ISSN 2042-2695

CENTRE for ECONOMIC PERFORMANCE

CEP Discussion Paper No 1244

October 2013

Trade Integration, Market Size and Industrialization: Evidence from China's National Trunk Highway System

Benjamin Faber

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract

Large scale transport infrastructure investments connect both large metropolitan centers of production as well as small peripheral regions. Are the resulting trade cost reductions a force for the diffusion of industrial and total economic activity to peripheral regions, or do they reinforce the concentration of production in space? This paper exploits China's National Trunk Highway System as a large scale natural experiment to contribute to our understanding of this question. The network was designed to connect provincial capitals and cities with an urban population above 500,000. As a side effect, a large number of small peripheral counties were connected to large metropolitan city regions. To address non-random route placements on the way between targeted city nodes, I propose an instrumental variable strategy based on the construction of least cost path spanning tree networks. The estimation results suggest that network connections led to a reduction in GDP growth among no n-targeted peripheral counties due to reduced industrial output growth. Additional estimation results present evidence that appears consistent with the existence of core-periphery effects of trade integration as found in increasing returns trade theory and economic geography.

Keywords: Trade integration, industrialization, road infrastructure
JEL Classifications: F12; F15; O18; R12

This paper was produced as part of the Centre's Globalisation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements

I am grateful to Lawrence Crissman at the ACASIAN Data Center at Griffith University for his support with the GIS data. The paper has greatly benefited from conversations with Robin Burgess, Dave Donaldson, Esther Duflo, Alexander Lembcke, Thierry Mayer, Guy Michaels, Henry Overman, Nancy Qian, Steve Pischke, Steve Redding, Frederic Robert-Nicoud, Thomas Sampson, Matti Sarvimaki, and Daniel Sturm, as well as the comments of four anonymous referees and the editor. This research was supported by the UK Economic and Social Research Council.

Ben Faber, Department of Economics, University of California Berkeley; Email: benfaber@econ.berkeley.edu.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

A large share of world trade takes place between regions within countries.[1] In this context, transport infrastructure investments have been a prominent policy tool that directly affects the degree of within-country trade integration.[2] These policies frequently combine national efficiency with regional equity objectives under the presumption that falling trade costs promote both national growth as well as the diffusion of economic activity to peripheral regions.[3]

Large scale transport investments almost inevitably connect both large metropolitan centers of production as well as small peripheral regions. This is especially the case in developing countries where spatial disparities are particularly pronounced (Kanbur and Venables, 2005). Are the resulting trade cost reductions between large and small markets a force for the diffusion of industrial and total economic activity to peripheral regions, or do they reinforce the concentration of production in space? Despite widespread policy interest in this question, our existing empirical knowledge is limited. The growing body of empirical literature on the evaluation of transport infrastructure has so far paid little attention to the role of asymmetric market sizes in trade integration, and the policy question of how peripheral regions are affected when connected to large metropolitan agglomerations.[4]

This paper exploits China's National Trunk Highway System (NTHS) as a large scale natural experiment to contribute to our understanding of this question. The NTHS policy objective was to connect provincial capitals, cities with an urban population above 500,000, and border crossings on a single expressway network (World Bank, 2007b). While the targeted metropolitan centers represented less than 1.5% of China's land area and only 14% of its population, they accounted for half of the country's non-agricultural production before the bulk of the network was built in 1997. The average pre-existing difference in market sizes between non-targeted peripheral counties and targeted metropolitan centers was 1:24 in terms of county GDP. The NTHS thus provides an empirical setting where a large number of relatively small peripheral counties where connected to China's major centers of economic activity as a by-product of the network's policy objective.

I exploit this policy setting to empirically estimate the economic consequences of NTHS network connections among the non-targeted peripheral counties. While peripheral counties were not explicitly targeted by the policy, it would nevertheless be a strong *ex ante* assumption that route placements on the way between targeted city centers were randomly assigned. In particular, both the NTHS planning process and descriptive statistics suggest

[1] For evidence using microdata on US plant transactions see Hilberry and Hummels (2008).

[2] Transport infrastructure has been the second most important spending category in World Bank lending over 2001-06, of which 73% were spent on highways and roads (www.worldbank.org).

[3] For example, from the World Bank's Transport Business Strategy 2008-2012 (2008, pp. 03): "*One of the best ways to promote rural development is to ensure good accessibility to growing and competitive urban markets.*" For a discussion of the combination of efficiency with regional equity objectives in the context of European Structural Fund spending on transport see Puga (2002).

[4] See the discussion of the related literature further below in this section.

that planners favored politically important and economically prosperous peripheral counties on the way between targeted destinations. To address these concerns I propose an instrumental variable (IV) strategy based on hypothetical least cost path spanning tree networks. These correspond to the question which routes planners would have built if the sole policy objective had been to connect all targeted city nodes subject to construction cost minimization. In a first step, I use remote sensing data on land cover and elevation in combination with Dijkstra's (1959) optimal route algorithm to compute least costly construction paths between any bilateral pair of targeted nodes. I then use these bilateral cost parameters in combination with Kruskal's (1956) minimum spanning tree algorithm to identify the subset of routes that connect all targeted nodes to minimize network construction costs. I also construct a more extensive, but less precise straight line spanning tree network that is subject to a different trade-off between route precision and the number of captured bilateral connections.

The baseline identifying assumption is that county location along an all-China least cost spanning tree network affects changes of county level economic outcomes only through NTHS highway connections, conditional on province fixed effects, distance to the nearest targeted city node, and controls for pre-existing political and economic characteristics. To assess the validity of the exclusion restriction, I report how baseline IV point estimates are affected by the inclusion of county controls, and test the heterogeneity of the connection effects across pre-existing county characteristics. In addition, the relatively recent nature of the NTHS also allows me to test for network connection effects on identical county samples both before and after the network was built as a placebo falsification test.

The estimation results suggest that NTHS connections have led to a reduction in local GDP growth among peripheral counties on the way between targeted metropolitan centers relative to non-connected peripheral counties. This effect appears to be mainly driven by a reduction in industrial output growth. These results are confirmed on local government revenue growth instead of production data. Furthermore, network connections do not appear to have significant effects on county populations.

After documenting the reduced form connection effects, the analysis proceeds to further investigate the channels at work. I consider three main channels that could be driving the observed effects: 1) Peripheral connections to metropolitan centers of production lead to core-periphery effects of trade integration as found in increasing returns trade theory and economic geography (Helpman and Krugman, 1985; Krugman, 1991; Fujita et al., 1999; Baldwin et al., 2003); 2) Peripheral connections lead to a process of urbanization and industrial decentralization to neighboring peripheral areas (Baum-Snow et al., 2012); and 3) Peripheral connections lead to deindustrialization due to locational fundamentals/comparative disadvantages in industrial production relative to metropolitan regions, while total production declines in the presence of barriers to the reallocation of factors (Goldberg and Pavcnik, 2007; Topalova, 2010) or due to capital outflows (Redding, 2012).

Motivated by the testable implications of the different mechanisms, I present a series of

additional estimation results. The main findings can be summarized as follows: i) NTHS connections do not appear to have significant effects on changes in urban population or urbanization among connected peripheral counties relative to non-connected ones; ii) The estimated adverse effects on production and government revenue growth do not appear to be driven by a process of decentralization of production to surrounding areas in the neighborhood of connected peripheral counties; iii) Peripheral counties with larger pre-existing market sizes and higher pre-existing trade costs to the metropolitan nodes are significantly less adversely affected by NTHS connections; and iv) Conditioning on pre-existing market sizes and trade costs, county differences in sectoral specialization, shares of skilled labor, urban status, prefecture capital status or GDP per capita do not appear to significantly affect the heterogeneity of the NTHS connection effect across peripheral counties. This additional qualitative evidence is consistent with a simple core-periphery model of integration between large and small markets, while it would be difficult to reconcile with a mechanism of urbanization and decentralization or endowment based specialization among peripheral regions on the NTHS network.

In conclusion, the paper establishes the following main novel insights. First, it presents empirical evidence in support of the hypothesis that falling trade costs between large and small markets can lead to reduced industrial and total output growth among peripheral regions, rather than diffusing production in space. Second, the additional estimation results present evidence that appears consistent with the existence core-periphery effects of trade integration as found in increasing returns to scale trade theory and economic geography. These findings provide empirical support for existing theoretically motivated policy discussions in the trade literature (Fujita et al., 1999; Baldwin et al., 2003; Combes et al., 2008), and serve to emphasize the importance of potentially unintended general equilibrium consequences when evaluating and planning large scale transport infrastructure policies.

This research is related to a growing empirical literature on the evaluation of transport infrastructure. Recent contributions have studied the economic effects on suburbanization (Baum-Snow, 2007), skill premia in local labor markets (Michaels, 2008), long term GDP effects (Banerjee et al., 2012), gains from trade (Donaldson, 2013), urban form (Baum-Snow et al., 2012), and city growth (Duranton and Turner, 2012). Relative to the existing literature, this paper draws attention to a different question of policy interest. In particular, I study China's NTHS policy as a large scale natural experiment to learn about the role of asymmetric market sizes in trade integration, and to provide empirical evidence on the question of how relatively small peripheral market places are affected by transport connections to large metropolitan centers of production.

The empirical strategy of this paper is related to recent empirical work on transport infrastructure in China (Banerjee et al., 2012; Baum-Snow et al., 2012; Baum-Snow and Turner, 2012). It is most closely related to Banerjee et al. (2012) who use straight line connections between nearest-neighbor pairs of Treaty Ports and historical cities in China to predict the construction of railway lines in the late 19th and early 20th century. Relative to

existing work, this paper exploits the network design of the NTHS policy and its targeting of a subset of major city nodes in order to propose an instrumental variable strategy that is based on minimum spanning tree networks. In addition to instrumenting for route placements on any given bilateral connection, such as by straight lines in Banerjee et al. (2012) or by applying Dijkstra's algorithm on land cover and elevation data as in this paper, the spanning tree approach allows to instrument for the choice of bilateral route connections covered by the network subject to construction cost minimization. This strategy can provide a useful empirical tool in the evaluation of a variety of different infrastructure policies whose design is based on targeting subsets of centers on national or regional grids, such as transportation networks, utility grids, electricity grids or telecommunication networks.

The remainder of the paper is structured as follows. Section 2 describes the policy background and data. Section 3 presents the empirical strategy. Section 4 presents estimation results. Section 5 presents additional estimation results to further investigate the channels at work. Section 6 concludes.

2 Background and Data

2.1 China's National Trunk Highway System

In 1992, the Chinese State Council approved the construction of the "7-5" network, consisting of seven horizontal and five vertical axes, under the National Trunk Highway Development Program (Asian Infrastructure Monthly, 1995; World Bank, 2007b) (see Figure 1). The NTHS was constructed at an estimated cost of US$ 120 billion over a 15-year period until the end of 2007, spanning approximately 35,000 km of high speed four-lane highways (Li and Shum, 2001; Asian Development Bank, 2007; World Bank, 2007a).

Its stated objectives were to connect all provincial capitals and cities with an urban registered population above 500,000 on a single expressway network, and to construct routes between targeted centers and the border in border provinces as part of the Asian Highway Network. NTHS routes are four-lane limited access toll ways. The common speed limit is 120 km/h, and a common minimum speed limit is 70km/h. Road quality, congestion, and driving speed of the modern expressways are in clear contrast to pre-existing national highways (speed limit 80-100 km/h) and provincial highways (speed limit 70 km/h) that can also be subject to road tolls.

The network was originally earmarked for completion by 2020, but was completed ahead of schedule by the end of 2007. Planners at the Chinese Ministry of Communications divide the construction into a "kick-off" phase between 1992-1997, and "rapid development" between 1998-2007 (World Bank, 2007a). The reason behind the acceleration of construction efforts in 1998 is that highway construction became part of the government's stimulus spending after the Asian financial crisis (Asian Development Bank, 2007).

To finance the great majority of NTHS routes, the central government encouraged

province and county level governments to raise funds by borrowing against future toll revenues. Roughly 70% was financed from province and county level debt, and 10-15% was contributed by the central government. Private sector participation was also encouraged with up to 5% of financing stemming from domestic and foreign investors (Asian Development Bank, 2007). Construction was undertaken almost entirely by Chinese state owned enterprises, part of which were assigned directly to particular localities, part of which were participating in contract auctions.[5] Given the progress in the construction of the NTHS ahead of plan, the State Council approved an even more ambitious follow-up blue print for highway construction in 2004. The so called "7-9-18" system has the stated objective to connect all cities with an urban registered population of more than 200,000. It is scheduled to be completed by 2020.

2.2 Data

This section describes the data and variables used in the estimations. A more detailed description can be found in the Online Appendix. Geo-referenced administrative boundary data for the year 1999 were obtained from the ACASIAN Data Center at Griffith University in Brisbane, Australia. These data provide a county-level geographical information system (GIS) dividing the surface of China into 2341 county level administrative units, 349 prefectures, and 33 provinces. Chinese administrative units at the county level are subdivided into county level cities (shi), counties (xian), and urban wards of prefecture level cities (shixiaqu).

County level socioeconomic records are taken from Provincial Statistical Yearbooks for the years 1990, 1997 and 2006, as well as the 1990 Chinese population census. The statistical yearbook records for 1997 and 2006 were obtained from the University of Michigan's China Data Center, and the 1990 census data as well as statistical yearbook data for 1990 were obtained from the China in Time and Space (CITAS) project at the University of Washington. The Provincial Statistical Yearbook series report county level GDP broken up into agriculture, industry, and services gross value added, as well as local government revenues and registered county populations. The 1990 Population Census provides county level data on population, education, and employment shares by sector.

These sources result in a database of 1748 historically consistent geo-referenced county units that have non-missing reporting values in the Provincial Statistical Yearbooks of 1997 and 2006 (75% of Chinese administrative units). Close to the entirety of this county sample (1706 of 1748) also report socioeconomic records in the 1990 Population Census, and 1238 of the 1748 report local government revenues in the CITAS Provincial Statistical Yearbooks for the year 1990.[6] Table 1 presents a set of descriptive statistics, and the Online Appendix describes the data sources and processing in more detail.

[5]Until the "Measures on Tenders and Bids for Contracts for Construction Projects" came into effect in May 2003, competitive bidding, was recommended but not mandatory (World Bank, 2007a).

[6]Only a fraction of the reporting counties in 1997 and 2006 report production data in the Provincial Statistical Yearbooks for 1990.

Geo-referenced NTHS highway routes were obtained from the ACASIAN Data Center. NTHS routes were digitized on the basis of a collection of high resolution road atlas sources published between 1998 and 2007. These atlas sources made it possible to classify NTHS segments into three categories that coincide with the main construction phases described by the Ministry of Communications: opened to traffic before mid-1997 (10% of NTHS), opened to traffic between mid-1997 and end of 2003 (81% of NTHS), and opened to traffic after 2003 (9% of NTHS).[7] A list of the atlas publications as well as a more detailed description of the data processing and NTHS classifications is given in the Online Appendix. Finally, land cover and elevation data that are used in the construction of least cost path highway routes were obtained from the US Geological Survey Digital Chart of the World project, and complemented by higher resolution Chinese hydrology data from the ACASIAN data center.

3 Empirical Strategy

The data described in the previous section is used to estimate the effects of NTHS network connections among peripheral counties between 1992-2003 on changes of economic outcomes between 1997-2006. The baseline estimation strategy is a difference in differences specification of the form:

$$\ln\left(y_{ip}^{2006}\right)\text{-}\ln\left(y_{ip}^{1997}\right)=\gamma_p+\beta\, Connect_{ip}+\eta \mathrm{X}_{ip}+\epsilon_{ip} \tag{1}$$

where y_{ip} is an outcome of interest of county i in province p, γ_{p} is a province fixed effect, $Connect_{\mathrm{ip}}$ indicates whether i was connected to the NTHS between 1992-2003, and X_{ip} is a vector of county control variables described below. I classify highway connections using GIS with a dummy indicator that takes the value of one if any part of county i is within a 10km distance of a NTHS highway that was opened to traffic before the end of 2003. Alternatively, I run specification (1) with a continuous treatment variable, $\ln DistHwy_{\mathrm{ip}}$, which stands for the logarithm of great circle distance to the nearest NTHS highway segment opened to traffic before the end of 2003, measured from the center of each county unit. Given that 89% of the reported NTHS connections until the end of 2003 were completed during the phase of "rapid development" between 1998-2003, the main source of variation used in the estimations stems from network connections during this five year period 1998-2003 and their effects on changes of economic outcomes over the nine year period 1997-06.

