devoted to the analysis of HSR network in China. As for the focus on Chinese High-Speed Railroad, the majority of in-depth research suggests that the expansion of HSR network has induced significant increase in city-level GDP[4], home price difference[18], tourist arrivals[10] and transformation of tourist market[14] via the reduction of travel costs and travel time. It seems like most of the research agrees upon the economic benefit induced by railroads, yet a closed look at the studies have generated the following two issues that are particularly challenging on identification.

One of the biggest obstacles that researchers have faced when dealing with the construction of HSR is arguably the endogeneity issue. Because the assignment and order of new HSR stations is often prioritized based on political reasons and unlikely random, treating the evaluation as a simple random experiment would be an oversimplification. Some studies completely ignore the endogeneity concern at all [3][16][17], so while their results have some indication of whether HSR network benefits local economy, their conclusion might be qualified if an IV approach is used. Several Instrumental Variable methods have been explored in the past literature, including dynamic panel analysis [9][12], using the shipment of military troops in the past as an exogenous shock [4][18], retrospecting to the historical network as an instrument [8], or constructing a least-cost path network in a clever way [6].

All the studies have run tests to validate the instrument, report a high first stage F-stat and show its exogeneity. An interesting phenomena to notice is that all the papers that use IV approach above also use DID methods, concerning only about whether a city is linked to the network or not in a given year. This finding is not surprising because most of the instruments are time-invariant, which is relatively easier to find compared to a time-variant instrument because the time frame can add significant complexity in the exclusion restriction assumption. One of the biggest challenges in my multi-state Difference-in-Difference approach is to find a suitable instrument that varies in time, and in the Results section below, I'll discuss why I come up with this new instrument and the reasons behind it.

Another issue that arises in previous work is the heterogeneity and spatial difference of any HSR network. In other words, simply treating each individual city on the network as identical entries would be a biased estimate because cities differ in sizes, population and past tourism development. The vast majority of the result agrees that smaller cities in China with low tourism development in the past experienced more significant increment of tourist inflows and tourism revenue [8][10][17][18]. In an quasi-experiment conducted on the assignment of HSR stations, researchers use generalized DID method to validate that new HSR does lead to significant and positive effect on the tourism arrivals. Interestingly, when various cities are grouped by different sizes, the treatment effect is more consequential in medium/small cities compared to large cities [8]. Li, Yang and Cui have examined both domestic and international tourist arrivals simultaneously. Apart from the overall positive impact of increased High-Speed Rail connectivity on local tourism, the result also suggests that lower-income and less-populated cities have a relatively higher increase in their tourist inflows for both domestic and international tourists [10]. Moreover, studies on the capital region of China suggest that smaller cities will receive more benefit on tourist inflows from HSR system than larger ones[17]. Some researchers have turned their eyes to look into the relationship between HSR and home price difference. Zheng and Kahn also find a rising living quality of urban population that stems from the construction of the network, but more importantly, they find that second-tier and third-tier cities underwent more advancement than first-tier big cities [18].

Spatial heterogeneity also appears when researchers look into province-level tourism growth. in A case study of Wuhan-Guangzhou HSR concludes that some provinces benefit from the HSR connections more than others, and the regional effect could even be negligible [16]. Even though none of the provinces had implanted any intercity HSR line before, which could potentially alleviate the rising tourism popularity, Hubei Province underwent close to zero stimulation on tourism development, while Hunan and Guangzhou Province experienced positive stimulation. Although the author does not take into account of the endogeneity issue, the result still raises a red flag to the plain treatment of stations on the network as the same entries. Another interesting topic a few studies have addressed is the spillover effect because the implementation of a HSR station in a large city might spill over to neighboring regions. However, an evaluation of city-level panel data in China shows that increasing centrality of a city in HSR network leads to vital economic growth of regional GDP, and the effect mainly attributes to local effect instead of spillovers [4].

My paper addresses the spatial difference concern in two ways: a) I redefine connectivity to add in diminishing impact with distance, so I could differentiate the relative weight of each city in the network, generating a spatial heterogeneity; b) I group cities based on their sizes and economic advancement in the study. This approach result in a discussion of whether small or medium scale cities receive a more significant boost in economy, which remains a controversy in the previous literature.

