Essays in International Trade

By

Junwei Tang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2022

Doctoral Committee:

Assistant Professor Dominick Bartelme, Co-Chair
Professor Andrei Levchenko, Co-Chair
Professor Jagadeesh Sivadasan
Associate Professor Sebastian Sotelo

Junwei Tang
tangjw@umich.edu
ORCID iD: 0000-0003-4543-1966

Dedication

To my beloved wife, Qianyu Chen

To my supportive mom, Yuxiang Tang

To my respectful dad, Shihong Tang

Acknowledgments

I am deeply grateful to Dominick Bartelme and Andrei Levchenko for their encouragement and guidance throughout my graduate study, and to my committee members, Sebastian Sotelo and Jagadeesh Sivadasan, for their invaluable feedback. I also thank Alan Deardorff, Gilles Duranton, John Leahy, Xuan Teng, Jiafu Wang, Xuan Wang and participants of seminars at University of Michigan for helpful comments.

Table of Contents

Dedication ii

Acknowledgments iii

List of Tables vi

List of Figures vii

Abstract viii

Chapter

1 Knowledge Diffusion Across Cities: Evidence from High-speed Railways in China 1

1.1 Introduction . . . 1
1.2 Motivational Evidence . . . 5
1.2.1 High-speed Rail . . . 5
1.2.2 Data . . . 7
1.2.3 Reduced-form Analysis . . . 9
1.2.4 Estimation Results and Robustness . . . 11
1.2.5 Channels at Work . . . 13
1.3 Model . . . 15
1.3.1 Workers . . . 15
1.3.2 Migration . . . 16
1.3.3 Production and Trade . . . 17
1.3.4 Unbalanced Trade and Market Clearing . . . 18
1.3.5 Model Solution . . . 19
1.3.6 Equilibrium . . . 20
1.4 Parameterization . . . 21
1.4.1 Migration Elasticity ν . . . 21

1.4.2 Joint Estimation of Productivity Spillover Parameters ρ and δ 24
1.5 Inferring Migration and Trade Costs 26
1.5.1 Migration Costs 27
1.5.2 Trade Cost 28
1.5.3 Estimation Results 29
1.6 Solution Algorithm 30
1.7 Counterfactual Analysis 34
1.7.1 Distributional Impacts of HSR 34
1.7.2 Sensitivity Analysis on the Strength of Productivity Spillovers 36
1.8 Conclusion 38
1.9 Appendix 44
1.9.1 Theoretical Appendix 44
1.9.2 Empirical Appendix 49
1.9.3 Robustness Checks for Reduced-form Analysis 54
1.9.4 Quantification Appendix 57

2 Does Comparative Advantage Predict Future Growth? 61
2.1 Introduction 61
2.2 *EXPY* 64
2.3 Replication and Extension of HHR 65
2.3.1 In-sample Performance 65
2.3.2 Out-of-sample Performance 65
2.4 Export-revealed CA Structure 68
2.4.1 Adjusted *EXPY* 68
2.4.2 Sector-level CA Structure 69
2.4.3 Prediction Results 70
2.5 Robustness Checks 73
2.6 Conclusion 77
2.7 Appendix 81

List of Tables

1.1 Summary Statistics for Transportation Data 8
1.2 Impact of HSR Construction on Prefectural Economic Development 12
1.3 Internal Migration Shares of China 23
1.4 Employment Share across Sectors 24
1.5 Calibrated Model Parameters 26
1.6 Estimates of Migration Costs and Trade Costs 31
1.7 Domestic Migration in China 49
1.8 List of Provinces by Region 50
1.9 List of Industries by Sector 51
1.10 List of Port Prefectures by Province 51
1.11 Robustness Check I for Impact of HSR Construction on Prefectural Economic Development 54
1.12 Robustness Check II for Impact of HSR Construction on Prefectural Economic Development 54
1.13 Regressions of Income Growth on Initial Economic Conditions 55
1.14 Robustness Check III for Impact of HSR Construction on Prefectural Economic Development 56

2.1 Panel growth regressions, OLS 66
2.2 Out-of-sample Performance 68
2.3 RMSE Report, Controlling for Initial GDP 72
2.4 HLN-DM Test Stats, Controlling for Initial GDP 73
2.5 RMSE Reports for Robustness Checks 75
2.6 HLN-DM Test Stats, Controlling for Initial GDP and Human Capital 76
2.7 HLN-DM Test Stats, Controlling for Initial GDP and Region Dummies 76
2.8 RMSE Report, Controlling for Initial GDP 86
2.9 RMSE Report, Controlling for Initial GDP and Human Capital 87
2.10 RMSE Report, Controlling for Initial GDP and Region Dummies 88

List of Figures

1.1 China's HSR Network as of 2015 6
1.2 Employment Density and Least Cost Path Spanning Tree Network 10
1.3 Distributional Impacts of HSR 35
1.4 Impacts of HSR on Labor Relocation 36
1.5 Slower Spatial Decay Rate of Productivity Spillovers 37
1.6 Construction Cost Raster 53

2.1 Decision Tree Illustration 84

Abstract

This dissertation studies economic geography and international trade. It discusses topics on productivity spillovers, agglomeration economies, transportation infrastructure, labor market dynamics, economic growth, comparative advantage and machine learning.