The error term ϵ_{ip} could be correlated across counties that were connected to a similar part of the network during a similar period between 1992-2003. I therefore cluster standard errors at the level of 33 Chinese provinces. Alternatively, I follow Conley (1999) and allow for spatial dependence to be a declining function over bilateral county distances without

[7] The available series of atlas sources did not allow to date the opening to traffic of each segment of the 35,000 km NTHS road network. See the Online Appendix for a listing of the atlas publications.

imposing parametric assumptions. Finally, due to the fact that the explicitly targeted network nodes are China's largest city regions that encompass multiple county level units, I exclude county observations within a 50 km commuting radius around the targeted city centers.[8]

3.1 Least Cost Path Spanning Tree Networks

Estimating specification (1) by OLS would imply the assumption that county connections between nodal cities were randomly assigned within provinces. Given the policy setting of the NTHS, this assumption would be strong. The NTHS was planned in 1992 to establish the backbone of a modernized road transport system for China. Province and county governments borrowed against future expressway toll revenues to finance its construction. This background raises the concern that planners targeted politically important and economically prosperous regions on the way between the network's nodal cities. This concern is supported by descriptive statistics presented in Table 1. Peripheral counties connected to the network by 2006 were on average larger, richer, more urbanized, and more industrialized than non-connected peripheral counties before the bulk of the network had been built in 1997.[9]

To address these concerns, I construct two hypothetical minimum spanning tree highway networks as instruments for actual route placements (see Figures 2 and 3). I refer to the first as least cost path spanning tree network, and to the second as Euclidean spanning tree network. Both instruments correspond to the question of which routes central planners would have been likely to construct if the sole policy objective had been to connect all targeted destinations on a single network in a least costly manner. The least cost path network yields more precise route predictions between any given bilateral connection on an all-China minimum spanning tree due to its use of land cover and elevation data, while the Euclidean network covers a larger set of the actually built bilateral network routes.

The following paragraphs describe the spanning tree instruments. A more detailed description can be found in the Online Appendix. The least cost path network depicted in Figure 2 is constructed in a two-step procedure. The first step is to compute least cost highway construction paths between all possible targeted destination pairs on the basis of remote sensing data on land cover and elevation. To this end, I adopt a simple construction cost function from the transport engineering literature that assigns higher construction costs to land parcels with steeper slope gradients and land cover classified as water, wetlands, or built structures (Jha et al., 2001; Jong and Schonfeld, 2003).[10] I use the remote sensing data to create a construction cost surface covering the PR China in a rectangular grid of

[8] See Garske et al. (2011) for a study of commuting patterns and distances in China.

[9] Reported differences are statistically significant at the 1% level.

[10] As discussed further below, I will also include these geographical characteristics used in the construction of the instrument as direct county level controls to address the concern that these might affect changes in economic outcomes directly.

cost pixels (see Online Appendix for details and illustrations).

I then implement Dijkstra's optimal route algorithm to construct least cost highway construction paths between all possible bilateral city connections as well as provincial capitals and the border in border provinces. In the second step, I extract the estimated aggregate construction cost of each bilateral connection and feed them into Kruskal's minimum spanning tree algorithm. This algorithm yields the minimum number of least cost connections (i.e. number of targeted nodes minus one) to connect all targeted destinations on a single continuous network to minimize aggregate construction costs.

To construct the Euclidean spanning tree network depicted in Figure 3, the first step is to compute great circle distances between all possible bilateral connections of the network. I then compute Kruskal's algorithm to identify the minimum number of edges that connect all targeted destinations subject to the minimization of total network distance. To compensate for the loss of route precision, I account for the fact that Chinese planners constructed many more than the minimum number of spanning tree connections. I therefore re-run Kruskal's algorithm within separate geographical subdivisions after dividing China into North-Center-South or East-Center-West geographical divisions.[11] These six additional spanning tree computations add nine bilateral routes in addition to the single all-China minimum spanning tree. The Online Appendix provides further details and additional illustrations of these computations.

3.2 Additional County Controls and Identifying Assumption

The minimization of a network construction cost objective function from which the instruments in Figures 2 and 3 are derived is aimed to address the concern of non-random highway placements on the way between targeted destinations. However, the exclusion restriction could be violated if locations along least cost road construction paths between major economic centers in China are correlated with economic county characteristics due to history and sorting. Furthermore, the instrument is likely to be mechanically correlated with distance to the nearest targeted metropolis. I therefore estimate regressions before and after including a set of additional county controls that could be correlated with the instrument while also affecting the change of economic outcomes between 1997-2006.

Counties closer to targeted city centers are mechanically more likely to lie on a least cost spanning tree path than counties situated farther away. Concerns about the exclusion restriction arise if distances to the major cities of China are correlated with economic county characteristics that also affect growth trajectories. I include the log distance between counties and the nearest targeted metropolitan city center to address this concern.

Conditional on county distance to the targeted centers, location on least cost road construction paths between major economic centers in China could be correlated with political

[11] I define these geographical areas on the basis of six geographic regions with official administrative recognition in China: East, North, North-East, North-West, South-Central, and South-West.

and economic county characteristics due to historical trade routes. To address such concerns, I include a set of observable controls for pre-existing county level political status and economic conditions. The two political controls are dummy variables indicating whether the county seat was a prefectural capital or has city rather than township status in 1990. The concern is that higher administrative status might be historically concentrated along least cost path routes between important economic centers.

Concerning pre-existing economic conditions, I use data from the 1990 Census at the county level which allows me to compute the share of agricultural employment in total county employment, the logarithm of county level urban registered population, as well as the share of above compulsory schooling attainment in total county population above 20 years of age in 1990.[12] These controls are aimed to address concerns that counties along least cost connections between major cities differ in terms of both their economic composition (shares of skilled labor and sectoral specialization), as well as in their mass of economic activity (urban populations).

The baseline identifying assumption is that county location along an all-China least cost spanning tree network affects changes in county level economic outcomes only through NTHS highway connections, conditional on province fixed effects, distance to the nearest targeted city region, administrative status and county-level economic conditions in 1990. I discuss in detail a series of robustness checks after reporting baseline estimation results.

4 Estimation Results

This section reports estimation results and robustness checks of specification (1) for a number of different county level economic outcomes. Table 2 presents the first stage results for the least cost path and the Euclidean network instruments as well as their combined first stage results. First stages are run for binary NTHS connection indicators as well as the log distance to the nearest NTHS segment. Both the least cost path and the straight line networks are strongly significant within province predictors of actual NTHS placements conditional on log distance to the nearest targeted node and the full set of pre-existing political and economic county characteristics. County controls are related to NTHS exposure mostly as expected. NTHS route connections are more likely for counties with lower distances to the targeted metropolitan city centers, larger pre-existing urban populations, and city status.[13]

Both instruments remain statistically significant when included simultaneously, confirm-

[12]Categories beyond the compulsory 9-year curriculum are senior middle school, secondary technical school, technical college, junior college and university.

[13]When individually included, higher pre-existing shares of agricultural employment decrease the likelihood of route placements, and higher shares of educated population increase it. These correlations are no longer significant when both controls are added simultaneously as reported in Table 2. Finally, the identifier for prefecture level capital status in 1990 enters in opposite sign than expected (decreased likelihood). This is due to the simultaneous inclusion of the city identifier, so that the coefficient is driven by approximately 1% of relatively remote prefecture level capitals that are not also classified as cities in 1990.

ing that the two spanning tree networks capture slightly different sources of the increased likelihood of route placements. While the least cost path instrument is a more precise predictor of placements on any given bilateral connection, the Euclidean tree instrument captures a higher proportion of the actually built network connections.

Table 3 presents OLS and IV results when regressing log changes of county level outcomes on the binary network treatment variable before and after including the full set of 1990 county controls. The instrumental variable estimates of the NTHS connection effect are negative and statistically significant for industrial output growth, non-agricultural output growth, local government revenue growth, as well as total GDP growth.

Two important patterns emerge from Table 3. The first is that the IV point estimates are more negative than the OLS estimates. The second is that the inclusion of additional county controls for pre-existing political status and economic conditions leads to more negative point estimates of the NTHS connection effect. These findings are in line with the discussed concern that planners targeted places with higher expected returns to infrastructure investments and/or higher expected traffic demand, which is also apparent in the descriptive statistics of Table 1.

The results in Table 3 suggest that county location along least costly road construction paths between major cities in China is at least partly correlated with pre-existing county characteristics, such as the size of urban populations, the share of educated labor, and the degree of industrialization. As noted in the previous subsection, these correlations could be driven by settlements and sorting along historical trade routes. However, the finding that conditioning on pre-existing county characteristics leads to more negative connection treatment effects on industrial output growth, total output growth, as well as local government revenue growth suggests that these characteristics are positively associated with economic growth, rather than negatively.

Nevertheless, the sensitivity of the IV point estimates to the inclusion of county controls in principle raises the concern that the estimated effects remain biased in either direction due to omitted unobserved differences that are correlated with the instrument. A related concern is that counties along an all-China spanning tree network had different pre-existing growth trends before the highway network came into effect. To address such concerns, I make use of the fact that the majority of the reporting county sample in 1997 and 2006 also reported local government revenues in the Provincial Statistical Yearbooks for the year 1990. If the exclusion restriction is satisfied conditional on the included county controls, then we should expect to find no significant relationship between NTHS treatments and local government revenue growth prior to the network, when estimated on the identical county sample for both periods.

Table 4 presents OLS and IV results for both instruments in both periods 1990-1997 and 1997-2006. For completeness, the table also reports results for the continuous NTHS exposure variable measured by the log county distance to the nearest NTHS route for both periods. The county sample is smaller than in the previous regressions firstly because not

all reporting counties in 1997-2006 have non-missing entries for local government revenue in 1990, and secondly because these regressions exclude counties that were connected to NTHS routes built between 1992-1997 (10% of the NTHS).

The connection indicator enters negatively and statistically significantly only for the NTHS period, and the log distance to the nearest NTHS segment enters positively and statistically significantly only for the NTHS period. Furthermore, these outcomes are not driven by differences in the size of the standard errors of the point estimates across the two different periods, but by changes in the point estimates themselves. These results provide a reassuring robustness check of the exclusion restriction conditional on the included set of pre-existing political and economic county controls.[14]

In terms of magnitudes of the estimated effects, the IV estimate using both spanning tree instruments and the full set of county controls as reported in the final column of Table 3 suggests that NTHS connections on the way between targeted destinations have on average reduced GDP growth by about 18 percent over a nine year period between 1997-2006 compared to non-connected peripheral counties. Local government revenue growth is reduced by approximately 23 percent. These adverse growth effects appear to be mainly driven by a decline in industrial output growth of approximately 26 percent over the nine year period.[15]

The results for agricultural GDP growth are close to zero and not statistically significant. This finding would point against a reallocation of factors of production from industry to agriculture. The result could be due to labeling both more and less industrialized activities as agricultural in county level economic accounts. An economic explanation could be related to factor market rigidities or adjustment costs and the frequent empirical finding that the inter-sectoral reallocation of resources following trade shocks fails to be confirmed in the data (Goldberg and Pavcnik, 2007; Topalova, 2010).

The final result from Table 3 is that NTHS connections appear to have no significant effect on county population growth. The point estimates are close to zero and statistically insignificant conditional on controls. This result is consistent with Chinese migration controls under the hukou system (e.g. Au and Henderson, 2006), and suggests that the estimated effects on output growth are not driven by significant differences in population growth across counties.[16]

In addition to the results reported in this section, the Online Appendix includes a series of additional robustness checks concerning the estimated average NTHS connection effects

[14]The system of government revenue collection underwent significant reforms in 1994 under the so called tax sharing system (Wong, 2000, Qiao et al., 2008; Lin, 2009). As an additional robustness check, I also report the placebo specification on the selected subsample of counties with long enough series of reported production data. As reported in the Online Appendix, the falsification results also hold for industrial output growth in this smaller subsample.

[15]The cited estimates correspond to point estimates as shown in the regression tables after converting log points back to percentage changes and rounding.

[16]The following section also reports a further robustness check on the possibility of unreported outmigration that could remain uncaptured when estimating the NTHS effect on registered county populations.

on production and government revenue growth. These address potential concerns about i) controlling for direct effects of the land cover and elevation characteristics used in the construction of the least cost path instrument, ii) controlling for construction activity already underway in 1997, iii) excluding mountainous regions due to least cost path endogeneity concerns, iv) excluding the Golmud-Lasa railway route completed over the same period, v) controlling for proximity to historical trade routes such as the Silk Road, and vi) controlling for 1997 county differences in distance weighted market access to all other counties. The results discussed in this section are robust in their magnitude and statistical significance in these additional estimations.

The Online Appendix also provides a discussion and estimation results concerning the proportion and observable characterization of the complier group of counties that drive the estimated local average treatment effects. Descriptive statistics and the pattern of coefficient estimates discussed above suggest that planners targeted economically prosperous counties on the way between targeted city regions. The additional estimations address the concern that least cost spanning tree location might have affected actual highway placements only for a subset of remote and economically stagnant counties on the way between targeted nodes, so that the estimated local average NTHS connection effects might systematically differ from population average effects. The results presented in the Online Appendix suggest little support for this concern, showing that the predictive power of the instruments does not significantly vary across observable pre-existing county characteristics. The reasons behind this finding are due to the nature of the spanning tree prediction errors compared to actual NTHS route placements which I further explore in a set of cartographic illustrations in the Online Appendix.

5 Channels at Work

The preceding section has presented empirical evidence suggesting that NTHS network connections have led to reduced GDP growth among peripheral counties on the way between targeted metropolitan city nodes relative to non-connected peripheral regions. These effects appear to be mainly driven by a reduction in industrial output growth. This section discusses three underlying channels that could be driving these empirical findings, and presents a series of additional estimation results that aim to further investigate the channels at work.

The first channel (CP) is based on core-periphery effects of trade integration between *ex ante* asymmetric markets as found in the literature on increasing returns trade theory and economic geography. In the presence of increasing returns to scale in industrial production, Dixit-Stiglitz monopolistic competition, and iceberg trade costs, Krugman (1980), Helpman and Krugman (1985), and Krugman (1991) have provided a microfoundation for the proposition that market size is a determinant of industrialization, and that falling trade costs between large and small markets can reinforce the concentration of production in the

larger market.[17]

This result can either be driven solely by the so called home market effect as in Krugman (1980) and Helpman and Krugman (1985), or by an interplay of forces including additional self-reinforcing agglomeration forces that arise in the context of labor mobility as introduced by Krugman (1991). The common foundation of the core-periphery effect of trade integration across different modeling settings in this literature is that while falling trade costs reduce the strength of both agglomeration forces (such as better access to consumers or to intermediate inputs in larger markets) and dispersion forces (such as less product market competition or lower factor prices in smaller markets), they tend to attenuate the agglomeration forces at a lower rate compared to the dispersion forces.

The second channel (UD) is related to recent findings of Baum Snow et al. (2012) in the urban economics literature. Using rich historical geographical information about China's urbanization process, they document that a process of urbanization and industrial decentralization has taken place over recent decades among the central city districts of Chinese prefecture level capitals. That is, while the populations have grown at higher rates in central cities compared to their urban peripheries, the opposite has been the case for industrial production growth. In this context, they present evidence that radial road and railway transport infrastructures can have significant effects in shaping this process of urbanization and decentralization across cities.

These findings could be relevant in two central ways for the interpretation of the presented reduced form effects of NTHS connections among peripheral counties in China. First, the radial network routes of the NTHS could have a causal effect on the urbanization and industrial decentralization of connected peripheral counties. Alternatively, it could be the case that the non-targeted peripheral counties include a large fraction of second tier cities which undergo a process of urbanization and decentralization, while least cost path location along the all-China spanning tree instrument could be correlated with the location of such city centers. In other words, the exclusion restriction could be violated because the instrument picks up decentralizing city centers relative to surrounding peripheral counties. In both scenarios, the negative reduced form effects reported in the previous section would be driven by a decentralization of industrial production from connected peripheral regions to neighboring areas, as opposed to adverse core-periphery effects of trade integration between the large metropolitan city regions and the non-targeted peripheral regions in the CP channel.