2.2 Worldwide HSR Systems

Given the prevalence of HSR around the world, Chinese HSR has not been the sole research interest in the past years. Several studies have examined the economic impacts of HSR in Spain [7][11] and France [13], where networks of High-Speed Train system have been built. Those studies have led to mixed results given the nuanced impacts of building HSR on various industries. Some research suggests that the introduction of HSR network in Madrid-Barcelona region in Spain will cause dramatic negative effect on the air transportation industry. The study forecasts that the share of air traffic will likely plummet to less than half of the original market proportion [11]. Therefore, the introduction of the railroad might alter the transportation industry landscape, and the outcome might not be desirable. A case study on Madrid-Toledo High Speed Line includes the design of the railway stations and the accessibility to urban population as the main factors that could potentially affect the interest of the riders. Although a large number of tourists tend to take HSL on weekends, a relatively low dependence on HSL occurs on weekdays, pointing out that temporal heterogeneity in the HSL design also correlates with tourism [7]. On the contrary, another case study on intermediate cities in France supports the construction of HSR because the implementation stimulates modernization at national, regional and local level. Nevertheless, the argument is qualified by the side effect that the HSR also accelerates polarization towards only the metropolises, which might widen the income and capital gap between small and large cities [13].

Not only has previous research examined the effect of High-Speed Railway System in one single country, several studies have been conducted by using the rich panel data on a multitude of countries with varied wealth level and population sizes. A continent-wise analysis via dynamic panel analysis and gravity model points out that transportation infrastructure could shift the attractiveness of destinations and international tourist arrivals. Moreover, given the abundance of available information of 28 countries, they use comparative analysis to differentiate between high-income and low-income countries as well as small and large countries based on GDP, CPI and Population [9]. The investigation shows the importance of transportation capital on tourism inflows, and international tourists put great value on the safety and efficiency of transportation infrastructure no matter what destinations they arrive at. In summary, the research does show some spatial heterogeneity across countries, so it should not be taken granted that the construction of HSR will necessarily boost the economy in all scales. A more careful nuanced approach needs to play the role in the detailed examination.

2.3 Network and Connectivity

A particular feature that stands out in the analysis of HSR development is the nature of the network. By treating each station as a node and each railroad between two cities as en edge, any HSR network can be reduced to an undirected graph, whose topological features could be a crucial component to the construction of the network. Nevertheless, even though plenty of research has been devoted to the analysis of HSR system in China, a surprisingly low amount of work has considered or even mentioned the graphical representation of the network. An overwhelming amount of literature has used the traditional Diff-in-Diff method and looks into the economic impact induced by the shift from no connection to some connection [2][3][4][8][18]. Admittedly, DID method is useful in this setting because the construction of a railway could be easily represented by the "switch": 0 if a city is not connected and 1 if the city is connected. However, not only did Bertrand, Duflo and Mullainathan warn that DID method could understate the standard deviation of the estimators by a significant amount [1], but also the dummy variable does not take any of the mathematical properties of the network into account.

A few studies have acknowledged this special feature of the railroad network and created their own definition of connectivity. For example, Chong, Qin and Chen use a comprehensive

National Railroad Schedule to construct a frequency table and measures the connectivity of one city as the normalized sum of the number of direct HSR train departures and arrivals on a daily basis [4]. The immediate benefit of utilizing the train frequency compared to using a generic connect/disconnect switch is to capture the degree centrality of a city in a network. In other words, if more high-speed trains are taking off from a city, the indication is that this city plays a more vital role in the network. A central city that connects multiple networks like Wuhan and Shanghai will enjoy a higher connectivity, while a small town with some major network passed through might not receive much boost in economic benefit induced by the increased connectivity. Therefore, their strategy successfully addresses the concern of spatial difference mentioned earlier.

Nevertheless, a limitation to their study is that only trains that directly depart from a station are counted, so the trains that pass through a station are not present in their analysis of connectivity. Secondly, even if two cities are connected via the HSR train, taking Harbin and Guangzhou as an example, that are thousands of miles away, the actual flow of tourist might be negligible. On the other hand, although inter-city trains like Beijing-Tianjin HSR has a high frequency of around twenty trains on a daily basis, the tourism induced by additional frequency might be diminishing fast. In my attempt to redefine connectivity, I value the distance between two cities more than the frequency of trains, and my approach consider both the direct connection as well as pass-by connection to address those concerns.

3 Data

In this section, I will present the data as well as the source for all subsequent analysis. In order to examine HSR network in China and acquire information on prefecture-level cities, I use the following major data sets: a) The construction and expansion of Chinese High-Speed Rail System at city-level; b) The schedule of HSR trains; c) Yearly time series of tourism revenue and population at city-level; d) Yearly updates to the Chinese AAAAA tourist attraction list; e) Yearly data on air transportation and the number of aircraft for departure and arrival. All of the yearly city-level time series range from 2007 to 2017, generating a 11-year panel data with 193 prefecture-level cities.