Chapter 1 measures productivity spillovers across cities by using the development of high-speed railways (HSR) in China as a natural experiment. HSR shortens inter-city passenger travel time, makes face-to-face communication easier and thus facilitates knowledge spillovers. I develop a dynamic spatial general equilibrium model that features intra- and international trade, frictional domestic migration and dynamics in labor markets. My structural estimation on the productivity spillover parameters show that production externalities are substantial but become negligible between cities that require more than 2 hours of travel time. I then calibrate the model to 2010 Chinese economy and characterize the out-of-steady-state dynamics of cities' employment and income. Quantitative results indicate that the HSR network completed in mainland China before 2015 will affect the location choice of 1.33% of the total workforce in the long run. It benefits southern and southeastern regions where both cities and HSR routes are densely located substantially more than the northern or western regions in terms of labor inflow, regional productivities and real income.

Chapter 2 (joint work with Dominick Bartelme) uses machine learning techniques to examine whether comparative advantage (CA) structure predicts GDP growth. We first show that Hausmann et al. (2007)'s *EXPY*, an aggregate index widely used by policy makers as GDP growth predictor, fails to have predictive power out of sample. We then examine if the failure of *EXPY* was due to a loss of information during the aggregation process by directly investigating the linkage between export-revealed sector-level CA structure and GDP growth while controlling for foreign demand shocks. To handle the high dimensionality problem, we adopt machine learning techniques exemplified by Random Forest. We find the sector-level CA structure outperforms *EXPY* when predicting GDP growth; nevertheless, its predictive power becomes limited after controlling for a few additional standard macro variables.

Chapter 1

Knowledge Diffusion Across Cities: Evidence from High-speed Railways in China

1.1 Introduction

The new economic geography following Krugman (1991) and Fujita et al. (1999) has been extremely popular in studying the interactions between economic agents across geographic space. It successfully captures the uneven distribution of economic activities between cities, and, among many other important factors, introduces agglomeration as a fundamental theoretical explanation. Nevertheless, most studies of this literature that interpret agglomeration as productivity spillovers model it as a function of local employment only. The productivity spillovers are thus by construction constrained within regions. Urban economists relax this restriction by enriching the local employment component with a travel-time weighted sum of employment density in surrounding city blocks. With this strategy, urban economists successfully establish the significance of productivity spillovers across blocks within a city, but they generally omit productivity spillovers across cities with the argument being travel time is too long for them to take effect (see Redding and Rossi-Hansberg (2017) for a review). In China, however, the development of high-speed rail (HSR) renders travel time across cities comparable to travel time between blocks within a city. This paper shows that productivity spillovers across cities play an important role in shaping the distribution of economic activities of China by using its HSR expansion between 2008 and 2015 as a large-scale natural experiment.

HSR generally refers to rail network that transports passengers across major cities with a top speed greater than 250 kilometers per hour. China initiated its HSR project in 2003 but did not

accelerate until 2008 when the Chinese government decided to make the HSR a cornerstone of its economic stimulus programs to confront the 2008 financial crisis. By the end of 2015, HSR had covered more than 50% of Chinese prefectures, including almost all the provincial capital cities and cities with a population greater than half million. Total HSR length reached over 19,000 km and annual passenger traffic exceeded 960 million persons.

I first show that HSR stimulates regional economic development with a reduced-form analysis over the impact of a reduction in travel time to big (employment-dense) cities led by HSR expansion on prefectural income growth. The analysis relies on the plausibly exogenous variation in each individual city's HSR access as the HSR project is planned and administered by the central government. To address the endogeneity issue of non-random HSR routes placement, I follow Faber (2014) to construct a hypothetical HSR network as an instrument for the actual HSR network. The hypothetical HSR network aims to answer the question of which HSR routes the Chinese central government would have built if its only objective had been to connect all targeted nodal cities while minimizing the total construction cost.

The exclusion restriction of the IV strategy could be violated if there are pre-existing prefectural economic conditions correlating with both locations along least cost paths and prefectural economic development. To examine this issue, I collect a large set of observable prefectural socioeconomic variables, apply LASSO to select the ones that are most related to prefectural income growth during the study period and report the estimation results after adding LASSO's selection into the estimation equation as controls. I also perform a placebo falsification test where the dependent variable is replaced with its counterpart before the HSR construction acceleration to settle the concern that pre-existing heterogeneity in prefectural growth trends might confound with the observable treatment effect.