The third channel (CA) derives from neoclassical constant returns to scale trade theory. In the presence of locational fundamentals that give rise to a comparative disadvantage in industrial production among peripheral regions relative to metropolitan centers, falling trade costs would lead to a decline in industrial production among connected peripheral regions. In turn, total production could be adversely affected in this setting in the presence

[17]See Fujita et al. (1999), Baldwin et al. (2003), and Combes et al. (2008) for reviews of this literature.

of either frictions in the reallocation of factors of production across sectors (Goldberg and Pavcnik, 2007; Topalova, 2010), or due to outflows of factors used in industrial production (e.g. capital) (Redding, 2012).

While in principle each of the above channels could give rise to the estimated reduced form effects on peripheral county outcomes, the mechanisms would also have different testable implications on a series of additional observable economic outcomes. Motivated by the channels discussed above, Table 5 presents a summary of six additional testable implications of peripheral network connections which the remainder of this section seeks to investigate empirically.

5.1 Additional Estimation Results

The first three additional estimations summarized in Table 5 are aimed to differentiate between the UD channel as opposed to either CP or CA. A mechanism based on urbanization and decentralization would imply that peripheral connections are correlated to increases in urbanization among connected peripheral counties relative to non-connected ones. It would also imply that counties in close proximity to the connected peripheral markets are affected more positively by NTHS connections in terms of industrial production and GDP relative to peripheral counties farther away from the direct neighborhood of the connected counties. As summarized in Table 5, neither of these predictions would hold under either the CP or the CA channels. Furthermore, while both CP and CA would imply that the adverse peripheral connection effect is decreasing in county distance to the nearest NTHS route, the UD channel predicts a non-monotonicity in the relationship between the network connection effect and county distance to the NTHS.

Table 6 presents the estimation results on these three testable implications. Columns 1 and 2 report the estimated NTHS treatment effect on urban population growth as well as changes in urbanization among connected peripheral counties after using both least cost path and Euclidean spanning tree instruments to instrument for route placements. The finding that the IV point estimates are close to zero and statistically insignificant in both cases presents initial evidence that a process of urbanization is unlikely to underlie the reduced form effects presented in the previous section.

The remaining Columns 3-11 of Table 6 further investigate the spatial pattern of the estimated NTHS connection effects among non-targeted peripheral counties. In particular, Columns 4, 7, and 10 report the IV estimate of the NTHS treatment effect on growth in industrial production, GDP and government revenues respectively after including an additional indicator variable which takes the value of one for peripheral counties that are next to connected peripheral counties and zero otherwise.[18] This specification implies a

[18]To instrument for the additional neighbor dummy, I use both an indicator for adjacency to a connected county on the LCP spanning tree and an indicator for adjacency to a connected county on the Euclidean spanning tree.

partition of the peripheral county sample into connected counties, neighboring counties, and counties farther away from connected counties which represent the new reference category in the estimation.

The finding that the baseline NTHS effect reported in Columns 3, 6, and 9 is virtually unchanged by the inclusion of the additional peripheral neighbor dummy does not support the prediction that surrounding peripheral areas are more positively affected by NTHS connections compared to counties farther away. Instead, the finding suggests that the estimated adverse effect of network connections is concentrated within connected county regions.[19]

The final columns of Table 6 and Figure 4 confirm this finding of a relatively steeply declining adverse connection effect over distance to the nearest NTHS route with respect to industrial output growth, GDP growth, as well as local government revenue growth. Columns 5, 8 and 11 report a statistically significant positive effect of log distance to the nearest NTHS route placement for all three dependent variables.[20] Figure 4 then plots the best fitting polynomial functional forms of the flexibly estimated relationships between county distance to the nearest NTHS segment and industrial output growth, GDP growth, or government revenue growth.[21] The graphs suggest a relatively steeply declining monotonic relationship of the adverse connection effect over distance to the NTHS, and once again there is no evidence of a non-monotonicity in this relationship around the median distance of peripheral counties in the direct neighborhood of connected counties. To summarize, the findings reported in Table 6 and Figure 4 provide evidence that appears consistent with the CP and CA channels, while they provide little support for the UD channel in the context of NTHS connections among non-targeted peripheral counties in China.

The final three additional estimations summarized in Table 5 are aimed at differentiating between the CP and CA channels. An increasing returns based CP mechanism and a neoclassical constant returns based CA mechanism give rise to different testable implications on the heterogeneity of the peripheral connection effects with respect to pre-existing county characteristics. In particular, a CA channel would imply that pre-existing differences in relative production costs and sectoral specialization among peripheral counties are significant determinants of both the sign and intensity of the effect of trade integration between peripheral regions and targeted metropolitan centers of production.

In contrast, the CP channel gives rise to two principal sources of heterogeneity in the effect of peripheral NTHS connections. The first is that the adverse connection effect should be less pronounced among peripheral counties with larger pre-existing market sizes. The second determinant is that, holding constant initial differences in market sizes, the adverse

[19] As pointed out by a referee, these results are also informative with respect to so called agglomeration shadows (Fujita and Krugman, 1995). In particular, I do not find evidence suggesting that connected peripheral counties attract industrial activity from neighboring non-connected counties.

[20] To instrument for distance to the nearest NTHS segment, I use both distance to the nearest LCP spanning tree and the nearest Euclidean spanning tree as instruments.

[21] The plotted relationships correspond to the best fitting specification according to the Akaike Information Criterion (AIC). Distance terms are instrumented with distances to LCP and Euclidean spanning trees.

connection effect should be less pronounced among peripheral counties whose initial level of trade costs *vis-a-vis* the metropolitan core regions is higher. The Online Appendix provides a formal exposition of these results in a simple core-periphery model based on Helpman and Krugman (1985). The key intuition behind both interaction effects is that falling trade costs attenuate the locational advantage of producing in a remote location at a faster rate than the agglomeration advantage of producing in the core, so that for a more pronounced initial core-periphery size gradient, or for lower initial trade costs, a given reduction in trade costs induces more industrial concentration in the core regions.[22]

To test the heterogeneity of the network's effects among peripheral counties, I estimate specification (1) after including interaction terms between the NTHS connection indicator and a number of theoretically motivated observable pre-existing county characteristics. Alternatively, I replace the binary NTHS connection indicator in these interactions by the continuous treatment variable of log distance to the nearest NTHS route as in the previous estimations reported in Table 6. In particular, county log distance to the nearest targeted metropolitan node and an indicator for counties with above mean 1990 total employment sizes are aimed to capture the initial degree of trade costs between targeted core cities and peripheral counties and the pre-existing degree of core-periphery size differentials respectively. Alternatively, 1990 county levels of agricultural output shares, the share of skilled workers, log government revenue per capita, city status and prefecture level capital status are aimed to capture pre-existing differences in county specialization and relative factor prices under the CA channel.

Table 7 reports second stage IV estimation results for industrial output growth as well as total GDP growth as dependent variables after instrumenting for NTHS connection and its interaction terms or for log NTHS distance and its interaction terms with the least cost path network instrument.[23] The first column for each dependent variable reproduces the average treatment estimate conditional on the full set of county controls as reported in Table 3. The second columns then introduce the additional two interaction terms motivated by the CP channel. In the binary network connection specifications both interaction terms enter positively and statistically significantly for both industrial output growth and GDP growth. When using the log distance to the nearest NTHS segment as continuous treatment variable instead, the interaction terms enter negatively and statistically significantly for both dependent variables.[24]

[22]These results are related to the concept of home market magnification in the trade literature (Baldwin et al., 2003). See Online Appendix for discussion.

[23]In the cross-derivative regressions the least cost path spanning tree instrument performs better in terms of first stage predictive power than the Euclidean instrument or both instruments combined. The Online Appendix reports these specifications using both instruments instead.

[24]A potential concern with these two stage least squares results is that the first stage F-statistics significantly drop after instrumenting for NTHS treatments as well as their interactions. The Online Appendix reports an additional set of results that compare two stage least squares (2SLS) estimates using both instruments to estimations by limited information maximum likelihood (LIML). The fact that the variation of first stage F-statistics between the different specifications across Columns 2-4 in Table 7 have little effect

Columns 3-5 of Table 7 then report results after including additional interaction terms that are aimed to capture pre-existing differences in relative production costs and sectoral specialization under the CA channel. The reported estimates now control for the heterogeneity of the NTHS connection effect on county level industrial or total output growth with respect to 1990 differences in the share of skilled population, the share of agricultural employment, county status as a city or a prefecture level capital, and local government revenue per capita. The finding is that the point estimates of the CP interaction terms in the second columns are hardly affected, whereas the additional CA interaction terms in Columns 3-5 enter statistically insignificantly across all reported specifications.

While the additional results on the heterogeneity of the NTHS connection effect in Table 7 do not represent formal tests or conclusive evidence, they contribute to a significantly richer understanding of the estimated reduced form effects in Section 4. In particular, the reported results provide empirical evidence of significant heterogeneity in the NTHS connection effects among peripheral counties that is consistent with the CP channel. While the majority of both large and small peripheral counties are estimated to have experienced negative GDP growth effects, a subset of large peripheral counties that are also subject to large initial trade costs with respect to the targeted metropolitan city centers are estimated to have experienced positive growth effects due to NTHS connections.[25] In contrast, when conditioning on the CP channel, I find little empirical support for a neoclassical CA mechanism in the context of the NTHS policy among peripheral counties in China.[26]

To summarize, the estimation results of this section summarized in Table 5 and reported in Tables 6 and 7 and Figure 4 provide additional qualitative evidence of the NTHS connection effects on urbanization, decentralization of production, as well as the heterogeneity of the effects with respect to pre-existing conditions among peripheral counties. In conclusion, the presented findings provide an empirical picture that appears consistent with the testable implications of a core-periphery channel of trade integration between large and small markets as found in increasing returns trade theory and economic geography.

---

on the point estimates of the two interaction terms of interest, and the fact that the reported LIML point estimates are slightly higher than the 2SLS estimates reported in the Online Appendix indicate that weak instrument bias is unlikely to be a concern. See for example Angrist and Pischke (2008, Section 4.6) for a discussion 2SLS and LIML estimates in the context of weak instrument concerns.

[25] The median and mean distances to the nearest targeted city node were 168 km and 203 km respectively.

[26] The results in the two final columns 5 in Table 7 can also be informative on another potential mechanism which involves unregistered outmigration. The concern would be that unregistered migrants move to equalize real wages across counties net of the costs of migration. Metropolitan centers have higher real wages, and NTHS connections tip the balance for a larger number of peripheral citizens to outmigrate compared to non-connected counties. Similar to the CA channel considered above, a testable implication would be that the heterogeneity is driven by pre-existing regional factor returns, rather than market size or the initial trade cost position. In particular, one would expect the interaction point estimates with respect to county size and metropolitan distance to be sensitive to the inclusion of the additional interaction with agricultural output shares, shares of skilled populations, urban status, or log local government revenue per capita reported in 1990 which serves as a proxy for regional GDP per capita differences. The results reported in Table 7 provide little evidence in support of this alternative mechanism.

6 Conclusion

Large scale transport investments connect both large metropolitan centers of production as well as small peripheral market places. This is especially the case in developing countries where spatial disparities are particularly pronounced. A common policy view is to combine national efficiency objectives of transport policies with regional equity objectives under the presumption that falling trade costs between large and small markets promote both national gains from trade as well as the diffusion of economic activity to peripheral regions. This paper exploits China's NTHS as a source of plausibly exogenous variation in trade cost shocks across a large number of ex ante asymmetric regions to provide empirical evidence on this question.

The presented analysis provides the following main novel insights. First, the paper presents empirical evidence in support of the hypothesis that falling trade costs between large and small markets can lead to a reduction of industrial and total output growth in peripheral regions, rather than diffusing economic activity in space. Second, additional estimation results present evidence that appears consistent with the existence of core-periphery effects of trade integration between large and small markets as found in increasing returns trade theory and economic geography. These findings provide empirical support for existing theoretically motivated policy discussions in the trade literature (Fujita et al., 1999; Baldwin et al., 2003; Combes et al., 2008), and serve to emphasize the importance of potentially unintended general equilibrium consequences when evaluating and planning large scale transport infrastructure policies.

Finally, it is important to emphasize that the presented empirical findings are perfectly consistent with potentially large aggregate efficiency improvements. Each of the discussed channels underlying the estimated reduced form effects would also imply aggregate gains from trade. In this context, it is also important to point out some of the main limitations of the presented analysis. In particular, while the NTHS policy design in combination with the proposed empirical strategy have the advantage to exploit plausibly exogenous variation in network connections among non-targeted peripheral counties, the current empirical setting would not be suited to evaluate the network's economic impact at the national level, which I leave open for future research.

References

Angrist, J., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Asian Development Bank. (2007). A retrospective analysis of the road sector 1997-2005. *Asian Development Bank Operations Evaluation Department Report.*

Asian Infrastructure Monthly. (1995). China paves the way for expressway network. *Asian Infrastructure Monthly, June.*

Au, C., & Henderson, J. (2006). How migration restrictions limit agglomeration and productivity in China. *Journal of Development Economics*, *80*(2), 350–388.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G., & Robert-Nicoud, F. (2003). *Economic geography and public policy*. Princeton University Press.

Banerjee, A., Duflo, E., & Qian, N. (2012). On the road: Access to transportation infrastructure and economic growth in China. *Yale Department of Economics mimeo*.

Baum-Snow, N. (2007). Did highways cause suburbanization? *Quaterly Journal of Economics*, *122*(2), 775.

Baum-Snow, N., Brandt, L., Henderson, J., Turner, M., & Zhang, Q. (2012). Roads, railroads and decentralization of Chinese cities. *mimeo, Brown University*.

Baum-Snow, N., & Turner, M. (2012). Transportation and the decentralization of Chinese cities. *mimeo, Brown University*.

Behrens, K., Lamorgese, A., Ottaviano, G., & Tabuchi, T. (2009). Beyond the home market effect: Market size and specialization in a multi-country world. *Journal of International Economics*, *79*(2), 259–265.

Brülhart, M., & Trionfetti, F. (2009). A test of trade theories when expenditure is home biased. *European Economic Review*, *53*(7), 830–845.

CIBC. (2009). Capitalizing on the upcoming infrastructure stimulus. *CIBC World Markets Occasional Report No. 66*.

Combes, P., Mayer, T., & Thisse, J. (2008). *Economic geography: The integration of regions and nations*. Princeton University Press.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, *92*(1), 1–45.

Davis, D. (1998). The home market, trade, and industrial structure. *American Economic Review*, *88*(5), 1264–76.

Davis, D., & Weinstein, D. (1996). Does economic geography matter for international specialization? *National Bureau of Economic Research Working Paper*.

Davis, D., & Weinstein, D. (1999). Economic geography and regional production structure: An empirical investigation. *European economic review*, *43*(2), 379–407.

Davis, D., & Weinstein, D. (2003). Market access, economic geography and comparative advantage: An empirical test. *Journal of International Economics*, *59*(1), 1–23.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Donaldson, D. (2013). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review, forthcoming*.

Duranton, G., & Turner, M. (2012). Urban growth and transport. *Review of Economic Studies*, *79*, 1407–1440.

Fujita, M., & Krugman, P. (1995). When is the economy monocentric?: von Thunen and Chamberlin unified. *Regional Science and Urban Economics*, *25*(4), 505–528.

Fujita, M., Krugman, P., & Venables. (1999). *The spatial economy: Cities, regions and international trade*. Wiley.

Garske, T., Yu, H., Peng, Z., Ye, M., Zhou, H., Cheng, X., & Ferguson, N. (2011). Travel patterns in China. *PloS one*, *6*(2).

Goldberg, P., & Pavcnik, N. (2007). Distributional effects of globalization in developing countries. *Journal of Economic Literature*, *45*, 39–82.

Hanson, G., & Xiang, C. (2004). The home-market effect and bilateral trade patterns. *American Economic Review*, 1108–1129.

Head, K., & Ries, J. (2001). Increasing returns versus national product differentiation as an explanation for the pattern of US-Canada trade. *American Economic Review*, 858–876.

Helpman, E., & Krugman, P. (1985). Market structure and international trade. *Cambridge (Mass.)*.

Hillberry, R., & Hummels, D. (2008). Trade responses to geographic frictions: A decomposition using micro-data. *European Economic Review*, *52*(3), 527–550.

Jha, M., McCall, C., & Schonfeld, P. (2001). Using GIS, genetic algorithms, and visualization in highway development. *Computer-Aided Civil and Infrastructure Engineering*, *16*(6), 399–414.