3.1 Main Data Sets

The first step to look into the HSR network is to see whether a city is connected to the network in a given year. Thus, by using the data provided by State Railway Corporation, the official administration for railroad construction and operation, I can build up a list of dummy variables, indicating when a particular city joins the network. However, because my research interest does not limit itself to a qualitative analysis based off the simple connection, I need additional information on the quantity of cities connected to a given city, the distance between two cities in the network, and whether a certain High-Speed Line originates from or passes by a city. Although the usage of extra information has raised the difficulty of finding a reliable source of data, I have combined data found on China's National Railway Administration website, www.12306.cn; an open database of National Railway Schedule found on ip138.com as well as all the paper copies of the schedule. The biggest challenge in the past research studies is to find information on cities that are not departure or arrival stations. In other words, the online database provides information mostly on the two ends of a line, but not on most of the connecting stations between them. Therefore, I read through the books published by National Railway Administration to search for intermittent stations in a network. This distinct feature of railroad connections is notably different from airlines connections, which could be treated as an edge between two nodes, instead of a series of edges connecting various nodes. Then, I can define connectivity with all the data acquired.

My outcome variable of interest is the tourism revenue at city level, which can be found on ceic.com, another open source online data set that summarize the travel expenditures and revenues at city, province and national level. However, given the data availability, not all the cites have data from 2007 to 2017, so the total number of 193 prefecture-level cities is used in this study. The vast majority of previous work uses data before 2013 because the comprehensive data on almost all cities is given. However, since the range network has expanded significantly from 2014 to 2017, a trade-off between time horizon and city horizon is necessary to capture the overall impact of HSR construction. Although the exclusion of some cities might produce some bias, yet the 193 cities expand to 28 provinces and the missing cities are at random regardless of size, economic growth or the endogeneous factors.

Other than the independent and dependent variables of interest, the data on various control variables is also vital. The information on city average residential population comes from *China City Statistical Yearbook*, a comprehensive summary of various economic indicators at city-level and it is available from 2007 to 2017. A nuance that is seldom discussed is the difference between end-of-the-year population and year-average population. Given the huge impact of Chunyun, a spring travel migration that occurs around Spring Festival that is particular in China, an averaged population would be more accurate. Furthermore, using residential population rather than household registration is also a treatment to acknowledge the fact that a significant number of people leave their hometown for major metropolitan areas, which often results in a non-negligible difference between the registered population and the residential population.

An indicator of tourism popularity, the list of 5A (AAAAA) tourist spots is published by China National Tourism Administration from 2008 to 2017. This list is updated on a yearly basis. In each year's report, several tourism spots in specific prefecture-level cities could be added or deleted from previous year's report. Therefore, it is a time-variant measure of a city's popularity to tourists, which also influences tourism revenue. Finally, the data on the number of aircraft taking off or landing on an airport associated with the city is published by Civil Aviation Administration of China each year. A nuance here is that a few mega cities have multiple airports, and some small cities share one airport that is built on their boundaries, so the data on those particular cities is combined to address that concern.

3.2 Defining Connectivity

As stated in the literature review, previous studies in economics on High-Speed Rail Network have not paid attention to the *network* perspective and rather focuses on the presence of a city in the network. More papers on mathematics and urban planning have delved into the structural design of network connectivity and topology, such as betweenness and centrality [5] or natural connectivity and global efficiency [15]. In those studies, the distance between two nodes, or the adjacency, plays a big role in the definition of connectivity. We can treat the conventional DID method in the HSR analysis that defines connectivity as follows:

$$connectivity_{numeric,it} = \begin{cases} 1 & \text{if city } i \text{ is connected to some cities in year } t \\ 0 & \text{if city } i \text{ is not connected to the network in year } t \end{cases} \tag{1}$$

Using this simplification for DID method is reasonable, concerning only the change in economic outcome before and after the implementation of the network. Nevertheless, the binary representation no longer makes sense when I begin taking the *quantity* of cities linked by the network into account, which I have emphasized throughout the paper. Therefore, I can define a generic representation of *numeric connectivity* as follows:

$$connectivity_{numeric,it} = \begin{cases} 1 & \text{if city } i \text{ is connected to 1-10 cities in year } t \\ 2 & \text{if city } i \text{ is connected to 10-50 cities in year } t \\ 3 & \text{if city } i \text{ is connected to more than 50 cities in year } t \\ 0 & \text{if city } i \text{ is not connected to the network in year } t \end{cases} \tag{2}$$

Equation (1) not only captures whether there is a link between a city and the network, but also grasps the total number of cities that a particular city can reach to. As suggested intuitively, the connectivity remains zero if there is no connection and becomes non-zero if any connection is constructed. Therefore, the numeric connectivity fully catches the essence of the dummy variable approach. By the notion of being "connected", one may wrongfully presume that as long as city i is in the network, there must be a connection between city i and any other city j in the same network. Nevertheless, this is not the case in most settings. If city i is not a central city of connection, then for most of the case, only a limited number of cities can be reached, and the linkage to other cities depends on various factors.