The estimation results suggest a reduction in travel time to big cities led by HSR expansion significantly boosts prefectural income growth. Such an empirical evidence lends support to the existence of agglomeration economies across cities. I then develop this point further by discussing three potential micro-founded mechanisms at work: improvement in market potential, labor market pooling and facilitation in knowledge spillovers, which are the core theories urban economists think about agglomeration economies (Glaeser and Gottlieb 2009). I first follow Donaldson and Hornbeck (2016) to construct an empirical approximation of Harris (1954)'s market potential and add it to the estimation equation. The previous reduced-form results remain significant implying there must be other channels through which HSR contributes to the observed agglomeration economies. I then argue that HSR could hardly result in labor market pooling in mainland China by introducing China's unique institutional feature, *Hukou*, that dominates its inter-city labor mobility frictions.

The rest of this paper focuses on discussing the knowledge spillover mechanism via a structural

model. The intuition is that a reduction in passenger travel time between cities led by HSR makes face-to-face communication easier, thus facilitating knowledge spillovers. As supporting evidence, there have been a number of empirical studies finding that HSR stimulates patent growth of connected cities (e.g., Yu et al. 2019; Bian et al. 2019; Ji and Yang 2020; Dong et al. 2020). Nevertheless, the magnitude, spatial scope and general equilibrium effects of the knowledge spillover mechanism remain unanswered.

I proceed with constructing a dynamic quantitative economic geography model that encompasses the knowledge spillover mechanism as a driving force of the agglomeration economies. The model features intra- and international trade, frictional domestic migration and dynamics in labor market. Inclusion of dynamics in labor markets allows for a richer study over the medium and long run general equilibrium effects when productivity spillovers across cities are interacting with frictional domestic migration. When determining the distribution of economic activities, the agglomeration force depends on the productivity spillovers and the dispersion force depends on the inelastic land supply.

I calibrate the model parameters using prefecture-level macro data from China, including socioeconomic information, domestic trade and migration data, and international trade statistics. I find substantial production agglomeration force with an estimated elasticity of productivity with respect to city size being 0.056 that is within the range of 0.02 to 0.10 generally reported in the literature (see e.g., Combes et al. 2012). I also find for cities that are 60 minutes away 97% of the productivity spillovers would decay on the road and for cities that are 120 minutes away the productivity spillovers become negligible. In addition, when inferring the bilateral migration frictions from the migration data, I am able to confirm that institutional, geographic and cultural barriers all impede the domestic labor mobility in China with the institutional barrier significantly dominating the others.

With the calibrated model, I conduct counterfactual exercises to explore the distributional impacts of the HSR network completed before 2015 on labor distribution, regional productivities and real income in mainland China. I find that in the long run, HSR affects the location choice of 1.33% of the total workforce or 10.11 million workers. It also enlarges the between-city inequality in regional productivities and real income by 1.94% and 3.16% respectively based on interquartile range (IQR) calculation. In particular, HSR incentivizes workers in the northeastern and northwestern regions to migrate to and settle in the southern, central and southeastern regions. For instance, Guangxi province located along the southern coast gains the most among all provinces with its 14 prefectures on average creating 8.59% more jobs due to HSR. The gains in the southern, central and southeastern regions are expected as both prefectural cities and HSR routes there are densely located. Such agglomeration effects assure higher regional productivities which lead to higher income and hence attract migrants. In terms of the dynamic impacts, HSR takes 20 years to change

the location choice of 0.49% of total workforce (3.73 million workers) and 50 years to change the location choice of 0.74% of total workforce (5.64 million workers).

This paper contributes to the broad literature examining the uneven distribution of economic activity across space. Its theoretical modelling part draws insights from recent development in quantitative spatial economics exemplified by Allen and Arkolakis (2014) and Redding (2016) as well as in structural urban economics (see e.g., Lucas and Ross-Hansberg 2012, Desmet and Rossi-Hansberg 2013, Monte et al. 2018). A closely related paper is Ahlfeldt et al. (2015) who construct a tractable quantitative model of internal city structures that embeds agglomeration and dispersion forces, and then use it to evaluate the impacts of Berlin's division and reunification. I extend their model to study productivity linkage across cities in a developing country with large geographic scope.