Jong, J., & Schonfeld, P. (2003). An evolutionary model for simultaneously optimizing three-dimensional highway alignments. *Transportation Research Part B: Methodological*, *37*(2), 107–128.

Kanbur, R., & Venables, A. J. (2005). *Spatial inequality and development*. Oxford University Press.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review*, *70*(5), 950–959.

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, *7*(1), 48–50.

Li, S., & Shum, Y. (2001). Impacts of the national trunk highway system on accessibility in China. *Journal of Transport Geography*, *9*(1), 39–48.

Lin, S. (2009). The rise and fall of China's government revenue.

Michaels, G. (2008). The effect of trade on the demand for skill: Evidence from the interstate highway system. *The Review of Economics and Statistics*, *90*(4), 683–701.

Puga, D. (2002). European regional policies in light of recent location theories. *Journal of Economic Geography*, *2*(4), 373.

Qiao, M.-V. J., B., & Xu, Y. (2008). The tradeoff between growth and equity in decentralization policy: China's experience. *Journal of Development Economics*, *86*(1), 112–128.

Redding, S. J. (2012). Goods trade, factor mobility and welfare. *NBER Working Paper*(w18008).

Redding, S. J., & Sturm, D. M. (2008, December). The costs of remoteness: Evidence from German division and reunification. *American Economic Review*, *98*(5), 1766-97.

South China Morning Post. (1994). Construction on the road to success. *South China Morning Post*, *September 08*.

Topalova, P. (2010). Factor immobility and regional impacts of trade liberalization: Evidence on poverty from India. *American Economic Journal: Applied Economics*, *2*(4), 1–41.

Wong, C. (2000). Central-local relations revisited the 1994 tax-sharing reform and public expenditure management in China. *China Perspectives*, 52–63.

World Bank. (2007a). China study tour by New Delhi transport office team. *World Bank Internal Report*.

World Bank. (2007b). Domestic trade impacts of the expansion of the national expressway network in China. *World Bank EASTR Working Paper No. 14*.

World Bank. (2008). Safe, clean, and affordable: Transport for development. *World Bank Transport Business Strategy 2008-2012*.

Appendix

Figures

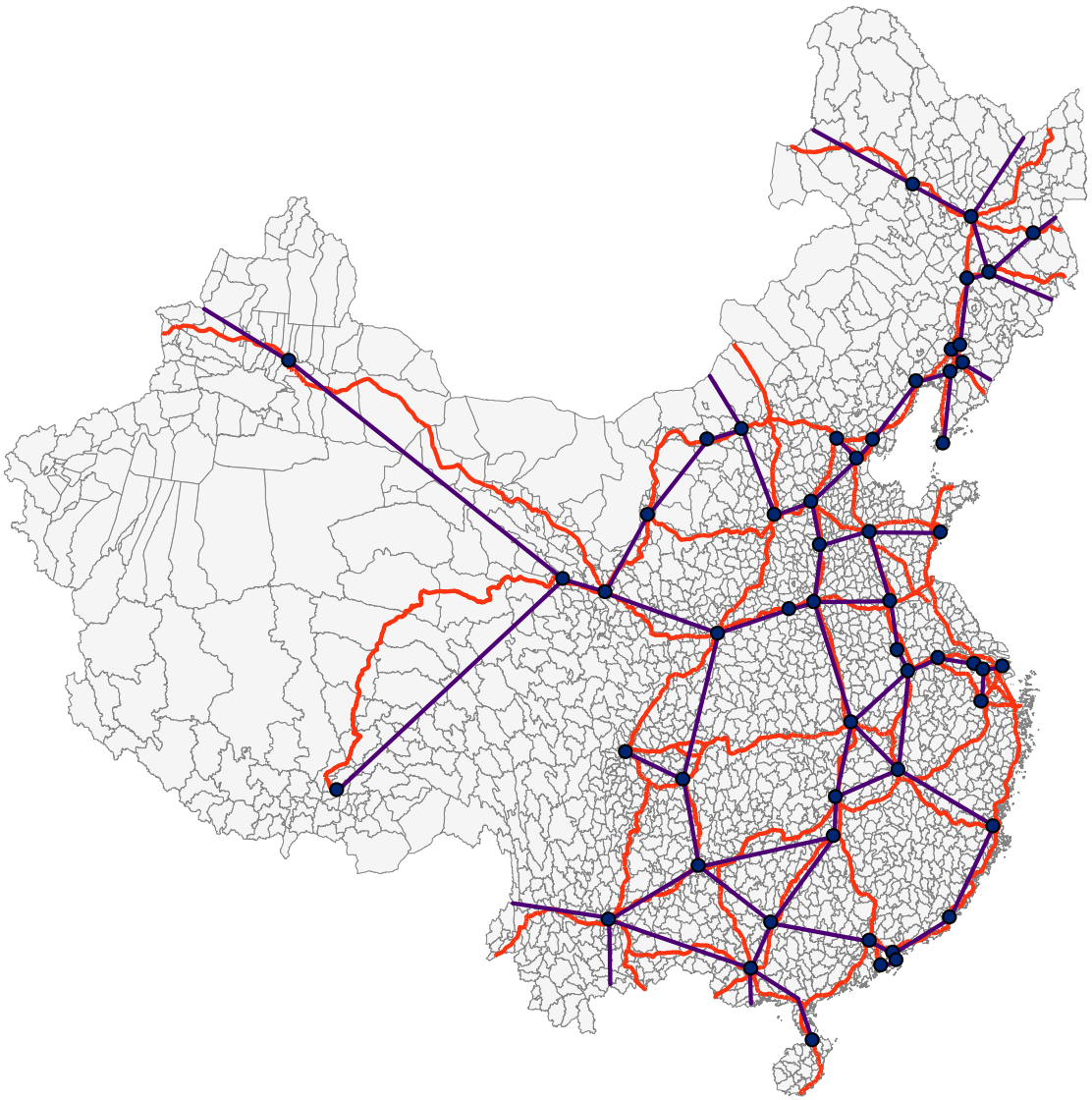
Figure 1: **China's National Trunk Highway System**

The figure shows Chinese county boundaries in 1999 in combination with the targeted city nodes and the completed expressway routes of the National Trunk Highway System (NTHS) in the year 2007.

Figure 2: **Least Cost Path Spanning Tree Network**

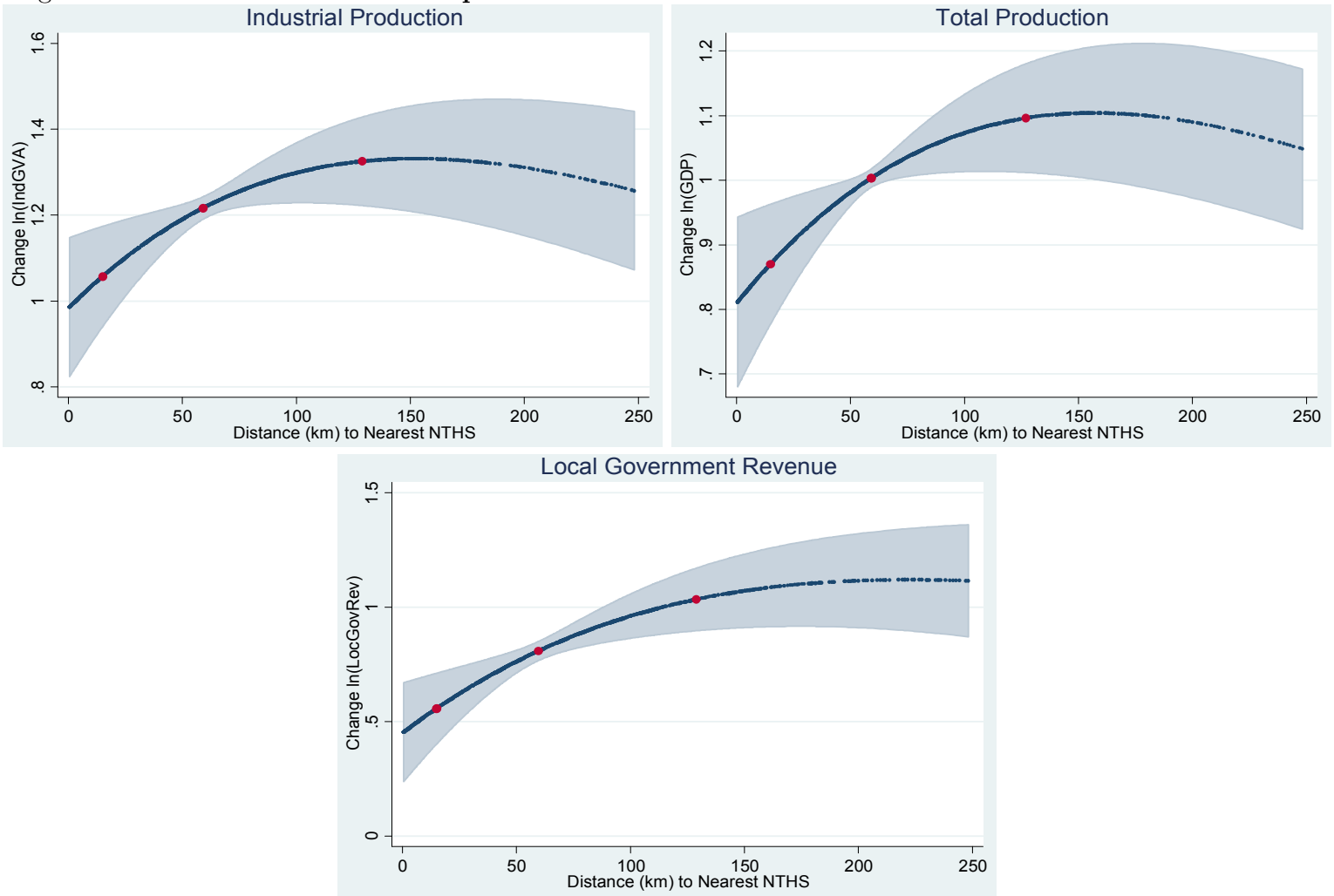
The network in red color depicts the completed NTHS network in 2007. The network in black color depicts the least cost path spanning tree network. The black routes are the result of a combination of least cost path and minimum spanning tree algorithms. In the first step Dijkstra's (1959) optimal route algorithm is applied to land cover and elevation data in order to construct least costly paths between each bilateral pair of the targeted centers. In the second step, these bilateral cost parameters are fed into Kruskal's (1956) minimum spanning tree algorithm to identify the minimum number of bilateral routes that connect all targeted cities on a single continuous network of the PR China to minimize total route construction costs. Border connections are least costly paths between provincial capitals to the border in border provinces. A detailed description of these computations and additional maps can be found in the Online Appendix.

Figure 3: **Euclidean Spanning Tree Network**



The network in red color depicts the completed NTHS network in 2007. The network in darker color depicts the Euclidean spanning tree network. The routes are the result of applying Kruskal's (1956) minimum spanning tree algorithm to bilateral Euclidean distances between targeted nodes. This algorithm is first run for the all-China network, and then repeated within North-Center-South and East-Center-West divisions of China. These repetitions add 9 routes to the original 53 bilateral connections. Connections between provincial capitals of border provinces and the border are minimum Euclidean distance paths. A detailed description of these computations and additional maps can be found in the Online Appendix.

Figure 4: **Estimated Effect of Peripheral Connections over Distance to the Nearest NTHS Route**



The graphs depict the flexibly estimated relationships between distance to the nearest NTHS route and peripheral county growth in industrial value added, total GDP, and local government revenue. The plots correspond to the best fitting polynomial functional form according to the Akaike Information Criterion (AIC). The functions and confidence intervals are based on IV estimates holding covariates at their mean. County distance to the NTHS and its polynomial terms are instrumented with distances to the LCP and Euclidean spanning trees and their polynomials. The red dots indicate median county distances to the nearest NTHS route among connected peripheral counties (left), peripheral counties neighboring a connected county (center), and the remaining peripheral counties farther away (right). The shaded areas indicate 90% confidence intervals. Standard errors are clustered at the province level.

Tables

Table 1: **Descriptive Statistics for 1997**

	Targeted City Centers	Connected Periphery	Non-connected Periphery	National Share of Targeted City Centers
Population (10,000)	233.24	56.96	38.48	0.14
Urban population (10,000)	179.69	10.77	5.83	0.44
GDP (100 Million Yuan)	517.86	32.58	15.09	0.42
GDP per capita (Yuan)	21435.06	5142.16	3637.09	-
Local government revenue (100 Million Yuan)	38.23	1.23	0.57	0.60
Industrial gross value added (100 Million Yuan)	194.61	14.93	5.58	0.39
Non-Agricultural gross value added (100 Million Yuan)	505.75	24.42	9.74	0.50
Agricultural output share	0.04	0.34	0.42	-
Land area (km^2)	1543.09	3057.47	4513.40	0.013
Number of counties	54	424	943	54

The first three columns present mean 1997 levels, and the fourth column presents national shares by county groups. Targeted city centers refer to the central city county units (shixiaqu) of targeted metropolitan regions. Peripheral counties are counties outside a 50 km commuting buffer around the targeted city centers.

Table 2: **First Stage Regressions**

Dependent Variable:	(1) Connect	(2) Connect	(3) Connect	(4) lnDistHwy	(5) lnDistHwy	(6) lnDistHwy
Least Cost Path IV	0.323***		0.254***	0.317***		0.245***
	(0.0574)		(0.0635)	(0.0645)		(0.0635)
Euclidean IV		0.243***	0.144**		0.280***	0.193***
		(0.0529)	(0.0560)		(0.0599)	(0.0657)
lnDistNode	-0.130***	-0.127***	-0.104***	0.588***	0.635***	0.426***
	(0.0376)	(0.0295)	(0.0323)	(0.130)	(0.112)	(0.122)
Prefect Capital	-0.124*	-0.129*	-0.120*	0.437**	0.429*	0.413*
	(0.0648)	(0.0736)	(0.0658)	(0.209)	(0.229)	(0.215)
City Status	0.0891**	0.0929**	0.0847**	-0.297***	-0.296***	-0.270***
	(0.0403)	(0.0437)	(0.0399)	(0.0946)	(0.103)	(0.0951)
lnUrbPop90	0.106***	0.115***	0.107***	-0.228***	-0.244***	-0.227***
	(0.0225)	(0.0217)	(0.0209)	(0.0691)	(0.0640)	(0.0636)
Educ90	-0.273	-0.303	-0.302	-1.671	-1.747	-1.626
	(0.598)	(0.656)	(0.601)	(1.697)	(1.804)	(1.666)
AgShare90	-0.170	-0.216	-0.167	0.0238	-0.00173	-0.0160
	(0.182)	(0.189)	(0.179)	(0.537)	(0.555)	(0.533)
Constant	-0.212	-0.314	-0.388	2.321**	2.627**	2.695**
	(0.335)	(0.299)	(0.293)	(1.103)	(1.049)	(0.992)
Obs	1342	1342	1342	1342	1342	1342
R^2	0.222	0.204	0.233	0.401	0.394	0.414
First stage F-Stat	31.61	21.07	20.31	24.09	21.82	15.00

All regressions include province fixed effects. Columns 1-3 report results for binary NTHS connection indicators among peripheral counties. Columns 4-6 report results for the log distance to the nearest NTHS segment among peripheral counties. lnDistNode is log county distance to the nearest targeted city node. Prefect Capital and City Status are binary indicators for respective county status in 1990. lnUrbPop90 is log 1990 county urban population. Educ90 is the 1990 county share of above compulsory schooling in 20+ population. AgShare90 is the 1990 county share of agricultural employment. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 3: **Network Connection Effects among Peripheral Counties**