The first factor is distance. It is mostly the case that no direct connection exists at all between two cities that are far away from another, and an indirect connection requires transferring at a major station. As a side note, the case of transfer is excluded in my study because: a) there is no data set available on transfer so far, and b) if city i indirectly connects to city k via transferring in city j, then it's likely to be double counted for separate analysis of two individual linkages from city i to j and j to k. Because of unnecessary complexity brought upon by it, transfer lines are not included. Secondly, the existence of a direct connection depends on the specific railroad corridor one city lies on. The Chinese HSR network mainly consists of Eight Verticals and Eight Horizontals as its main grid system. It is unlikely that a particular line crosses multiple corridors without transferring at a transfer station. Therefore, instead of treating *all* stations in the network as reachable, I use specific time schedule of each line to include only the *reachable* city via direct link.

Based on those factors for consideration, I twist the generic version of connectivity by setting up thresholds. Based on the data set, most of newly built lines consist of fewer than 10 cities, which is the reason I set my first threshold. As more new HSR lines join the network, it becomes possible to reach more destinations, and 50 is an estimate of the connectivity of a non-central city, which means only one railroad passes through the city. If a city is connected to more than 50 cities, then it is likely that the city is a central city connected to multiple lines, or the city lies on one of the major Four Verticals and Four Horizontals corridors. Thus, those cities have the highest connectivity overall. However, there are several issues that arise with the above numeric definition. While the choice of 10 cities and 50 cities captures some heterogeneity among cities, using a specific number to set up a threshold seems too dictating. Cities on edge of the threshold would have a sudden jump even if only a few new stations are connected, while the tourism revenue might not have a significant response to the shift in connectivity.

Apart from the quantity of connected cities, we come back to the key factor that influences connectivity and also creates city-level heterogeneity, which is the *distance* between two cities. As suggested by Wang, although travel time and cost is significantly reduced with the expansion of the HSR system, the effect it induces decreases with the distance [14]. Intuitively, if two cities are thousands of miles apart, say Urumqi and Shanghai, the fact that these two cities are connected via HSR might actually have negligible impact on the tourist flow from one city to the other given the huge geographical disparity. Tourists from Shanghai might plan to visit Urumqi regardless of whether there exists an HSR network because of the time cost. Thus, one assumption I make is that the connectivity decreases with distance, which is not picked up by the previous definitions.

To wrap up my discussion on how to define connectivity, my study focuses on three vital components: a) whether a city is connected to HSR or not; b) how many cities are linked to a city via HSR; c) longer distance between cities leads to decreasing connectivity. Based on those assumptions and concerns, I propose the final definition of *spatial connectivity* as follows:

$$connectivity_{spatial,it} = \sqrt{\sum_{j \neq i} \frac{1}{d(i,j)}} \tag{3}$$

where $d(i,j)$ represents the *minimum distance* between city i and city j on the undirected graph. In other words, if we represent each station as a *node* and the connection between them as an *edge*, $d(i,j)$ is the minimum number of edges available to travel from node i to node j directly via HSR. I assume the functional form to be square root instead of just a linear sum, so the concavity captures the diminishing impact on tourism revenue as more cities are connected to city i.

Equation (3) successfully addresses all three concerns for connectivity. Besides, rather than using geographical distance between two cities, a spectral measure of distance on the simple graph is more reasonable, since a small geographical distance between cities does not necessarily mean a direct HSR connection. By using the available data on all the cities that are passed by HSR lines, I am able to construct such spectral measure of distance. I will use spatial connectivity in my following empirical analysis to explore how different forms of network linkage impacts tourism.

Table 1 provides information on all variables of interest:

	Mean	SD	Min	Max
Tourism	28537.12	46576.7	51	512240
Numeric_Connectivity	.8097033	1.222457	0	3
Spatial_Connectivity	.76405	1.111509	0	3.793415
Attraction	.2783797	.4483069	0	1
Airport	27082.11	77178.63	0	760360
Population	447.6179	293.6058	47.1	2506.4
N	2123			

Table 1: Summary Statistics of Variables for 193 Cities from 2007 to 2017

4 Methods

4.1 Base Model

After providing new definitions of railroad connectivity in a complex network, I can use a baseline model to investigate the economic impact of HSR connectivity. The dependent variable of interest is the logarithm of tourism revenue in city i in year t. Given the nature of the variable and the availability of the full panel data set, I use a Difference-in-Difference method as follows:

$$\ln(TourismRevenue)_{it} = \beta Connectivity_{it} + \gamma_1 Attraction_{it} + \gamma_2 \ln(Airport)_{it} + \gamma_3 \ln(Population)_{it} + FE_i + FE_t + \epsilon_{it} \quad (4)$$

where $Connectivity_{it}$ represents *spatial connectivity* I have defined as the key explanatory variable I am interested in. I have also included both city fixed effect and year fixed effect because they are crucial components of the identification. Several control variables are included to address potential omitted variable bias: $Attraction_{it}$ is the dummy variable that indicates whether city i has an AAAAA tourist attraction in year t, $\ln(Airport)_{it}$ is the logarithm of the total number aircraft that departed from and arrived at an airport associated with city i in year t, and $\ln(Pop)_{it}$ is the logarithm of yearly average residential population of city i in year t.