The empirical part of this paper relates to the literature studying China's spatial economy, such as Au and Henderson (2006), Tombe and Zhu (2019), Fan (2019) and You and Wu (2020). A prominent feature emphasized by this literature is that migration frictions in China, largely due to the strict internal migration policy (*Hukou*), are outstanding. These frictions strongly impact the distribution of economic activity in China, e.g., rendering a large fraction of cities in China undersized as argued by Au and Henderson (2006). This paper directly models frictional migration between cities in a discrete choice framework. More importantly, it adds in the dynamics in labor markets following Artuç et al. (2010) and Caliendo et al. (2019) to account for the observation that migration, besides being costly, is a forward-looking decision that depends on future labor market opportunities.

This paper also contributes to the growing literature focusing on transportation infrastructure projects in developing countries. For instance, Donaldson (2016) examines the effects on domestic trade costs, inter-regional price gaps and economic welfare brought by the colonial India's railroad network. Sotelo (2020) looks into how changes in trade opportunities led by a policy of paving roads in Peru affect aggregate productivity and individual farmer's welfare. One strand of this literature directly speaks to the economic influence of large-scale transportation infrastructure investments, the national highway system in particular, undertaken in China during the past several decades (e.g., Faber, 2014; Baum-Snow et al., 2017; Baum-Snow et al., 2019; Banerjee, Duflo and Qian, 2020). My paper draws attention to another type of transportation infrastructure—HSR. It allows passenger transportation but not freight transportation and hence reshapes the distribution of economic activities through a communication facilitation and knowledge spillovers channel rather than the direct decrease in trade costs as in the other studies.

The introduction of HSR to the study of transportation infrastructure is not completely new. Charnoz et al. (2018) examine the decreases in communication costs due to the reduction in travel time between headquarters and affiliated plants induced by the HSR in France. Ahlfeldt and

Feddersen (2018) show the HSR in Germany causes GDP growth of counties with intermediate stops. Bernard et al. (2019) study how firms outsource tasks and search for suppliers using data from the HSR in Japan. These works either entirely or partially abstract away from the labor markets. Lin (2017) discusses the HSR in China and links the HSR connection of a city to its market access changes. He performs a difference-in-differences exercise and finds that an HSR connection significantly increases urban employment. My paper uses a GE approach to examine both direct effects and indirect effects that HSR have on regional economy. Recently there is a growing number of works studying the GE impacts of HSR on spatial distribution of economic activities (e.g. Hayakawa et al. (2021) for Japan and Tian and Yu (2021) for China). Compared to them, I specifically feature dynamic labor markets in the structural model to account for the complicated migration patterns in a developing country as China.

My project is most closely related to Xu (2018) with deviations lying in three important aspects. First, Xu's paper follows Eaton, Kortum and Kramarz (2016) and Bernard et al. (2019) by focusing on firm-level matching and outsourcibility, whereas mine aims to look at the effects of HSR from a more macro perspective—productivity network across cities. Second, my paper embeds a richer labor-market structure to study the medium and long run dynamics in economic activity distribution of China when productivity spillovers are interacting with frictional domestic migration. Third, I argue for city-level heterogeneity of HSR's impacts which depends on each city's pre-HSR connectivity and market size, whereas Xu assumes homogeneity of HSR's impacts among cities within a province.

1.2 Motivational Evidence

This section begins with an introduction of the high-speed rail with a focus on its expansion in China between 2008 and 2015. It then empirically establishes the contributions of HSR network to regional economic development through shortening inter-city passenger travel time and hence suggests the existence of agglomeration economies at the city cluster level. It continues with examining the channels at work by focusing on three core theories urban economists generally think about agglomeration economies as pointed out by Glaeser and Gottlieb (2009)—improvement in market potential, labor market pooling and facilitation in knowledge spillovers. These discussions motivate the general equilibrium structural model in the subsequent section.

1.2.1 High-speed Rail

"High-speed rail" (HSR) generally refers to rail transport that can carry trains with a top speed greater than 250 kilometers per hour (International Union of Railways (UIC)). It uses an integrated

Figure 1.1: China's HSR Network as of 2015

Note: The figure illustrates the high-speed rail network (in red) and prefecture boundaries of China at the end of 2015. Data Source: Li (2016).

system of specialized rolling stock and dedicated tracks, and mostly transports passengers across major cities. The first HSR is the Tokaido Shinkansen (widely known as the "bullet train") introduced by Japan in 1964. Since then, over twenty countries and regions have built and developed HSR networks (UIC 2018). East Asia has the longest HSR network in the world so far with China alone accounting for two-thirds of the world total. Europe has HSR cross international borders with Spain having the world's second longest HSR network as of 2013 bypassing Japan. Middle east and Central Asia, led by Turkey and Uzbekistan, are also planning and developing their own HSR systems.