Dependent Variables		(1) OLS No Controls	(2) OLS With Controls	(3) LCP IV No Controls	(4) LCP IV With Controls	(5) Euclid IV No Controls	(6) Euclid IV With Controls	(7) Both IVs No Controls	(8) Both IVs With Controls	(9) Both IVs With Controls
	Connect	-0.0529	-0.0356	-0.284**	-0.304**	-0.246*	-0.287*	-0.272***	-0.297***	-0.297**
Change ln(IndGVA)		(0.0418)	(0.0499)	(0.118)	(0.145)	(0.148)	(0.154)	(0.0965)	(0.108)	(0.121)
1997-2006	Obs	1302	1280	1302	1280	1302	1280	1302	1280	1280
	R^2	0.242	0.255							
	Connect	-0.0411	-0.0266	-0.243**	-0.252**	-0.270**	-0.296**	-0.251***	-0.268***	-0.268***
Change ln(NonAgGVA)		(0.0335)	(0.0375)	(0.0983)	(0.117)	(0.122)	(0.131)	(0.0877)	(0.0969)	(0.0100)
1997-2006	Obs	1285	1262	1285	1262	1285	1262	1285	1262	1262
	R^2	0.270	0.284							
	Connect	-0.0497*	-0.0914***	-0.0542	-0.223*	-0.175	-0.315**	-0.0926	-0.257***	-0.257***
Change ln(GovRevenue)		(0.0285)	(0.0295)	(0.109)	(0.120)	(0.117)	(0.132)	(0.0893)	(0.0996)	(0.0800)
1997-2006	Obs	1290	1285	1290	1285	1290	1285	1290	1285	1285
	R^2	0.275	0.334							
	Connect	-0.00204	-0.0144	-0.106	-0.177*	-0.178	-0.254**	-0.127	-0.203**	-0.203**
Change ln(GDP)		(0.0245)	(0.0276)	(0.0830)	(0.0942)	(0.112)	(0.116)	(0.0824)	(0.0886)	(0.0800)
1997-2006	Obs	1297	1272	1297	1272	1297	1272	1297	1272	1272
	R^2	0.228	0.264							
	Connect	-0.00344	-0.00790	0.000194	-0.0252	-0.0305	-0.0597	-0.00865	-0.0371	-0.0371
Change ln(AgGVA)		(0.0210)	(0.0220)	(0.0631)	(0.0789)	(0.0672)	(0.0728)	(0.0545)	(0.0630)	(0.0668)
1997-2006	Obs	1335	1313	1335	1313	1335	1313	1335	1313	1313
	R^2	0.202	0.208							
	Connect	0.00488	-0.00217	0.0395**	0.0264	0.0183	0.0104	0.0333*	0.0207	0.0207
Change ln(Population)		(0.00456)	(0.00568)	(0.0188)	(0.0234)	(0.0242)	(0.0262)	(0.0183)	(0.0215)	(0.0208)
1997-2006	Obs	1337	1314	1337	1314	1337	1314	1337	1314	1314
	R^2	0.234	0.271							

Each point estimate stems from a separate regression. All regressions include province fixed effects. LCP IV stands for the least cost path spanning tree instrument. Euclid IV stands for the straight line spanning tree instrument. No controls columns refer to regressions on NTHS treatment and log county distance to the nearest targeted city node. w/ controls indicates a full set of 1990 county controls (city status and prefecture capital dummies, log urban population, share of agricultural employment, and share of above compulsory school attainment in 20+ population). The dependent variables in order as listed are county level industry gross value added, manufacturing plus services gross value added, local government revenue, total GDP, agricultural gross value added, and population. Column (9) reports Conley (1999) standard errors, estimated by GMM to adjust for spatial dependence without imposing parametric assumptions. I allow for spatial dependence up to the distance given by the diameter of the average province area. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 4: **Falsification Test Before and After the Network Was Built**

Dependent Variable: Change ln(LocGovRev)	(1) OLS 1990-97	(2) OLS 1997-06	(3) LCP IV 1990-97	(4) LCP IV 1997-06	(5) Euclid IV 1990-97	(6) Euclid IV 1997-06	(7) Both IVs 1990-97	(8) Both IVs 1997-06
Panel A: Binary								
Connect	0.0154	-0.0848**	0.0143	-0.151	0.117	-0.282**	0.0563	-0.204***
	(0.0410)	(0.0360)	(0.0853)	(0.0974)	(0.107)	(0.129)	(0.0647)	(0.0467)
Obs	894	894	894	894	894	894	894	894
R^2	0.274	0.339						
First stage F-Stat			19.635	19.635	19.091	19.091	14.930	14.930
Panel B: log Distance								
ln(DistHwy)	-0.0114	0.0160	-0.0409	0.0854*	-0.00442	0.185**	-0.0274	0.122***
	(0.0142)	(0.0190)	(0.0350)	(0.0470)	(0.0573)	(0.0783)	(0.0329)	(0.0430)
Obs	894	894	894	894	894	894	894	894
R^2	0.275	0.336						
First stage F-Stat			18.696	18.696	17.306	17.306	11.259	11.259

Each point estimate stems from a separate regression. All regressions include province fixed effects and a full set of county controls. LCP IV stands for the least cost path spanning tree instrument. Euclid IV stands for the straight line spanning tree instrument. Panel A presents results for binary NTHS connection indicators and Panel B presents results for log distance to the nearest NTHS segment. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 5: **Summary of Channels**

	Core-Periphery	Urbanization and Decentralization	Comparative Advantage
Effect on industrial and total production among connected peripheral counties relative to non-connected	(-)✓	(-)✓	(-)✓
(1) Effect on urbanization of connected peripheral counties relative to non-connected counties	(?)✓	(+)	(?)✓
(2) Relationship between peripheral connection effect and distance to the nearest NTHS route	(+)✓	(?)	(+)✓
(3) Effect on peripheral counties in the neighborhood of connected ones relative to non-connected counties farther away	(?)✓	(+)	(?)✓
(4) Heterogeneity of peripheral connection effect among initially larger peripheral markets	(+)✓	(?)	(?)
(5) Heterogeneity of peripheral connection effect among counties with higher initial trade costs to the metropolitan city regions	(+)✓	(?)	(?)
(6) Heterogeneity of peripheral connection effect among initially more industrial, more skilled, more urban, or richer counties	(?)✓	(?)✓	(-)

(-) and (+) indicate predictions of a negative or a positive effect. (?) indicates either no or an ambiguous prediction consistent with a non-significant estimation result. ✓ indicates significant empirical evidence in favor of the prediction as found in Tables 3, 6, and 7 and Figure 4 of the paper. See Section 5 for discussion, and the Online Appendix for a simple core-periphery model of trade integration based on Helpman and Krugman (1985).

Table 6: **Did Connections among Peripheral Counties Lead to Urbanization and Industrial Decentralization?**

Dependent Variable:	Change ln(UrbPop) 1997-06	Change ln(Urb/Pop) 1997-06	Change ln(IndGVA) 1997-06			Change ln(GDP) 1997-06			Change ln(GovRevenue) 1997-06		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Connect	0.0350	0.0137	-0.297***	-0.262**		-0.203**	-0.193**		-0.257***	-0.253***	
	(0.0953)	(0.0925)	(0.108)	(0.113)		(0.0886)	(0.0919)		(0.0996)	(0.0961)	
Neighbor				0.153			0.0907			0.0535	
				(0.214)			(0.132)			(0.195)	
lnDistHwy					0.113*			0.0845*			0.177***
					(0.0615)			(0.0480)			(0.0667)
First Stage F-Stat	13.374	13.374	18.886	5.016	13.852	17.425	4.840	12.989	19.055	5.383	13.879
Obs	1,072	1,072	1,280	1,280	1,280	1,272	1,272	1,272	1,285	1,285	1,285

All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage estimates using the least cost path and the Euclidean spanning tree networks to instrument for NTHS connections, neighboring peripheral counties, or distance to the nearest NTHS segment. Columns 1 and 2 report connection effects on peripheral county changes in log urban population and urbanization respectively. Neighbor indicates peripheral counties neighboring a connected peripheral county. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 7: **Testing the Heterogeneity of Peripheral Connection Effects**

Dependent Variable:	Change ln(IndGVA) 1997-2006					Change ln(GDP) 1997-2006				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Panel A: Binary										
Connect	-0.304**	-4.281***	-4.236***	-3.147	-6.001**	-0.177*	-3.571***	-3.483***	-3.218**	-4.597**
	(0.145)	(1.569)	(1.620)	(2.146)	(2.575)	(0.0942)	(1.011)	(1.023)	(1.533)	(1.848)
Connect*ln(DistNode)		0.748***	0.740***	0.784***	0.891*		0.636***	0.626***	0.649***	0.759**
		(0.270)	(0.277)	(0.267)	(0.472)		(0.172)	(0.172)	(0.170)	(0.312)
Connect*Emp90Dum		0.450*	0.473*	0.468*	0.689		0.404**	0.485***	0.412**	0.642**
		(0.255)	(0.253)	(0.262)	(0.444)		(0.196)	(0.184)	(0.193)	(0.315)
Connect*City90			-0.113					-0.267		
			(0.403)					(0.211)		
Connect*PrefCap90			0.153					0.182		
			(0.374)					(0.245)		
Connect*Educ90				-2.901					-1.216	
				(3.956)					(2.698)	
Connect*AgShare90				-1.221					-0.331	
				(2.038)					(1.305)	
Connect*LocGovRevCap90					0.191					0.0490
					(0.215)					(0.132)
Obs	1280	1280	1280	1280	1020	1272	1272	1272	1272	1024
First stage F-Stat	29.966	3.462	1.765	2.601	1.317	27.972	4.724	1.717	2.517	1.553
Panel B: log Distance										
lnDistHwy	0.0954	1.465***	1.495***	1.067	1.689***	0.0639	1.105***	1.109***	0.861*	1.238***
	(0.0674)	(0.455)	(0.463)	(0.669)	(0.562)	(0.0434)	(0.318)	(0.324)	(0.507)	(0.437)
lnDistHwy *ln(DistNode)		-0.236***	-0.239***	-0.245***	-0.239***		-0.181***	-0.181***	-0.180***	-0.184***
		(0.0748)	(0.0762)	(0.0737)	(0.0821)		(0.0494)	(0.0501)	(0.0502)	(0.0619)
lnDistHwy*Emp90Dum		-0.266***	-0.250***	-0.250***	-0.258**		-0.192***	-0.188***	-0.184***	-0.192**
		(0.0823)	(0.0823)	(0.0821)	(0.111)		(0.0693)	(0.0673)	(0.0675)	(0.0833)
lnDistHwy*City90			-0.0743					-0.0291		
			(0.103)					(0.0689)		
lnDistHwy*PrefCap90			0.0409					0.0321		
			(0.105)					(0.0701)		
lnDistHwy*Educ90				0.776					0.468	
				(0.895)					(0.610)	
lnDistHwy*AgShare90				0.459					0.230	
				(0.424)					(0.240)	
lnDistHwy*LocGovRevCap90					-0.0485					-0.0235
					(0.0301)					(0.0164)
Obs	1280	1280	1280	1280	1020	1272	1272	1272	1272	1024
First stage F-Stat	22.367	4.649	2.720	2.286	2.004	21.698	4.842	2.876	2.549	2.355

All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage estimates using the LCP spanning tree to instrument for NTHS connections as well as their reported interaction terms. lnDistNode is log county distance to the nearest targeted city node. Emp90Dum is a dummy for counties with above mean levels of county employment in 1990. City90, PrefCap90, Educ90, AgShare90, and LocGovRevCap90 are 1990 city status, prefecture capital status, skilled labor shares, agricultural specialization, and log government revenue per capita. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Appendix - Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System

Benjamin Faber*

17 July 2013

Abstract

This appendix proceeds in five sections. Appendix 1 presents a simple three region core-periphery model based on the home market effect in Helpman and Krugman (1985). Appendix 2 presents estimation results concerning the proportion and characterization of complier counties that drive the local average connection effects estimated in the paper. Appendix 3 presents additional estimation and robustness results for both average NTHS connection effects and their heterogeneity with respect to pre-existing county characteristics. Appendix 4 describes the datasets and construction of variables. Appendix 5 describes the construction of the least cost path and Euclidean spanning tree networks.

*Department of Economics, University of California Berkeley; Email: benfaber@econ.berkeley.edu.

Appendix 1: Core-Periphery Model

The model is based on Helpman and Krugman (1985) and adapted to a setting with multiple and ex ante asymmetric regions. In addition, I introduce capital as an input to industrial production and allow this factor to be mobile across regions as in Martin and Rogers (1995). This serves to adapt the original cross-country model without factor mobility to a within country setting with partial factor mobility without altering the original set of microeconomic forces at play. The exposition closely follows the footlose capital model in Baldwin et al. (2003).

The economy is populated by a continuum of agents who are distributed over R regions. There are two sectors of production, labeled agriculture (A) and industry (M), and two factors of production labeled labor (L) and capital (K). The former is assumed to be immobile across regions, while the latter is mobile. Mobile stocks of capital are owned by workers, and returns to capital are repatriated across regions.

Preferences

The representative consumer in each region has two-tier preferences, where the upper tier is a Cobb-Douglas nest of consumption of agriculture (which will be the numeraire good) and a composite of industrial varieties. Industrial goods enter as a constant elasticity of substitution (CES) sub-utility function defined over a continuum of industrial varieties i(i=1,2,...N). Consumer utility in region j(j=1,2,...R) is given by:

$$U_j = C_{Mj}^{\mu} C_{Aj}^{1-\mu} \qquad C_{Mj} = \left(\int_{i=0}^{N} c_{ij}^{1-1/\sigma} \right)^{\frac{1}{1-1/\sigma}} \qquad 0 < \mu_{hk} < 1 < \sigma$$

C_{Mj} and C_{Aj} are consumption of industry and agriculture in region j respectively, c_{ij} is consumption of manufacturing variety i in region j, μ is the expenditure share on industry, and σ is the elasticity of substitution between varieties. Standard utility maximization yields a constant division of expenditure between sectors and CES demand for an industrial variety i in region j:

$$c_{ij} = \frac{p_{ij}^{-\sigma}}{\int_{i=0}^{N} p_{ij}^{1-\sigma} di} \mu Y_j$$

Y_j is total regional factor income of labor (L_j) and capital (K_j), with wage rate w_j and capital return π_j:

$$Y_j = w_j L_j + \pi_j K_j$$

Technology

The numeraire agricultural sector requires a_A units of labor to make one unit of A. It is subject to perfect competition, constant returns to scale and faces no trade costs. Marginal cost pricing implies that $p_{Aj} = a_A w_j$ and costless trade equalizes prices and wages across regions so that $p_{Aj} = p_A$ and $w_j = w$ as long as some positive fraction of A is produced in every region.[1] The industrial sector M is subject to increasing returns, Dixit-Stiglitz monopolistic competition and iceberg trade costs. Each firm of a continuum of industrial producers requires one fixed cost unit of capital K, and a_M units of L to produce a unit of M. This implies a cost function $\pi + wa_M x$, where x is firm level output. It is costless to ship industrial goods within a region, but $\tau_{jk} - 1$ units of the good are used up in transportation between two regions j and k. It is assumed that τ_{jk}=τ_{kj}. It proves convenient to define $\phi_{jk} = \tau_{jk}^{1-\sigma}$ as the "freeness" of trade ranging from 0 (prohibitive costs) to 1 (costless trade). Dixit-Stiglitz monopolistic competition and the above demand imply that mill pricing is optimal for industrial firms, so that the price ratio of a variety in an export region k over its local market price in j is τ_{jk}. For a variety i produced in region j but also sold in another region k this is:

$$p_{ij} = \frac{wa_M}{1 - 1/\sigma}, \quad p_{ik} = \tau_{ik} \frac{wa_M}{1 - 1/\sigma}$$

Equilibrium

Because the marginal cost of industrial firms depends on the immobile factor whose price is pinned down by costless trade in the numeraire sector, industrial f.o.b. prices are equalized across regions and consumer prices differ only by transport costs. As capital enters as fixed cost component in industrial production, this also implies that capital returns are equal to the operating profit of a typical variety. Under Dixit-Stiglitz competition, this is equal to the value of sales divided by σ: $\pi = px/\sigma$. Normalizing the price of agriculture to be the numeraire and choosing units of A such that $p_A = a_A = w = 1$, we can use demand and mill pricing to solve for the equilibrium returns to the mobile factor:[2]

$$\pi_j = \left(\sum_k \frac{\phi_{jk} S_{Yk}}{\sum_m \phi_{mk} S_{Nm}} \right) \frac{\mu}{\sigma} \frac{Y}{K}$$

S_Y represents regional shares of total expenditure, and S_N are regional shares of the

[1] Factor price equalization in the Helpman and Krugman (1985) structure implies a focus on firm relocation as the adjustment channel to equalize profits across regions. An alternative adjustment channel arises in the absence of factor price equalization through wage adjustments across regions (captured by so called wage equations in this literature). See Chapter 12 in Combes et al. (2008) for a discussion and formalization of these alternative adjustment channels.