The coefficient of interest is β, which indicates percentage impact that an unit increase in connectivity has on tourism revenue. Based on the definition, the increase in connectivity is inversely proportional to the minimum distance between cities, so the increment diminishes when cities are further apart. Therefore, β captures the average of the diminishing effect of connectivity when more cities join the network.

I separate all 193 cities into two groups: The treatment group consists of cities that had some HSR connection in a given year, while the control group is made of cities with no HSR connection. The city fixed effect captures fixed difference between cities, and the year fixed effect controls for tourism revenue trends that are invariant across time. By using the Diff-in-Diff method, I am able to identify the effect brought upon by the construction of High-Speed Rail without including

4.2 Lagged Effects of Connectivity

According to previous studies, lagged effects of HSR connections as well as the attractiveness of a city might have a non-trivial influence on tourist arrivals and revenue [2][10]. It is reasonable to investigate lagged effects of the construction of High-Speed Rails because its manufacture in one year might not only result in increasing tourist flow in that year *only*, but also cause subsequent impact on the decisions of tourists to travel in later years. This effect is not captured by the base model, which only finds out what the permanent effect on tourism is. The appearance of HSR offers a less time-consuming choice to travel to places that would not be on their wished list without the option, but they do not have to take that option to use it right away. However, on the other hand, the boost in tourism for a given city in the year of construction might also be more substantial than the impact in the later years. The immediate psychological stimulus of the new bullet trains operated with high speed might induce more traffic in the same year that the line is connected, but less so as people grow accustomed to the expanding network. Since both explanations could influence the regression, it is necessary to investigate the lagged effect of HSR connectivity of a city. The identification is specified as follows where I have included the lagged connectivity for one year given the time scale and data availability:

$$\begin{aligned}\ln(TourismRevenue)_{it} = \beta_1 Connectivity_{it} + \beta_2 Connectivity_{it-1} + \gamma_1 Attraction_{it} \\ + \gamma_2 \ln(Airport)_{it} + \gamma_3 \ln(Population)_{it} + FE_i + FE_t + \epsilon_{it} \quad (5)\end{aligned}$$

To interpret the coefficients of lagged effects, I consider two separate cases: a) The Dynamic Marginal Effect; b) The Cumulative Effect. In this first-order lag model, β_1 captures the immediate impact of spatial connectivity on tourism, since both belongs to the same time horizon. On the contrary, β_2 measures the dynamic marginal effect of spatial connectivity at one lag. In other words, β_2 assesses the effect of a *temporary* change in connectivity on the level of tourism, in contrast to the *permanent* change described by β_1. If we add β_1 to β_2, the sum measures the *cumulative* effect at one lag. Therefore, those regression coefficients inform us about the timing and magnitude of a shift in spatial connectivity. A temporary change of connectivity leads to a temporary change in tourism revenue, and this change will disappear after two periods of time. By contrast, a permanent change induced by connectivity will *not* vanish over time.

4.3 Addressing Endogeneity

Although some studies suggest that the assignment of HSR railroad stations can be treated as an quasi-experiment [8], the expansion of the network is still not a true randomized trial. Therefore, it is still possible that endogeneity of the main explanatory variable might affect the result. Therefore, I use an IV (Instrumental Variable) approach to address this issue. Based on the data set I have obtained, I can construct a dummy variable that indicates whether or not city i is linked to the *capital* of the province it is located at time t. One main goal of the construction of the network is to link the capitals of each province, regardless of the cities that the link passes through [6]. Therefore, the main reason that a city is connected to its provincial capital is probably not related to developing tourism, but is simply based on its geography because the location of the city is between two capitals. This assumption satisfies the exclusion restriction, since this specific connection does not have a direct influence on tourism revenue.

One might argue that the linkage to the provincial capital could have an indirect effect on local tourism. Capitals have higher connectivity than other cities, so higher tourist inflow into the capital could trigger neighborhood effect and spread into other cities. However, previous studies have used spatial econometric model to show that spillover effect is negligible compared to local effect [4]. Also, even if more tourists arrive at the capital, the most efficient way to travel to another city is through the channel of taking the HSR train, which is exactly the explanatory variable. Given the distance within a province, travelling by air is too costly, and driving on highway is too time-consuming. If a tourist chooses to take a normal train instead of HSR, then the slow train not only makes a stop at prefecture-level cities, but also county-level cities, which vastly increases the time cost. Therefore, using connectivity to the capital within the same province as an instrument, I am able to reduce the potential effect of endogeneity that biases the result.