China initiated its HSR construction in 2003 with a 404 km HSR line between Qinhuangdao and Shenyang (Ollivier et al. 2014). In 2004, the Chinese central government issued the "The Medium- and Long-Term Railway Network Plan (2004)" where it specified a "4-4" HSR network of four horizontal and four vertical corridors. The original goal was to complete a total length of 12,000 km of HSR routes by 2020. Starting from 2008, the Chinese government accelerated the HSR network construction as it decided to make the HSR network a cornerstone of its economic stimulus programs to confront the 2008 financial crisis ("The 2008–09 Chinese Economic Stimulus Plan")[1]. The HSR program gained momentum quickly. By the end of 2015, it had covered more

[1]The HSR line between Qinhuangdao and Shenyang was the only route China built between 2003 and 2007

than 50% of Chinese prefectures[2], including almost all the provincial capital cities and cities with a population greater than half million. Total HSR length reached over 19,000 km and total passenger traffic exceeded 960 million persons (China Statistical Yearbook 2015). The "4-4" HSR network was completed ahead of time. Figure 1.1 plots the HSR routes in operation by the end of 2015 as well as the Chinese prefecture boundaries. Given the rapid completion of the "4-4" HSR program, the Chinese central government initiated a more comprehensive "8-8" HSR network plan in 2016 where it aimed to link all middle- and large size cities via eight horizontal and eight vertical corridors by 2030.

1.2.2 Data

Geo-referenced HSR routes come from the CHGIS dataverse supported by Harvard University. They are compiled using publicly available information on Google Earth and Open Street Map by Yifan Li in 2016. I trace and adjust every HSR line launched between 2008 and 2015 to the China High-speed Railways Network Layout published yearly by the National Railway Administration of China.

The most direct impact of the HSR development is the reduction in inter-city passenger travel time. To quantify that, I merge the HSR routes data with the geographical information system (GIS) data for several other transportation modes namely national highways, provincial highways and regular railroads from Baum-Snow et al. (2017)[3]. Baum-Snow et al. (2017) digitize a series of large-scale Chinese transportation maps published by SinoMaps Press, a national-level map publisher in China. I adopt their 2010 routes and keep the routes unchanged during my study period[4]. As a result, the only difference between the 2007 and 2015 passenger transportation networks in my data is the HSR routes that started operation between 2008 and 2015.

I then compute the travel time between any two prefectural cities by tracing the quickest route in between their city halls, which can consist of only one or any combination of the four transportation modes. I assume 250 km/hr, 100 km/hr, 60 km/hr and 70 km/hr as the speed to HSR, national highways, provincial highways and railroads respectively. Travel time from a prefecture to itself is set to be 0 min. Table 1.1 summarizes the results. I exclude the 7 prefectures in Tibet due to data availability and end up with 333 prefectural cities (55,278 unique city pairs). With the

(Ollivier et al. 2014).

[2]Prefectures are sub-province administrative regions in China. There are 340 of them in 2015.

[3]I leave out waterways and air transportation due to data availability. They account for a total of roughly 1.5% of passenger traffic and 13% of freight traffic (in volume) in China (China Statistical Yearbook 2010).

[4]From 2008 to 2015, the regular railway system in China remained relatively invariant but the highway network had been under rapid development. Omitting the advances in highway may seem nocuous to my analysis. However, most of the highway construction during this period was to link counties, townships or villages (administrative divisions lower than prefectures) to the pre-existing highway network; the transport time between prefectural cities were not affected significantly given that most of the prefectures had already been linked by highways before 2008.

Table 1.1: Summary Statistics for Transportation Data

	2007	2015
HSR		
Number of routes	1	64
Number of cities linked	6	190
Total length (km)	404	>19,000
Passenger Traffic (million)	<1	>960
Travel Time		
Mean (min)	1,102	664
Std dev (min)	666	437
Max (min)	4,239	3,077
Number of city pairs	55,278	55,278
Avg number of cities within 1-hour radius	0.65	2.03
Avg number of cities within 2-hour radius	3.44	9.14

Notes: Travel time is based on author's calculation using HSR routes data from Li (2016) and national highways, provincial highways and railroads routes data from Baum-Snow et al. (2017).

development of HSR from 2008 to 2015, passenger travel time between two Chinese prefectural cities decreased by 40% on average. The average number of cities a Chinese prefectural city could travel to within 1 hour more than tripled from 0.65 to 2.03.

Geographic data used to construct the least cost path HSR routes such as elevation and land usage are from the Resource and Environment Science Data Center, Chinese Academy of Sciences. Prefecture-level socioeconomic data, including GDP, population, fixed asset investments, industrial output and government revenue come from the annual China City Statistical Yearbook from 2010 to 2017. Import and export data come from the 2010 China Statistical Yearbook for Regional Economy. Land area data come from the second National Land Survey conducted between 2006 and 2009. Labor data including employment, migration, sectoral composition and education attainment come from the 2000, 2010 China National Population Census and the 2005, 2015 Population Micro Census (1% random sample survey) complemented by the annual China City Statistical Yearbooks[5]. Consumer price indices come from the 2010-2017 provincial statistical yearbooks complemented by the yearly Statistical Communique on the Economic and Social Development of the prefectural governments.