[2] Industrial sales in market j become $\sum_k p_{jk} x = \left(\sum_k \frac{\phi_{jk} S_{Yk}}{\sum_m \phi_{mk} S_{Nm}} \right) \mu \frac{Y}{K}$. Capital returns in market j are thus $\pi_j = \frac{1}{\sigma} \sum_k p_{jk} x$.

mass of total industrial varieties. Y and K stand for total expenditure and the total capital endowment across all regions respectively. Given repatriation of capital returns to immobile owners, regional expenditure shares are a deterministic function over regional shares of capital owners and labor endowments, S_K and S_L respectively:

$$S_{Yj} = \left(1 - \frac{\mu}{\sigma}\right) S_{Lj} + \frac{\mu}{\sigma} S_{Kj}$$

Because capital is freely mobile across regions, there are two possible types of equilibria: core-periphery outcomes where S_N can be 0 or 1, and interior location equilibria. Given all regions maintain some positive fraction of industrial activity, capital returns are equalized so that the long run equilibrium location condition is given by $\pi_j = \pi$ for $0 < S_{Nj} < 1$. The profit equation coupled with inter-regional profit equalization yield a system of R equations that can be solved for an R∗1 vector of regional industrial production shares as a function of an R∗R bilateral trade cost matrix and an R∗1 vector of regional expenditure shares that are in turn determined by regional endowments.[3]

Predictions

The empirical estimations of the paper are based on the comparison of changes of economic outcomes among peripheral county regions that were connected to new NTHS routes relative to non-connected peripheral counties. Given that in general equilibrium it would be a strong assumption that non-connected regions are not affected at all by the network, the most basic policy scenario thus requires at least three regions. Consider two initially identical peripheral regions and one larger metropolitan core region, denoted by superscripts P1, P2 and C respectively, that are identical in terms of tastes, technology, and initial bilateral trade costs. Geometrically, one can think of this scenario as three regions located on the endpoints of an equilateral triangle. The profit equation in the first peripheral region becomes:

$$\pi^{P1} = \left(\frac{S_Y^{P1}}{S_N^{P1} + \phi\left(1 - S_N^{P1}\right)} + \phi \frac{S_Y^{C}}{S_N^{C} + \phi\left(1 - S_N^{C}\right)} + \phi \frac{S_Y^{P2}}{S_N^{P2} + \phi\left(1 - S_N^{P2}\right)} \right) \frac{\mu}{\sigma} \frac{Y}{K}$$

Profits in the core region are isomorphic, and profits in the second peripheral region are given by $\pi^{P2} = 1 - \pi^{P1} - \pi^{C}$. Initial peripheral symmetry implies that $S_Y^{P1} = S_Y^{P2}$, and ϕ is the identical bilateral trade freeness between all three regions at an initial period. We now introduce asymmetric trade integration in the most convenient way. Let $\alpha\phi$ denote the bilateral trade freeness between peripheral region 1 and the core region after a negative bilateral trade cost shock, while ϕ is the unchanged initial trade freeness between all regions. Initially, α=1, while after the trade cost shock takes effect, α is in the range $1 < \alpha <$

[3]Notice that total profits must be equal to total payments to capital. Also, $\frac{\mu}{\sigma}Y$ must equal to profits. This leaves us with R-1 independent equations. Using those and the fact that the sum of the shares S_N must sum to one, we can solve for S_N in all markets as a function of exogenous trad costs and exogenous expenditure shares.

$(1/\phi)$. Using peripheral symmetry $S_Y^{P1} = S_Y^{P2}$, introducing the asymmetric trade cost shock $\left(\phi^{P1} = \alpha\phi\right)$, and solving for the equilibrium difference of industrial activity between connected and non-connected peripheral regions subject to profit equalization $\pi^{P1} = \pi^{P2} = \pi^C$, we get:

$$S_N^{P1} - S_N^{P2} = \left(\left(\frac{1}{2} - \frac{3}{2}S_Y^C\right)\frac{(\alpha-1)\,\phi}{1-\alpha\phi} + 1\right)\frac{1+\alpha\phi-2\phi^2}{(1-\phi)\,(1+\alpha\phi-2\phi)} - \frac{3\phi}{1+\alpha\phi-2\phi} - 1$$

This provides a closed form solution for peripheral differences in industrial activity as a function of relative market sizes, initial levels of trade costs, and the degree of asymmetric trade integration. At the initial α=1 position, perfect symmetry between peripheral regions leads to $S_N^{P1} - S_N^{P2} = 0$. The question is what happens to industrial production in the connected peripheral county relative to the non-connected one after the trade cost shock materializes. The derivative of interest is $\frac{\partial\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha}$. The sign of this derivative in principle depends on the extent of the pre-existing core-periphery gradient summarized in S_Y^C, the level of pre-existing trade integration ϕ, as well as the extent of asymmetric trade integration captured by α. It is clear from the expression above that for any given scenario of core-periphery integration $1 < \alpha < (1/\phi)$ and initial trade costs ϕ, the difference in industrial production shares between the integrating and the non-integrating periphery becomes more negative as the core-periphery size asymmetry (summarized by S_Y^C) increases. Using this insight, one can solve for the necessary degree of the core-periphery gradient at which $\frac{\partial\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha} < 0$ holds for any combination of initial trade costs and trade cost shock asymmetry. From the expression above, this is the case as long as the metropolitan region is at least twice the size of an individual peripheral region.

Prediction 1: *Falling trade costs between a sufficiently uneven core-periphery pair of regions lead to a reduction of industrial production in the integrating periphery relative to a non-integrating peripheral control region.*

Descriptive statistics in Table 1 of the paper indicate that the model's size asymmetry threshold is clearly exceeded when comparing non-targeted peripheral counties to the targeted metropolitan city regions. The intuition behind this results is as follows. Equilibrium profits are a positive function of access to consumer expenditure, and decreasing in access to competing industrial producers. The former enters as agglomeration force and the latter as a dispersion force. On one hand, lower trade costs decrease the relative disadvantage of higher product market competition in the larger market because the relative increase in competition is stronger for the smaller region. On the other hand, lower trade costs also decrease the market access advantage of the larger region because the relative increase in market access is stronger for the smaller region. The microfoundation of the home market channel is that falling trade costs attenuate the dispersion force at a faster rate than the agglomeration force.

The model makes additional predictions about county level changes to overall GDP

and agricultural output. Aggregate GDP moves in parallel to industrial output, but less than proportional because labor formerly used in industry remains productive in the region. Prediction 1 thus holds for aggregate production, but we expect a lower point estimate on the elasticity of peripheral GDP to trade cost reductions compared to industrial activity. Conversely, the reallocation of labor to the agricultural numeraire sector implies that falling trade costs have the opposite effect on agricultural output growth.[4]

Prediction 2: *The negative effect of integration holds, but to a lesser extent, for total regional production, and is reversed in sign for agricultural production.*

In addition to the predictions on the average effects of integration among peripheral counties, the richness of the empirical setting also allows to test how the home market channel should affect peripheral counties differently. The first cross-derivative prediction is that the home market effect should be more pronounced among peripheral counties whose initial level of trade costs vis-a-vis the core region is lower: $\frac{\partial^2\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha\partial\phi} < 0$. For a given trade cost reduction, the marginal effect on industrial and aggregate production should be more negative at higher initial levels of ϕ.

Prediction 3: *The negative effects of integration on industrial and total production are more pronounced among peripheral regions with initially lower trade costs to the larger core region.*

This interaction effect is related to what the trade literature has referred to as home market magnification (Baldwin et al., 2003). The intuition is that falling trade costs attenuate the peripheral location advantage of less market crowding at a faster rate than the metropolitan market access advantage, so that at lower initial trade costs between core and periphery a given trade cost reduction will require a larger relocation of industrial production to equalize the rate of capital return. The second cross-derivative prediction is that, holding initial trade freeness constant, the home market effect is stronger among peripheral counties whose size differential vis-a-vis the core is more pronounced: $\frac{\partial^2\left(S_N^{P1}-S_N^{P2}\right)}{\partial\alpha\partial S_Y^C} < 0$.

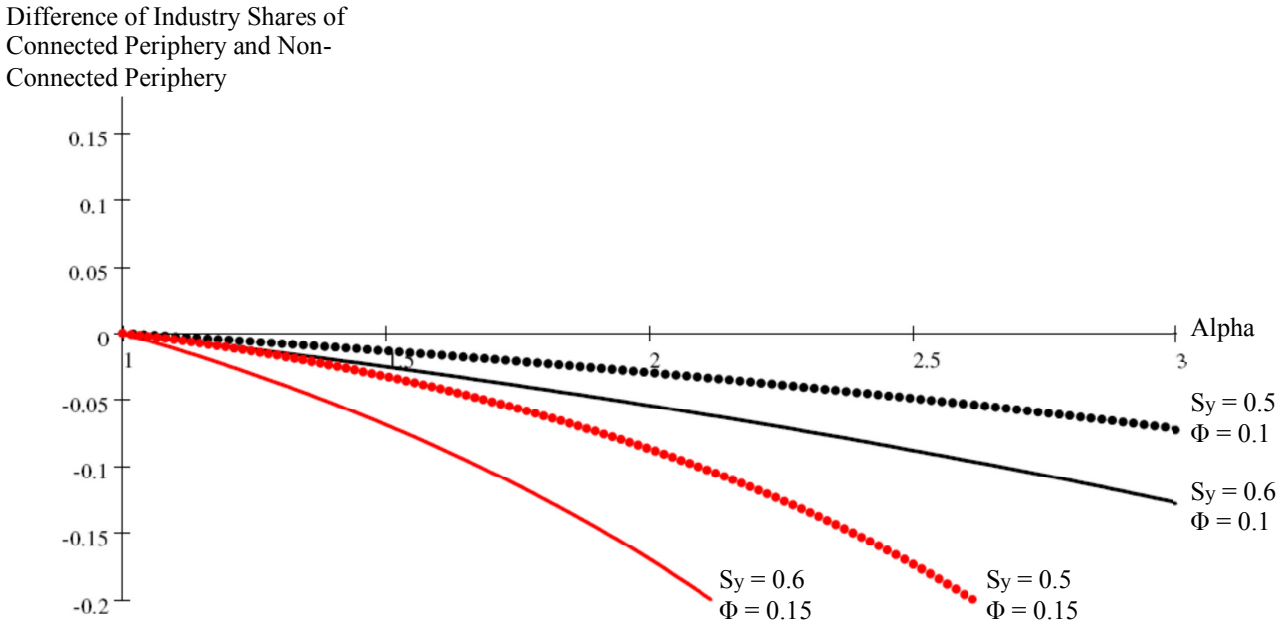
Prediction 4: *The negative effects of integration on industrial and total production are more pronounced among peripheral regions with an initially stronger market size differential to the core region.*

The prediction that the home market channel should operate more strongly among smaller peripheral regions is also intuitive. Falling trade costs weaken the dispersion force at a faster rate than the agglomeration force, so that for a larger core-periphery size gradient, and thus higher initial levels of agglomeration and dispersion forces, a given trade cost reduction requires more industrial concentration in the core to equalize profits.

Figure 1.1 provides a graphical illustration of Predictions1-4.

[4]To see this more clearly, we can write differences in total production and differences in agricultural output as a function of differences in industrial production as follows: $GDP^{P1} - GDP^{P2} = \left(S_N^{P1} - S_N^{P2}\right)\frac{\mu}{\sigma}Y + \left(S_L^{P1} - S_L^{P2}\right)L$ and $GDP_{Ag}^{P1} - GDP_{Ag}^{P2} = \left(S_L^{P1} - S_L^{P2}\right)L - \left(S_N^{P1} - S_N^{P2}\right)\mu\frac{\sigma-1}{\sigma}Y$.

Figure 1.1: Plotting Predictions 1-4



The x-axis displays the degree to which the policy treatment lowers the trade cost of the connected peripheral region to the core region relative to the non-connected peripheral region. The axis starts at the initially identical trade freeness *vis-à-vis* the metropolitan core (α=1). The y-axis displays the difference of industrial production shares between the connected and the non-connected peripheral regions. Sy is the share of total expenditure located in the metropolitan region, and Φ is the initial trade freeness parameter between all regions.

Appendix 2: Local Average Connection Effects

The instrumental variable estimates presented in the paper represent the local average treatment effect (LATE) of network connections among peripheral counties whose treatment status is affected by location along the all-China least cost spanning tree network. The evaluation literature refers to the latter category as "compliers", as opposed to "always taker" counties that were connected despite their location away from the spanning tree paths.[5]

Descriptive statistics and the pattern of coefficient estimates discussed in the paper suggest that planners targeted economically prosperous counties on the way between targeted city regions. In this empirical context, the concern addressed in this additional set of estimations is that least cost spanning tree location might have affected actual highway placements only for a subset of remote and economically stagnant counties on the way between targeted nodes, so that the estimated local average NTHS connection effects might systematically differ from population average effects.

While it is not possible to identify the complier status of individual counties in the county sample, it is possible to estimate the proportion of compliers among the treated counties

[5] An implicit assumption is that there are no "defiers" in this context, as location along advantageous construction routes does not cause counties not to be on the network.

as well as their observable characteristics (Abadie, 2003; Angrist and Pischke, 2008). The proportion of compliers among all actually treated NTHS counties is given by:

$$P\left(C_{1i} > C_{0i} | C_i = 1\right) = \frac{P\left(C_i = 1 | C_i > C_{0i}\right) P\left(C_{1i} > C_{0i}\right)}{P\left(C_i = 1\right)}$$

$$= \frac{P\left(z_i = 1\right)\left(E\left(C_i | z_i = 1\right) - E\left(C_i | z_i = 1\right)\right)}{P\left(C_i = 1\right)}$$

where C_i is actual NTHS connection status of county i, C_{1i} and C_{0i} are the connection status in cases where the instrument predicts treatment or not, P and E are probability and expectation operators, and z_i is the treatment value of county i as predicted by the instrument. The second equality makes use of the two facts that the total size of the complier group is given by the Wald first stage, and that by independence $P\left(z_i = 1 | C_{1i} > C_{01}\right) = P\left(z_i = 1\right)$. The proportion of compliers among treated counties can then be expressed as the product of the first stage estimate and the proportion of predicted treatments, divided by the proportion of actually treated counties. As presented in the first Column of the Table 2.1, this proportion is estimated to be 22% for both least cost path as well as the Euclidean spanning tree instruments.

The critical question is to what extent these complying counties could be systematically different from the rest of the treated counties. In the case of binary treatments and binary dependent characteristics we know that the relative likelihood of compliers falling into the binary observable category is given by the ratio of the first stage Wald estimated for a particular subgroup over the full sample first stage estimate. To this end, Columns 2-6 of Table 2.1 report first stage point estimates in stated order for counties with above mean 1997 levels of population, urban population, the share of urban population, GDP, and GDP per capita.

If the concern was true that the estimated local average treatment effects are unrepresentative of the population average effects, then we would expect the first stage predictive power of the instruments to differ significantly across observable pre-existing county characteristics. As discussed above, this would constitute evidence of significantly different likelihoods of observables among always takers as opposed to complying counties. In particular, in the present setting one would be concerned that the instrumental variable connection predictors would have a lower estimated effect on actual NTHS route placements among the large, urbanized, and rich county groups represented in Columns 2-6.

The reported results provide evidence against this concern. In particular, the first stage point estimates do not significantly differ from the full sample first stage estimate when estimated for different subsamples of counties as indicated across the columns. Figure 2.1 then takes a closer cartographic inspection of actual as opposed to predicted route placements to offer two plausible explanations for the absence of clear observable differences

between compliers and always takers.

The figure depicts two snapshots at the county level of the PR China in which counties are color coded according to their nominal levels of GDP in 1997. Both cases compare actual NTHS route placements to predictions from the least cost path spanning tree instrument. Case A illustrates the first point. It is evident that the least cost path algorithm is subject to prediction errors for both bigger urban and smaller rural counties, even in cases where no obvious incentive for deviations from the least cost path is evident. This has to do with the fact that entire bilateral route segments might have been differently picked by the instrument as opposed to NTHS routes, and most importantly, because planners built many more bilateral routes than the minimum number of edges that the spanning tree algorithm picks.

Case B illustrates the second point. When planners do seem to have deviated from the least cost path for an obvious reason (e.g. connection of a prefecture level capital city on the way as indicated in the figure), then this deviation for one "important" county leads to prediction errors for several "unimportant" counties on the way to this target. Both of these features that are evident from the GIS snapshots tend to work against any systematic correlation between the predictive power of the instrument in the first stage and the potential heterogeneity of the highway effect.