5 Results and Discussion

5.1 Result using Spatial Connectivity

In this section, I will present the main result using the definition of spatial connectivity defined earlier. Table 2 summarizes the outcome from the regression specified in Equation (4). Overall, regardless of the control methods I use, an increase in High-Speed Rail spatial connectivity leads to a significant positive increase in tourism revenue. All models use city fixed effect to capture the time-invariant characteristics of individual cities. For example, the presence of Karst topography in a city is one of the biggest challenges to HSR construction, but cities like Guilin in Guangxi Province actually attracts tourists because of the existence of this special geological nature. If no entity fixed effect is included, the topography of a city would mingle with the causality of connectivity and tourism revenue. Similarly, all models also include year fixed effect to control for city-invariant shocks across time such as inflation or national economy recession.

The first column does not include IV on all 193 cities. The baseline conclusion is that a 1 unit increase in the spatial connectivity would lead to the increment of 9 percent of tourism revenue with a p-value less than 0.01. The positive impact on tourism seems in line with most of the previous literature[2][8][10][18]. Not surprisingly, a city with popular tourism spots and busier air traffic also generates more on tourism revenue. An interesting thing to notice is that the population does not have a significant impact on the tourism. A possible reason is that even if a city is more populated than others, it might as well be the case that a greater outflow of people will mitigate the local effect. As stated in the Data section, I take the average residential population to exclude the case of a massive shift of population at Spring Festival. Therefore, even though Beijing is one of the most populated cities in China with many tourist attractions, an average resident in Beijing might choose to take the HSR train and travel outside the city where he or she has not visited before.

To minimize complications from endogeneity, I have included IV in all other regressions, and all the large first-stage F-statistics validates the strength of the instrument. Thus, a city's connection to its provincial capital does not correlate with the error terms which could cause tourism to rise through other channels than the increased HSR connectivity. By comparing the result from Column 1 and Column 2, I have noticed that the impact of spatial connectivity shrinks from 9 percent to 7 percent yet still being significant at 1 percent level, and the standard deviation is almost the same. A similar result can be seen with control variables. Without a significant shift in the coefficient, the result echoes the conclusion drawn by Hou, in which the construction of HSR stations could be treated as a quasi-experiment with a certain extent of randomness.

Also, given the nature of a panel data, the independence between observations might not be satisfied when common traits are shared by several entities and errors are correlated for a group of cities. Therefore, the clustering of standard errors is essential to capture the correlations of unobservables within group and independence across groups, which reinforces the robustness of the result. In my case, I cluster standard errors at *city* level individually. The first reason is that the total number of cities is 193. There exists possible downward bias of standard errors when the number of clusters is small, but 193 is considered large enough to avoid such bias. Furthermore, since the data ranges for 11 years, it is likely that the model errors are correlated across time for a single city, while being uncorrelated across cities. If we fail to cluster them, the result could show deceptive small standard errors, which give rise to false narrow confidence intervals.

We *could* consider clustering at *province* level. For instance, based on the data set, all

Variable	**1**	**2**	**3**	**4**
SpatialConnectivity	0.0911***	0.0725***	0.0593***	0.0814***
	(0.0103)	(0.00960)	(0.0124)	(0.0140)
Attraction	0.528***	0.345***	0.284***	0.392***
	(0.0509)	(0.0522)	(0.0668)	(0.0779)
log(Airport)	0.103***	0.0723***	0.0858***	0.0674***
	(0.0101)	(0.00926)	(0.0196)	(0.0108)
log(Population)	-0.118	-0.166*	-0.324	-0.155
	(0.126)	(0.0985)	(0.345)	(0.162)
Constant	9.561***	10.17***	11.45***	9.963***
	(0.748)	(0.584)	(2.280)	(0.897)
IV	No	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes
First Stage F-stat	-	18.95	25.37	26.91
Observations	2,123	2,123	803	1,320
Number of Cities	193	193	73	120

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: The Impact of HSR Connectivity on Tourism Revenue

cities in Liaoning Province experienced a plummet in tourism revenue in 2015 with no obvious reasons, while the whole Hunan Province underwent a skyrocketed tourism growth in the same year. It is unlikely some shock at national level that generates such vast heterogeneity, but some provincial-level shock that affects every city regardless of its HSR connection. However, only 28 provinces exist in China, and it would be risking a significant downward bias of standard errors to trade off the identification of common correlations within a province. By comparison, I choose to cluster standard errors on city level.