Data for the trade flows between Chinese provinces come from the 2010 interregional input-output tables of China compiled by Liu et al. in 2014. The 2010 interregional IO table bases itself on the interregional and regional input-output (IRIO) model proposed by Isard (1951) and utilizes the Provincial Input-Output Tables provided by the National Bureau of Statistics of China and

[5]See Appendix for detailed labor data compilation.

goods transportation data provided by the Ministry of Transport of China. GDP, land area and employment data for the rest of world (ROW) are from the World Bank.

1.2.3 Reduced-form Analysis

The rapid expansion of the HSR network between 2008 and 2015 provides plausibly exogenous variation in each individual city's HSR access and hence travel time to other cities, as it is planned and administered by the central government with little influence from the local government ("The Medium- and Long-Term Railway Network Plan (2004)"). I use this information to estimate the impact of reduction in travel time to the big cities between 2007 and 2015 on prefecture economic development between 2010 and 2017.

The baseline estimation strategy is a difference in differences specification as following:

$$ln(y_{ip}^{2017}) - ln(y_{ip}^{2010}) = \beta(BigCityAccess_{ip}^{2015} - BigCityAccess_{ip}^{2007}) + \gamma_p + \eta X_{ip} + \epsilon_{ip} \quad (1.1)$$

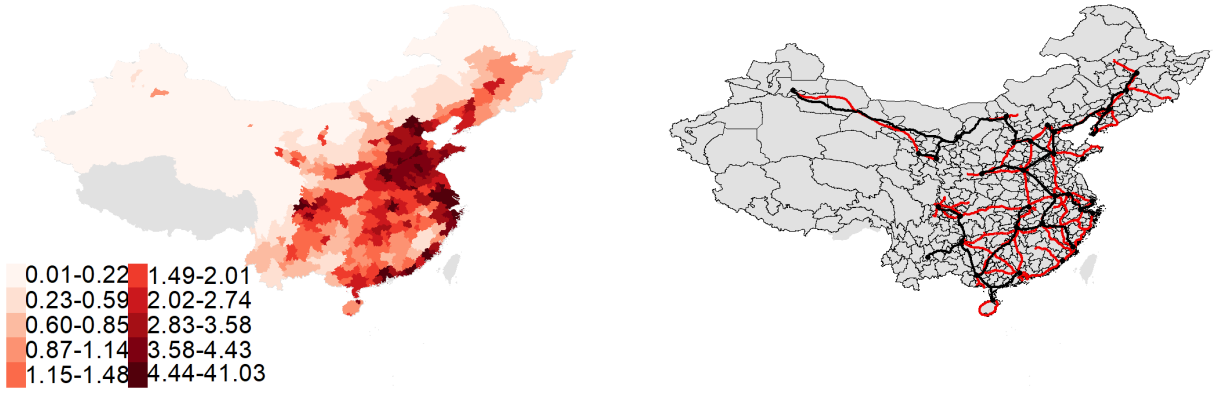
where y_{ip} is the real GDP per worker for prefecture i in province p, $BigCityAccess_{ip}$ is the number of top 50 cities (rank based on employment density in 2010 shown in Figure 1.2a) prefecture i can travel to within 1 hour, γ_p is province fixed effects and X_{ip} is a vector of prefecture control variables discussed later. I use employment density, defined by workers per square kilometer of administrative area, as a measure of city scale following the urban economics literature. I also cluster the error term ϵ_{ip} at the provincial level as it could be correlated across prefectures that were connected to a similar section of the HSR network.

Least Cost Path Spanning Tree Network

Directly estimating equation (1) by OLS requires the assumption that placement of HSR routes between nodal cities was random within provinces. However, this assumption seems too strong given the political setting of the HSR network. As documented in "The Medium- and Long-Term Railway Network Plan (2004)", the Chinese government wanted to use the HSR network to facilitate urbanization and promote the economic growth of less developed regions from the very beginning. To deal with this endogeneity issue, I follow Faber (2014) to construct a hypothetical HSR network as an instrument for the actual HSR routes (Figure 1.2b). I then assume it was the hypothetical HSR network rather than the actual HSR network that took place between 2008 and 2015 and recompute the inter-city travel time to obtain the IV for $\Delta BigCityAccess_{ip}$.