Table 2.1: Estimated Proportion of Compliers and Relative Likelihoods of Observable Characteristics

	(1) Full Sample	(2) Pop 97	(3) Urban Pop 97	(4) %Urban Pop 97	(5) GDP 97	(6) GDP Cap 97
Panel A: LCP IV						
Connect 1st Stage Point Estimate	0.418***	0.383***	0.432***	0.494***	0.399***	0.433***
	(0.0601)	(0.0821)	(0.0704)	(0.0599)	(0.0873)	(0.0869)
F-Statistic p-value [Coef=0.418]		0.677	0.839	0.214	0.832	0.864
Obs	1367	650	662	633	673	664
Estimated Proportion of Compliers Among Treated Counties	0.222					
Panel B: Euclid IV						
Connect 1st Stage Point Estimate	0.314***	0.354***	0.375***	0.328***	0.365***	0.337***
	(0.0492)	(0.0690)	(0.0822)	(0.0776)	(0.0784)	(0.0712)
F-Statistic p-value [Coef=0.314]		0.567	0.462	0.860	0.521	0.750
Obs	1367	650	662	633	673	664
Estimated Proportion of Compliers Among Treated Counties	0.221					

Each point estimate stems from a separate regression. The table presents first stage point estimates for regressions of binary NTHS connections on spanning tree connections and province fixed effects across different county samples. All regressions include province fixed effects. LCP IV stands for the least cost path spanning tree instrument. Euclid IV stands for the straight line spanning tree instrument. The first column presents the full sample first stage estimate. The following columns (in stated order) present this estimate for counties with above median 1997 levels of population, urban population, shares of urban population, GDP, and GDP per capita. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Figure 2.1: Cartographic Inspection of the Instrument

Case A **Case B**

The network in red color depicts actual NTHS expressway routes. The network in black color depicts the least cost path spanning tree network. Crosses indicate targeted metropolitan nodes. Counties are color coded according to their nominal levels of GDP in 1997, where darker colors represent higher values. Striped areas indicate missing 1997 GDP data.

Appendix 3: Additional Estimation and Robustness Results

This section presents additional estimation and robustness results.

Table 3.1 presents estimation results for a series of additional robustness specifications concerning the average NTHS connection effects on county growth discussed in the paper. The first row of results reproduces the baseline estimates of the NTHS connection effect for industrial output growth, non-agricultural output growth, GDP growth, and local government revenue growth for the preferred specification with both instruments and the full set of pre-existing county controls.

The second row of results addresses the concern that the geographical characteristics used in the construction of the least cost path instrument could directly affect county growth and thereby lead to a violation of the exclusion restriction. To address this concern, the table reports results after including the average terrain slope gradient, the percentage of water coverage, the percentage of wetlands coverage, and the percentage of developed land coverage as additional county controls. The NTHS connection effect estimates are unaffected by the inclusion of these additional controls, and show a very slight increase.

The third row of results addresses the concern that location along least costly paths might be subject to stronger endogeneity concerns in mountainous provinces where valleys provide natural advantages for settlements and economic development. The presented results are estimated after excluding the mountain provinces of Gansu, Qinghai, Sichuan, Tibet, and Xinjiang. The exclusion of these regions also address the concern that due to the mountainous terrain a new long distance railway route was built following closely the route of the NTHS between Golmud and Lasa over the same period. The NTHS connection effects are confirmed in sign and statistical significance for all dependent variables when estimated on the restricted county sample.

The fourth row of results addresses the concern that least costly route locations between the major city regions of China are likely to be correlated with historical trade routes. To this end, I obtained geo-referenced routes for the Northern and the Southern routes of the trans-Asian Silk road from the Old World Trade Routes (OWTRAD) Project.[6] The Southern routes of the Silk Road are sometimes referred to as the Tea Horse Road instead. Figure 3.1 provides an illustration of the NTHS network and the Silk Road routes. Reported estimation results include the log distance to the nearest Silk Road segment as an additional county control. The baseline NTHS coefficients are hardly affected by the inclusion of this additional control, indicating that the baseline controls for pre-existing political and economic characteristics have effectively captured county proximity to historical trade routes.

The fifth row addresses the concern that the spanning tree instruments might be picking

[6] See www.ciolek.com/owtrad.html.

up county locations with preferential market access positions in the preceding period. These locations could be especially well suited for the process of urbanization and decentralization that Baum Snow et al. (2012) have found to be the case for prefecture level central city districts in China during the 1990s. To this end, I compute each county's log market potential in 1997 following Harris (1954) as the distance weighted sum of all other county seat populations in China where the weights are equal to inverse distances. The fact that the baseline point estimates are virtually unaffected by the inclusion of this additional control suggests that omitted differences in pre-existing market access are not confounding the IV estimates.

The final row of results addresses the concern that the initial period output levels among NTHS connected peripheral counties might have been inflated by construction activity already underway in 1997. The concern is that the significant negative effects are driven in part by this inflation of the initial levels of economic activity. To address this concern, I include a dummy indicator for road construction underway in 1997 that I collect from the 1998 Atlas source described in the Data Appendix below. The inclusion of this additional county control hardly affects the baseline point estimates of the NTHS connection effects, indicating that road construction activity underway in 1997 did not lead to a spurious negative growth effect among NTHS connected counties.

Table 3.2 reports results of the placebo specification with industrial production instead of local government revenue growth as outcome variable. This implies a substantial reduction in the number of counties, as only a selected subsample of counties consistently report production data starting in 1990-2006. The table reports instrumental variable results using both LCP and Euclidean spanning tree instruments to estimate the effects of NTHS routes both in the period before the expressways had been built and afterward on the identical county sample. The placebo result appears to hold for industrial production in this restricted subsample. In particular, no significant effect is found in the preceding period, whereas the point estimate changes to negative and statistically significant during the NTHS period. These findings confirm the main results reported in the paper in terms of local government revenue growth.

Table 3.3 reports additional results of the estimations on interaction effects comparing two stage least squares (2SLS) estimates with limited information maximum likelihood (LIML) estimates when using both spanning tree networks to instrument for NTHS connections and its interaction terms with respect to pre-existing county characteristics. The concern addressed in these estimations is that the drop in the first stage F-statistics among specifications with interaction effects could lead to weak instrument bias.

Panel A of the table reports 2SLS results, and Panel B reports LIML results for identical specifications. As was the case for the just identified specifications in the paper, the fact that the point estimates on the interaction terms remain stable across the different specifications across Columns 2-4 with varying levels of first stage F-statistics offers a first reassurance against the concern of weak instrument bias. Furthermore, the table indicates that LIML

coefficient estimates on the main NTHS effect and its interaction terms are slightly higher across both dependent variables (industrial output growth and GDP growth), as well as across all reported specifications. Given that the LIML estimator has been shown to be less affected by weak instrument bias, this finding provides a second robustness check against this concern.[7]

Table 3.1: Additional Robustness Specifications

Robustness specifications		(1) Change ln(IndGVA) 1997-06	(2) Change ln(NonAgGVA) 1997-06	(3) Change ln(GDP) 1997-06	(4) Change ln(GovRevenue) 1997-06
Baseline estimates	Connect	-0.297***	-0.268***	-0.203**	-0.257***
		(0.108)	(0.0969)	(0.0886)	(0.0996)
	Obs	1280	1262	1272	1285
Control for direct effects of geographical variables used in least cost path construction	Connect	-0.308***	-0.277***	-0.209**	-0.236**
		(0.110)	(0.0988)	(0.0908)	(0.100)
	Obs	1280	1262	1272	1285
Exclude mountain provinces and Golmud-Lasa Railway	Connect	-0.363***	-0.369***	-0.297***	-0.216*
		(0.122)	(0.100)	(0.0940)	(0.123)
	Obs	1043	1032	1040	1039
Control for log distance to historical trade routes	Connect	-0.300***	-0.272***	-0.206**	-0.251**
		(0.106)	(0.0951)	(0.0874)	(0.102)
	Obs	1280	1262	1272	1285
Control for market access in 1997	Connect	-0.289**	-0.255***	-0.208**	-0.262***
		(0.114)	(0.0961)	(0.0879)	(0.0921)
	Obs	1280	1262	1272	1285
Control for construction underway in 1997	Connect	-0.296***	-0.266***	-0.201**	-0.256**
		(0.111)	(0.0969)	(0.0897)	(0.0996)
	Obs	1280	1262	1272	1285

Each point estimate stems from a separate regression. All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage IV estimates using the least cost path and the Euclidean spanning tree networks as instruments for NTHS connections. The dependent variables in order of the columns as listed are log changes of county level industrial gross value added, non-agricultural gross value added, total GDP, and local government revenue. Controls for geographical characteristics used in the construction of the least cost path spanning tree instrument are average county slope, and county percentage of wetland water, or developed coverage. Mountainous provinces refer to Gansu, Qinghai, Sichuan, Tibet, and Xinjiang. Historical trade routes are the Northern and Southern routes of the Silk Road (see map presented below). The control for market access in 1997 is the log of a county's market potential according to Harris (1954), i.e. it is the weighted sum of all other county seat populations in China, where the weight is equal to the inverse of bilateral distances. Control for construction in 1997 refers to a dummy variable indicating all counties with reported "under construction" expressway routes in 1997. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

[7]See for example Angrist and Pischke (2008, Section 4.6) for a discussion 2SLS and LIML estimates in the context of weak instrument concerns.

Figure 3.1: The Northern and Southern Routes of the Silk Road

The network in red color depicts the completed NTHS network in 2007. The green routes represent the Northern and Southern Routes of the Silk Road.

Table 3.2: Placebo Specification on Subsample with Longer Series of Production Data

	(1)	(2)
Dependent Variable:	Both IVs	Both IVs
Change ln(IndGVA)	1990-97	1997-06
Connect	0.0976	-0.219**
	(0.142)	(0.102)
Obs	513	513
First stage F-Stat	39.271	39.271

The table reports 2nd stage estimation results using both LCP and Euclidean spanning tree instruments to predict NTHS connections. The first column reports results on industrial production growth for the period preceding the NTHS network, and the second column reports results for the paper's estimation period 1997-2006. Regressions include province fixed effects and county controls. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Table 3.3: Comparing 2SLS and LIML Estimations of Interaction Effects

Dependent Variable:	Change ln(IndGVA) 1997-2006					Change ln(GDP) 1997-2006				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Panel A: 2SLS										
Connect	-0.297***	-3.876***	-3.649***	-2.783	-5.856**	-0.203**	-3.496***	-3.389***	-3.004**	-4.729***
	(0.108)	(1.333)	(1.295)	(1.811)	(2.431)	(0.0886)	(0.948)	(0.966)	(1.389)	(1.762)
Connect*ln(DistNode)		0.680***	0.639***	0.744***	0.917**		0.623***	0.605***	0.641***	0.803***
		(0.232)	(0.225)	(0.246)	(0.441)		(0.161)	(0.164)	(0.162)	(0.306)
Connect*Emp90Dum		0.400	0.378	0.467*	0.749*		0.396**	0.389**	0.394**	0.698**
		(0.247)	(0.235)	(0.271)	(0.440)		(0.196)	(0.193)	(0.197)	(0.338)
Obs	1280	1280	1280	1280	1020	1272	1272	1272	1272	1024
First stage F-Stat	18.886	2.047	2.004	1.718	1.204	17.425	2.147	2.149	1.510	1.267
Panel B: LIML										
Connect	-0.297***	-3.914***	-3.706***	-2.872	-6.035**	-0.205**	-3.540***	-3.502***	-3.039**	-4.844***
	(0.108)	(1.353)	(1.327)	(1.878)	(2.556)	(0.0893)	(0.970)	(1.025)	(1.412)	(1.838)
Connect*ln(DistNode)		0.686***	0.649***	0.767***	0.946**		0.631***	0.624***	0.648***	0.822**
		(0.235)	(0.231)	(0.259)	(0.462)		(0.165)	(0.174)	(0.166)	(0.319)
Connect*Emp90Dum		0.406	0.388	0.489*	0.778*		0.404**	0.409**	0.402**	0.718**
		(0.250)	(0.240)	(0.284)	(0.461)		(0.200)	(0.204)	(0.201)	(0.351)
Obs	1280	1280	1280	1280	1020	1272	1272	1272	1272	1024
First stage F-Stat	18.886	2.047	2.004	1.718	1.204	17.425	2.147	2.149	1.510	1.267

All regressions include province fixed effects and a full set of county controls. Reported results are 2nd stage estimates using the least cost path and the Euclidean spanning tree networks to instrument for NTHS connections as well as their reported interaction terms. lnDistNode is log county distance to the nearest targeted city node. As in Table 7 of the paper, Columns 2 do not include additional interaction effects. Columns 3 introduce additional interaction terms with respet to 1990 city status and prefecture level capital status. Columns 4 include additional interactions with Educ90 and AgShare90. Columns 5 include an additional interaction with respect to log 1990 government revenue per capita. Standard errors are clustered at the province level and stated in parentheses below point estimates. ***1%, **5%, and *10% significance levels.

Appendix 4: Data Appendix

GIS Data

Geo-referenced administrative boundary data for the year 1999 was obtained from the ACASIAN Data Center at Griffith University in Brisbane, Australia. These data provide a county-level geographical information system (GIS) dividing the surface of mainland China into 2341 county level administrative units, 349 prefectures, and 33 provinces. Chinese administrative units at the county level are subdivided into county level cities (shi), counties (xian), and urban wards of prefecture level cities (shixiaqu).

Administrative units in China are identified by a system of guo biao codes that allows the matching of records across the GIS and socioeconomic datasets. In addition to guo biao codes, the combination of prefecture and county names were used to double check the consistent matching of administrative units across the datasets. I use reported data on the county area under administration in km^2 from the Provincial Statistical Yearbook series to identify significant boundary changes over time. The historically consistent county sample for estimations on changes 1997-2006 are defined as counties without administrative area changes in excess of 5%. For the placebo falsification test that is estimated on the identical county sample for both the pre- and post-NTHS periods, 1990-1997 and 1997-2007, the same threshold is applied to changes for both periods.

Geo-referenced NTHS highway routes as well as Chinese transport network data were obtained from the ACASIAN Data Center. NTHS highway routes were digitized on the basis of a collection of high resolution road atlas sources published between 1998 and 2007 that is listed below.

(1) China Newest Public Road Atlas (1998), Ha Na Bin Map Publishing Company
(2) China Road Atlas (2002), Shadong Map Publishing Company
(3) China Public Road Atlas (2002), Shandong Map Publishing Company
(4) China Expressway Atlas (2003), People's Transport Press
(5) China Transportation Network Atlas (2003), Guangdong Map Publishing Company
(6) China Road Atlas (2003), Xue Yuan Map Publishing Company
(7) China Automobile Map (2003), China World Map Publishing Company
(8) Chinese People's Road Atlas (2005), Globe Publishing Company
(9) China Road Atlas (2007), Shandong Map Publishing Company

These atlas sources made it possible to classify NTHS segments into three categories: opened to traffic before mid-1997 (10% of NTHS), opened to traffic between mid-1997 and end of 2003 (81% of NTHS), and opened to traffic after the end of 2003 (9% of NTHS). In particular, Source (1) was used to digitize a baseline layer of NTHS routes that were in place by mid year in 1997, and Source (8) was used to digitize a baseline layer of NTHS routes that were in place by the end of 2003. These baseline route maps were then cross-referenced with route information provided in the remaining listed atlas sources. In cases where the

remaining atlas sources were at odds with the information of the baseline maps (i.e. routes present in 1997 but not in 2000 or thereafter, or routes present in 2003 but not thereafter), a decision was taken on the basis of the majority of sources (for the 1997 layer), or after tracking down highway openings through press releases on highway opening ceremonies for a small number of cases where Sources (8) and (9) were at odds.

Finally, land cover and elevation data that are used in the construction of least cost path highway routes were obtained from the US Geological Survey Digital Chart of the World (DCW) project, and complemented by higher resolution Chinese hydrology data from the ACASIAN data center. The higher resolution hydrology data from ACASIAN was used to assure that rivers were not interrupted by grid cells coded as mostly covered by land in the lower resolution raster data on land cover obtained from the DCW.