Column 3 and 4 analyze spatial heterogeneity that is brought upon in various research [8][10][13]. I use the Small and Medium City Development Index designed by *China Society of Urban Economy Media and Small Economic Development Committee* to categorize 193 cities into two groups: 73 cities as large cities and 120 cities as small/medium cities. The index takes not only the population size into account, but also factors such as capital, real estate and health level as indicators of a city's economic development. The result suggests that while the tourism in both large and small/medium cities benefit from higher HSR

Variable	**5**	**6**	**7**
SpatialConnectivity	0.0585***	0.0590***	0.0607***
	(0.0104)	(0.0136)	(0.0149)
L1SpatialConnectivity	0.0369***	0.0177	0.0438***
	(0.0110)	(0.0108)	(0.0162)
Attraction	0.261***	0.259***	0.255***
	(0.0545)	(0.0636)	(0.0874)
log(Airport)	0.0557***	0.0668***	0.0514***
	(0.00975)	(0.0187)	(0.0118)
log(Population)	-0.162*	-0.571*	-0.188
	(0.0912)	(0.306)	(0.153)
Constant	10.31***	13.23***	10.31***
	(0.543)	(2.034)	(0.849)
First Stage F-statistic	27.83	30.74	23.23
Observations	1,737	648	1,089
Number of Cities	193	72	121
Group of Cities	All	Large	Small/Medium

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: The Impact of Lagged Connectivity of Tourism Revenue

connectivity, lower-income cities with less advancement in past development enjoy higher tourism stimulus via HSR. The situation is reversed for air traffic, which is plausible since smaller cities have more limited access via air travel and smaller airports than larger cities. As for the high-speed train, the infrastructure paired with the construction of rail stations is less costly, and the continuous graphical property of HSR network allows for more flexible travels and reachable cities compared to the discrete air network, where one flight can only reach one or two destinations.

5.2 Exploring Lagged Effect

In this section, I will examine whether lagged effect is important in the analysis of connectivity and tourism. By "lagged effect", the term captures the spatial connectivity of city i in year $t-1$ relative to the current year t. One reason that introducing lagged terms might generate imprecise result is due to multicollinearity, where the set of lagged independent variables could be predicted using a linear equation. Therefore, I only include one lagged year to analyze the immediate lagged effect on tourism revenue.

Again, I run the regression on all cities first and then categorize cities into large and small/medium entities respectively. The result in column 5 shows that increasing spatial connectivity by a unit leads to a positive 5.85 percent permanent increase in tourism revenue in the same year, which is lower than the estimate without the lagged effect. On the contrary, the lagged connectivity has a 3.69 percent temporary increase in the revenue. When I group the cities into large and small/medium ones, the comparison between the results shows that the two groups share a relatively similar permanent effect. However, the temporary effect of High-Speed Rail is significantly smaller for large cities compared to small/medium cities.

One explanation to the spatial disparity is that the increment of connectivity to High-Speed Rail leads to similar increase in tourism permanently regardless of the characteristics of a city, such as the size or prior tourism development. However, for the small or medium cities, High-Speed Rail connection serves as an essential transportation carriage for travelers. In contrast, large cities are equipped with more developed public transportation systems, so High-Speed Rails are among one of the potential ways to travel, but not a necessary component. Therefore, it is reasonable that the temporary effect of High-Speed Rail is higher in smaller cities because those cities become reachable to many travelers for the first time. The shift from no connectivity to some connectivity induces tourism boost. On the other hand, large cities are already reachable with or without High-Speed Rails, so the temporary effect would be negligible. Thus, to conclude, spatial connectivity itself induces a similar permanent change to tourism revenue, but the temporary change is more significant for cities with smaller size.

6 Conclusion

One crucial thing that most of the previous investigation on the impact on tourism by the expanding High-Speed Rail System in China has not acknowledged for is the special property of a railroad network. A significant amount of work has delved into the Diff-in-Diff method and only addresses whether a city is connected to the network or not. However, not only does the connection itself that matters, but the quantity of total reachable cities via HSR, the minimum distance between cities, and the centrality of a city relative to the network can also affect prefecture-level city tourism revenue. I look into the detailed railway schedule and obtain data on all the stations that lie on a HSR lane. Then, by converting the Chinese HSR network into a simple graph with nodes and edges, I can use the minimum number of edges between two nodes as a proxy of the distance between two cities. Only direct connections are counted because transfers could result in double counting and unnecessary complications.