Specifically, the hypothetical HSR network is based on a least cost path spanning tree network. It aims to answer the question of which HSR routes the Chinese central government would have built if their only objective had been to connect all targeted nodal cities while minimizing the

Figure 1.2: Employment Density and Least Cost Path Spanning Tree Network



(a) Employment Density in 2010 (b) Least Cost Path Spanning Tree Network

Notes: In Panel (a), employment density is computed as 100 workers per km^2 of administrative area for 333 prefectures. The 10 legends correspond to the deciles with darker color indicating higher values. Gray shaded areas are excluded from the analysis due to limited data. The top 50 employment-dense cities are mainly located in the eastern and southern regions, although there are also a few scattered in the western region. In Panel (b), red routes are the actual HSR network in operation at the end of 2015. Black routes are the hypothetical HSR network IV constructed using least cost path and minimum spanning tree algorithm.

total construction cost. Step-by-step construction of the hypothetical HSR network follows Faber (2014). I first implement Dijkstra's optimal route algorithm to compute least cost HSR paths between all targeted nodal city pairs based on remote sensing data on land cover, land use and elevation. I classify all the provincial capital cities and sub-provincial municipalities as the nodal cities given these are the cities targeted by the Chinese planners in the HSR network plan[6]. I assign higher construction costs to land with steeper slope gradients and land covered with built structures, water or wetland (see Appendix for detail). Then I apply Kruskal's minimum spanning tree algorithm to identify the single continuous HSR network that connects all targeted nodal cities while minimizing the total construction cost.

Control Variables and Identifying Assumption

The least cost path IV aims to address the concern that the placement of HSR routes between nodal cities might not be random. However, the exclusion restriction could be easily violated if there are pre-existing prefectural economic conditions correlating with both locations along least cost paths due to historical reasons and income growth between 2010 and 2017 [7]. I thus need to

[6]There are 30 provincial capital cities and 15 sub-provincial municipalities among the 333 prefectural cities in my data. 11 of the 15 sub-provincial municipalities are also provincial capital cities.

[7]For example, a city on a plain may be passed by the hypothetical HSR network due to the low construction costs. Also, since it is on a plain, the city may have a large agriculture population which may be related to its income

control for the pre-existing economic conditions.

A follow-up question is which socioeconomic variables could capture the pre-existing economic conditions. I prefer not to take a stand on this issue beforehand; instead, I collect the 2010 value of a large set of observable prefecture-level variables and apply Double-LASSO (Belloni et al. 2014) to select the ones that are most strongly correlated to prefectural income growth between 2010 and 2017[8]. The variable candidates consist of logarithm of GDP level, GDP per worker, population, employment density, total import and export value, investment in fixed assets, industrial output, government revenue as well as share of agriculture employment in total employment and share of college-educated workers. Double-LASSO picks the logarithm of GDP per worker only, indicating prefectural initial income level is strongly associated with income growth. Therefore, I use the 2010 value of GDP per worker as the control for prefectural pre-existing economic conditions and plug it into X_{iq} in the estimation equation (1.1).

The baseline identifying assumption then is that reduction in travel time to big cities based on the hypothetical least cost spanning tree HSR network contributes to prefectural income growth only through the reduction in travel time due to actual HSR routes, conditional on province fixed effects and prefectural initial economic conditions.

1.2.4 Estimation Results and Robustness

Table 1.2 columns (1) and (3) present the baseline estimation results. As shown by the first stage F-stats, the least cost path is a strong IV for the actual HSR routes. In terms of estimated coefficients, the change in number of big cities within 1-hour radius variable has positive and statistically significant values in both OLS and IV regressions suggesting improvement in access to big cities induced by HSR development contributes to prefectural income growth. The IV estimate shows that one more big city reachable within 1 hour between 2007 and 2015 on average increased prefecture-level growth rate of real GDP per worker between 2010 and 2017 by 16%, which is in-line with the literature studying HSR in China (e.g., Xu 2018). The OLS estimates a smaller boost of 5%. The underestimation of the OLS is consistent with the hypothesis that after targeting nodal cities, the Chinese government may intentionally place the HSR network near less developed regions with the hope that it would promote their economic development.

A fundamental concern in the transportation infrastructure literature is that regions may have different growth trends before the transportation network took place and these pre-existing

growth.

[8]I implement Double-LASSO by performing LASSO on both the dependent ($\Delta ln(y_{ip})$) and independent variable ($\Delta BigCityAccess$) with the full set of observable prefecture-level variables as covariate candidates. The union of the covariate sets selected by the two individual LASSO exercises gives the final control for prefectural pre-existing economic conditions. Belloni et al. (2014) show that the post-double-selection LASSO estimator is consistent and asymptotically normal under some mild conditions.