Socio-Economic Data

The Provincial Statistical Yearbook series report production approach county GDP broken up into primary, secondary, and tertiary gross value added. Value added is reported as gross output value less intermediate inputs and value added tax. Traditionally, construction is included together with manufacturing under the secondary industrial sector. The county level data are collected from local establishments under the supervision of the provincial governments, and the Provincial Statistical Yearbooks constitute a separate process of data collection from the national Statistical Yearbook series that is undertaken by the National Bureau of Statistics.

The reported production output data collected by local governments in principle cover the entirety of producing establishments located in the area of the county authority. This is in contrast to central government production statistics that are based on a cut-off of 5 million Yuan annual revenues for so called directly reporting industrial enterprises.[8] The data are collected by teams of local bureaucrats in the form of surveys that are filled out by the establishments located in the jurisdiction of the county.

The Provincial Statistical Yearbooks also provide local government revenues that are reported from the revenue accounts of local authorities. Government revenues mainly consist of industrial and commercial taxes (including value added tax) as well as corporate income taxes (Lin, 2009). The population records contained in the yearbook series refer to locally registered populations under the household registration system.

The CITAS data from the 1990 Population Census provide county level data on population broken up by urban and non-urban, education, and employment shares at the county level. The 1990 Census was the fourth census conducted by the National Bureau of Statistics, and the information therein was recorded on the basis of household questionnaires that were collected locally. Population figures refer to registered county level populations, and

[8]This difference is sometimes cited as one of the reasons for discrepancies between the national and the sum of province level economic accounts.

agricultural employment shares, as well as above compulsory schooling shares of the population are computed using the county totals and subtotals reported in the CITAS data. The control variable for urban population in 1990 is registered residents in urban wards taken from the CITAS population records.[9]

Appendix 5: Construction of Spanning Trees

This section describes the construction of the least cost path and Euclidean spanning tree networks depicted in Figures 2 and 3 in the paper. The stated objectives of the NTHS in 1992 were to connect all provincial capitals and cities with an urban registered population above 500,000 and connect targeted nodes to border segments as part of the Asian Highway Network in border provinces. In the 1990 Chinese Population Census, 54 cities correspond to these criteria.[10]

The following computation steps have been executed in ESRI's ArcGIS software. To construct the least cost path spanning tree network depicted in Figure 2 of the paper, I adapt a simple construction cost function from the transport engineering literature (Jha et al., 2001; Jong and Schonfeld, 2003).[11]

$$c_i = 1 + slope_i + 25 * Developed_i + 25 * Water_i + 25 * Wetland_i$$

c_i is the cost of crossing a pixel of land i, $Developed_i$ indicates whether the pixel is covered by built structures, $Slope_i$ is i's average slope gradient, $Water_i$ and $Wetland_i$ are dummies indicating whether i is covered by water or wetland. This simple specification of the construction cost function implies that shorter and flatter routes will be preferred, while high costs are assigned to crossing water bodies, wetlands, or built structures. I use the remote sensing data on land cover and elevation described in the previous section to compute c_i for a continuous grid of land parcels covering the PR China. For computational feasibility, I reclassify the original resolutions of the elevation and land cover grids from 30 arc seconds (approximately 0.82x0.82 km^2) to 2x2 km^2 grid cells.[12] This yields an isotropic

[9]This definition of urban population used in the control variable does not directly correspond with the official Chinese administrative definition. Traditionally two characteristics are used to classify urban residents for Chinese official use. The first is that at least one person of the household holds a "non-agricultural occupation" (industry or services). The second is that the sub-county level administrative unit ("zhen"=ward) is classified urban as opposed to rural.

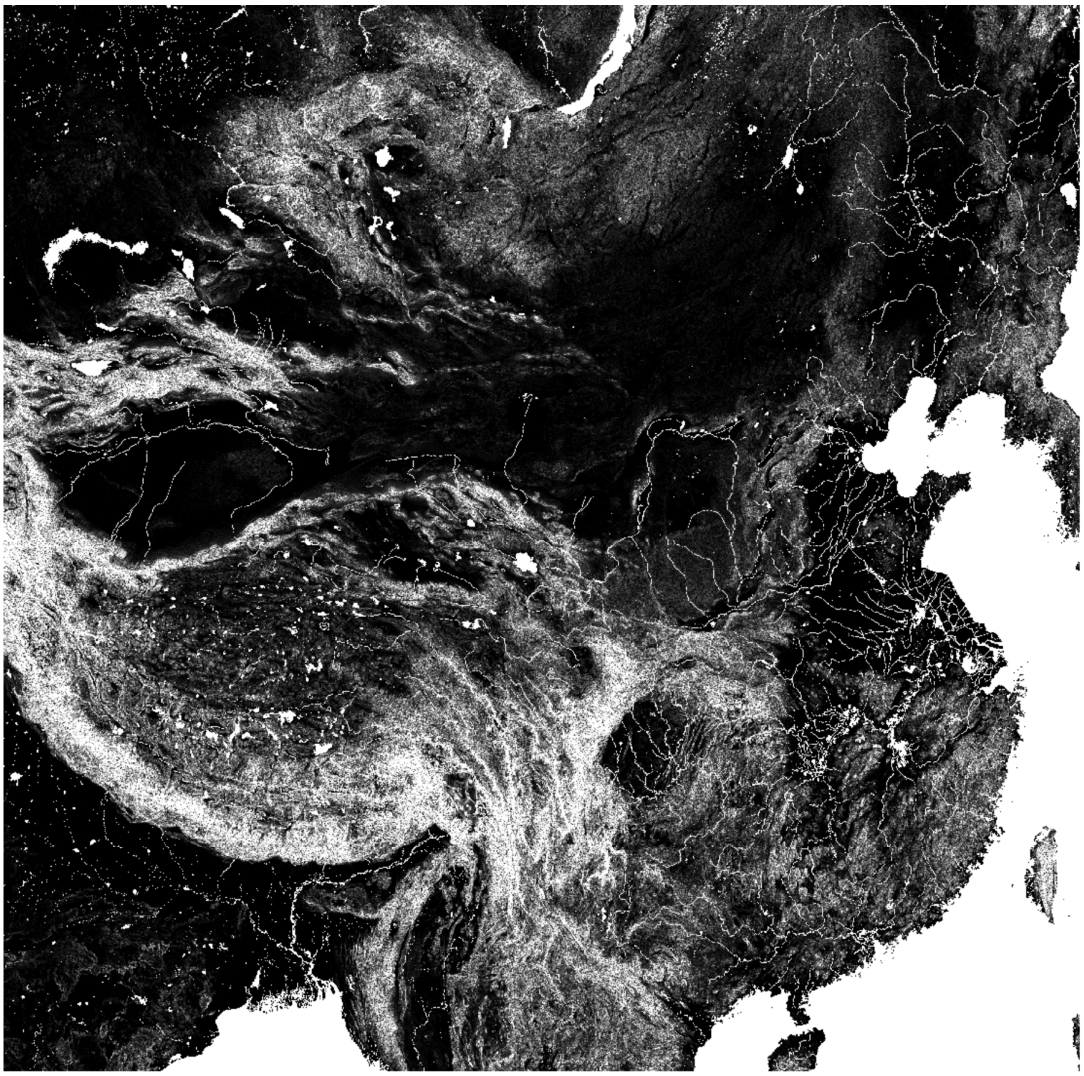
[10]The records of the 1990 Population Census became available for administrative use in 1991, and constituted the highest quality and most recent information about population registries at the ward level ("zhen") across China at the time of decision making for the NTHS in 1992. According to the Chinese administrative definition, the urban registered population of a central city is the sum of households with urban occupations in the wards ("zhen") of the central city county level units (shixiaqu) of the municipality. 1990 Census population and occupation data at the sub-county ward level was provided by the ACASIAN Data Center from archival records held at Griffith University library.

[11]The choice of the cost factor to be 25 is informed by empirical estimates of the per lane mile construction cost of highways relative to bridges. See for example WSDT (2002).

[12]To assure the continuation of rivers, the reclassified grid cells were classified as covered by water if any

cost surface grid covering the PR China in a rectangle of approximately 4.7 million 2x2 km grid cells. Figure 5.1 provides a graphical illustration of this construction cost surface.

Figure 5.1: Construction Cost Raster



The figure depicts the construction cost raster used as input into the least cost path algorithm. The color scale ranges from white (very high cost of crossing a parcel of land) to black (very low cost of crossing a square km parcel of land). The cost assignment is based on land gradient (slope) as well as land cover (water, wetlands, and developed land), and described in more detail in the text.

I then proceed to construct least cost highway paths between all 1431 $\left(\frac{54*53}{2}\right)$ possible bilateral pairs of targeted city nodes. To achieve this, I follow the accumulative cost minimization procedure pioneered by Douglas (1994). The first step is to compute Dijkstra's (1959) optimal route algorithm to identify the least costly path between each one of the 54 nodes and every cell center of the grid covering the PR China's surface. To calculate the cost of moving from the center of an origin cell to the center of one of the eight directly adjacent cells, there are two types of cost functions subject to which Dijkstra's algorithm is computed:

of the area was classified as water in the higher resolution grid.

$$c_{od1} = \frac{c_o + c_{d1}}{2}\pi, \text{ and}$$

$$c_{od2} = \sqrt{2}\frac{c_o + c_{d2}}{2}\pi$$

where c_{od1} is the cost of moving from the origin cell to one of four horizontally or vertically adjacent cells, c_{od2} is the cost of moving to one of four diagonally adjacent cells, c_o, c_{d1} and c_{d2} are the assigned construction costs of the respective cells, and π is the km cell resolution. These computations result in 54 separate cumulative cost rasters, each containing about 4.7 million 2x2 km pixels covering the PR China. Each cell is assigned the accumulative cost associated with Dijkstra's optimal route solution between any origin cell on the raster to one of the 54 nodal destinations. In addition, the computations yield 54 separate directional backlink raster files, assigning a code between 1-8 to each cell on the grid that indicates the moving direction from any cell to one of its eight adjacent cells on the identified least cost path to the particular targeted node of the grid.

The accumulative cost rasters and directional backlink rasters for each of the 54 nodes ultimately enable me to construct 1431 hypothetical least cost highway construction paths between all possible nodal connections. In the second stage, I then extract the aggregate construction cost of each possible bilateral connection in order to compute Kruskal's minimum spanning tree algorithm. This algorithm identifies 53 least cost connections that connect each of the 54 targeted cities on a single network. This yields an all-China spanning tree network.

The final step to constructing the network depicted in Figure A.3 in the paper is to apply the least cost path algorithm to find least costly connections between capitals of border provinces and segments of China's border. Least costly paths to any segment of the border within the same compass quadrant (NE, SE, SW, NW) as NTHS routes were constructed without imposing ex ante restrictions on the end points located on the border.

To construct the straight line spanning tree network depicted in Figure 3 of the paper, the first step is to compute great circle distances between all possible 1431 bilateral connections of the network, which is done by applying the Haversine formula to bilateral coordinate pairs. I then compute Kruskal's algorithm to identify the minimum number of edges that connect all targeted cities subject to the minimum aggregate distance impedance on the network. To account for the fact that Chinese planners construct many more than the minimum spanning tree connections, I re-run Kruskal's algorithm after dividing China into North-Center-South, as well subject to East-Center-West geographical areas.[13] These two additional estimations add 9 bilateral routes in addition to the 53 connections that resulted

[13]I define these geographical areas on the basis of six geographic regions with official administrative recognition in China: East, North, North-East, North-West, South-Central, and South-West.

from the all-China estimation. The final step is to include minimum great circle distance connections from provincial capitals of border provinces to the nearest border segment within the same compass quadrants as NTHS routes.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, *113*(2), 231–263.

Angrist, J., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton Univ Pr.

Baldwin, R., Forslid, R., Martin, P., Ottaviano, G., & Robert-Nicoud, F. (2003). *Economic geography and public policy*. Princeton University Press.

Baum-Snow, N., Brandt, L., Henderson, J., Turner, M., & Zhang, Q. (2012). Roads, railroads and decentralization of Chinese cities. *mimeo, Brown University*.

Combes, P., Mayer, T., & Thisse, J. (2008). *Economic geography: The integration of regions and nations*. Princeton University Press.

Douglas, D. (1994). Least-cost path in gis using an accumulated cost surface and slopelines. *Cartographica The International Journal for Geographic Information and Geovisualization*, *31*(3), 37–51.

Harris, C. D. (1954). The, market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers*, *44*(4), 315–348.

Jha, M., McCall, C., & Schonfeld, P. (2001). Using GIS, genetic algorithms, and visualization in highway development. *Computer-Aided Civil and Infrastructure Engineering*, *16*(6), 399–414.

Jong, J., & Schonfeld, P. (2003). An evolutionary model for simultaneously optimizing three-dimensional highway alignments. *Transportation Research Part B: Methodological*, *37*(2), 107–128.

Lin, S. (2009). The rise and fall of China's government revenue.

Martin, P., & Rogers, C. A. (1995). Industrial location and public infrastructure. *Journal of International Economics*, *39*(3-4), 335-351.

WSDT. (2002). *Highway construction cost comparison survey final report*. Washington State Department of Transport.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1243	Scott R Baker Nicholas Bloom	Does Uncertainty Reduce Growth? Using Disasters as Natural Experiments
1242	Jo Blanden Paul Gregg Lindsey Macmillan	Intergenerational Persistence in Income and Social Class: The Impact of Within-Group Inequality
1241	Richard Murphy Felix Weinhardt	The Importance of Rank Position
1240	Alex Eble Peter Boone Diana Elbourne	Risk and Evidence of Bias in Randomized Controlled Trials in Economics
1239	Richard Layard Dan Chisholm Vikram Patel Shekhar Saxena	Mental Illness and Unhappiness
1238	Laura Jaitman Stephen Machin	Crime and Immigration: New Evidence from England and Wales
1237	Ross Levine Yona Rubinstein	Smart and Illicit: Who Becomes an Entrepreneur and Does it Pay?
1236	Jan-Emmanuel De Neve Ed Diener Louis Tay Cody Xuereb	The Objective Benefits of Subjective Well-Being
1235	Pascal Michaillat Emmanuel Saez	A Model of Aggregate Demand and Unemployment
1234	Jerónimo Carballo, Gianmarco I.P. Ottaviano Christian Volpe Martincus	The Buyer Margins of Firms' Exports
1233	Daniel Fujiwara	A General Method for Valuing Non-Market Goods Using Wellbeing Data: Three-Stage Wellbeing Valuation
1232	Holger Breinlich Gianmarco I. P. Ottaviano Jonathan R. W. Temple	Regional Growth and Regional Decline
1231	Michalis Drouvelis Nattavudh Powdthavee	Are Happier People Less Judgmental of Other People's Selfish Behaviors? Laboratory Evidence from Trust and Gift Exchange Games

1230	Dan Anderberg Helmut Rainer Jonathan Wadsworth Tanya Wilson	Unemployment and Domestic Violence: Theory and Evidence
1229	Hannes Schwandt	Unmet Aspirations as an Explanation for the Age U-Shape in Human Wellbeing
1228	Bénédicte Apouey Andrew E. Clark	Winning Big But Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health
1227	Alex Gyani Roz Shafran Richard Layard David M Clark	Enhancing Recovery Rates: Lessons from Year One of the English Improving Access to Psychological Therapies Programme
1226	Stephen Gibbons Sandra McNally	The Effects of Resources Across School Phases: A Summary of Recent Evidence
1225	Cornelius A. Rietveld David Cesarini Daniel J. Benjamin Philipp D. Koellinger Jan-Emmanuel De Neve Henning Tiemeier Magnus Johannesson Patrik K.E. Magnusson Nancy L. Pedersen Robert F. Krueger Meike Bartels	Molecular Genetics and Subjective Well-Being
1224	Peter Arcidiacono Esteban Aucejo Patrick Coate V. Joseph Hotz	Affirmative Action and University Fit: Evidence from Proposition 209
1223	Peter Arcidiacono Esteban Aucejo V. Joseph Hotz	University Differences in the Graduation of Minorities in STEM Fields: Evidence from California
1222	Paul Dolan Robert Metcalfe	Neighbors, Knowledge, and Nuggets: Two Natural Field Experiments on the Role of Incentives on Energy Conservation
1221	Andy Feng Georg Graetz	A Question of Degree: The Effects of Degree Class on Labor Market Outcomes
1220	Esteban Aucejo	Explaining Cross-Racial Differences in the Educational Gender Gap

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7404 0612
Email info@cep.lse.ac.uk Web site http://cep.lse.ac.uk