By defining spatial connectivity, I find that connectivity is positively related to tourism revenue, generating a 7.25 percent boost in the revenue. In order to address challenges to the analysis, I use an Instrumental Variable approach to solve endogeneity issue, include both time and city fixed effects to account for omitted variable bias, cluster standard error at city level to reduce error correlations across cities, group cities by large and small/medium measures to compare and contrast the effects, and finally include lagged connectivity to investigate the permanent and temporary effect on tourism. The result shows that spatial connectivity has a significant positive impact on tourism revenue regardless of the methods I use, which is in line with the majority of work done on Chinese HSR network. The temporary effect of HSR on tourism is larger for low-income and underdeveloped cities, while the permanent effect is about the same for both groups, which results in an overall disparity of connectivity effects on tourism. The result provides some insights into the future design of HSR construction, where new lanes need not be simply connecting the capitals, but branch out to connect more small/medium scale cities because of the more significant impact on local tourism.

Future research could be done to elaborate on my definition of connectivity because I have only incorporated the minimum distance between cities, but not the frequency of high-speed trains, or other means of travel. By using intercity distance, quantity of connected stops and train frequency and creating a more comprehensive measure of connectivity, we can capture most of the nuances of the network system that is not discussed in previous work. Also, we can look into the lagged effect in particular and analyze through which channel do small and medium cities gain more on tourism via the connection of HSR. Since the High-Speed Rail system in China has only been operating for eleven years, with more available data in the future, we can further investigate the neighborhood effect and analyze the potential channels for the spillovers.

References

[1] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How Much Should We Trust Differences-in-Differences Estimates? Technical Report w8841, National Bureau of Economic Research, Cambridge, MA, March 2002.

[2] Zhou Bo and Li Ningqiao. The impact of high-speed trains on regional tourism economies: Empirical evidence from China. *Tourism Economics*, 24(2):187–203, March 2018.

[3] Zhenhua Chen and Kingsley E. Haynes. Impact of high-speed rail on international tourism demand in China. *Applied Economics Letters*, 22(1):57–60, January 2015.

[4] Zhaohui Chong, Chenglin Qin, and Zhenhua Chen. Estimating the economic benefits of high-speed rail in China: A new perspective from the connectivity improvement. *Journal of Transport and Land Use*, 12(1), April 2019.

[5] Daniel DeLaurentis, En-Pei Han, and Tatsuya Kotegawa. Network-Theoretic Approach for Analyzing Connectivity in Air Transportation Networks. *Journal of Aircraft*, 45(5):1669–1679, September 2008.

[6] Benjamin Faber. Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System. *The Review of Economic Studies*, 81(3):1046–1070, July 2014.

[7] B. Guirao and F. Soler. Impacts of the new high speed rail services on small tourist cities: the case of Toledo (Spain). In *Sustainability V*, pages 465–473, Skiathos, Greece, August 2008.

[8] Xinshuo Hou. High-speed railway and city tourism in china: A quasi-experimental study on hsr operation. *Sustainability*, 11:1512, 03 2019.

[9] Jameel Khadaroo and Boopen Seetanah. The role of transport infrastructure in international tourism development: A gravity model approach. *Tourism Management*, 29(5):831–840, October 2008.

[10] Leona S.Z. Li, Fiona X. Yang, and Chuantao Cui. High-speed rail and tourism in China: An urban agglomeration perspective. *International Journal of Tourism Research*, 21(1):45–60, January 2019.

[11] Andrés López-Pita and Francesc Robusté. Impact of High-Speed Lines in Relation to Very High Frequency Air Services. *Journal of Public Transportation*, 8(2):17–35, May 2005.

[12] Tiago Neves Sequeira and Paulo Maçãs Nunes. Does tourism influence economic growth? A dynamic panel data approach. *Applied Economics*, 40(18):2431–2441, September 2008.

[13] José M. Ureña, Philippe Menerault, and Maddi Garmendia. The high-speed rail challenge for big intermediate cities: A national, regional and local perspective. *Cities*, 26(5):266–279, October 2009.

[14] Xin Wang, Songshan Huang, Tongqian Zou, and Hui Yan. Effects of the high speed rail network on China's regional tourism development. *Tourism Management Perspectives*, 1:34–38, January 2012.

[15] Zhongzhi Xu, Qingpeng Zhang, Dingjun Chen, and Yuxin He. Characterizing the Connectivity of Railway Networks. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2019.

[16] York Qi Yan, Hanqin Qiu Zhang, and Ben Haobin Ye. Assessing the Impacts of the High-Speed Train on Tourism Demand in China. *Tourism Economics*, 20(1):157–169, February 2014.

[17] Ping Yin, Francesca Pagliara, and Alan Wilson. How Does High-Speed Rail Affect Tourism? A Case Study of the Capital Region of China. *Sustainability*, 11(2):472, January 2019.

[18] S. Zheng and M. E. Kahn. China's bullet trains facilitate market integration and mitigate the cost of megacity growth. *Proceedings of the National Academy of Sciences*, 110(14):E1248–E1253, April 2013.