Table 1.2: Impact of HSR Construction on Prefectural Economic Development

	2010-2017				2000-2007			
Dependent variable	OLS	OLS	LCP IV	LCP IV	OLS	OLS	LCP IV	LCP IV
$\Delta ln(rGDP/worker)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ BigCityAccess	0.051***	0.057***	0.160**	0.139***	0.022	0.023	0.076	0.071
	(0.018)	(0.018)	(0.068)	(0.052)	(0.024)	(0.026)	(0.050)	(0.046)
lngdppw10	-0.196***	-0.202***	-0.229***	-0.228***	0.163***	0.162***	0.146***	0.147***
	(0.047)	(0.046)	(0.047)	(0.047)	(0.043)	(0.042)	(0.042)	(0.043)
ΔlnMP		4.916**		5.288***		1.247		1.464
		(2.148)		(1.859)		(1.354)		(1.310)
First stage F-Stat			20.54	22.10			20.54	22.10
Obs	333	333	333	333	333	333	333	333
R^2	0.87	0.88			0.94	0.94		

Notes: All regressions include province fixed effects. LCP refers to the least cost path spanning tree network. lngdppw10 is log of GDP per worker in 2010. ΔlnMP is log change of market access between 2010 and 2017. Standard errors are clustered at the province level. ***1%, **5%, and *10% significance levels.

trend differences are likely to be confounding with the observable treatment effect. To examine this issue, I perform a falsification test where I replace the dependent variable with prefecture-level growth rate of real GDP per worker across the seven years before the acceleration in HSR construction, $ln(y_{ip}^{2007}) - ln(y_{ip}^{2000})$. Specifications of the explanatory variables remain unchanged. If the exclusion restriction was satisfied, there should be no significant relationship between changes in travel time to big cities induced by HSR and real income growth prior to the network. Table 1.2 columns (5) and (7) present the estimation results of the falsification test. The change in access to big cities variable becomes statistically insignificant in both regressions as expected and hence the exclusion restriction is further supported.

As for robustness checks, I redefine the $BigCityAccess$ variable so that it picks the top 50 cities based on GDP level in 2010. To check the sensitivity of the control variable selection, I expand X_{ip} to include log of total import and export value and log of industrial output in 2010[9]. The supplementary Appendix reports the results. Improvement in access to big cities induced by HSR development has positive and statistically significant effects on prefectural real income growth across all these specifications. Furthermore, I test the geographic scope of the big city accessibility by changing $BigCityAccess$ variable to count the number of top-50 employment density cities within 2-hour radius instead of 1-hour. In this case, the coefficient for $\Delta BigCityAccess$ becomes no longer significant at 5% significance level. The relationship between the existence of big cities and a prefecture's income growth weakens when they are farther away. If put into a knowledge

[9]Log of GDP per worker, total import and export value and industrial output in 2010 are the three variables with 5% significance level when regressing $ln(y_{ip}^{2017}) - ln(y_{ip}^{2010})$ on the full set of potential prefecture-level controls. See Appendix.

spillover channel which I discuss in details in the following section, this result is consistent with the empirical studies that find knowledge spillover decays over space[10].

1.2.5 Channels at Work

The previous section presents empirical evidence showing that a reduction in travel time to big cities led by HSR network development helps with prefectural income growth. The conclusion stands robust after controlling for province fixed effects and prefecture-level preexisting economic conditions. This finding lends support to the geographical clusters of cities; specifically, it suggests the existence of agglomeration economies at the city cluster level. In this section, I discuss three possible micro-founded mechanisms at work: improvement in market potential, labor market pooling and facilitation in knowledge spillovers. These three channels are the core theories urban economists think about agglomeration economies (Ottaviano 2008, Glaeser and Gottlieb 2009).

A transportation infrastructure could make a city more appealing to firms by increasing its relative size or improving its accessibility to other trading markets. Both dimensions are captured by the city's "market potential", a measure of customer proximity originally introduced by Harris (1954). The HSR in China does not allow freight transportation ("The Medium- and Long-Term Railway Network Plan 2016 Revision"), thus it is supposed to have no direct impact on trade costs or the relative centrality of a city in the network of trading markets. In other words, HSR affects a city's market potential mainly through changing its number of consumers. To empirically examine whether improvement in market potential works as a source of the observed agglomeration economies, I follow Donaldson and Hornbeck (2016) to construct an approximation of Harris's market potential for city i in year t as $MP_i^t \approx \sum_{n \neq i} (d_{in})^{-1} L_n^t$. d_{in} is the great-circle distance between cities i and n and L_n^t is the number of workers of city n in year t[11]. I then add the log change of market potential between 2010 and 2017 as an additional control variable to the estimation equation. I also treat the 2010 value of log of market potential as an additional candidate for Double-LASSO selection when choosing the appropriate variables to control for prefecture-level preexisting economic conditions.

Columns (2) and (4) of Table 1.2 report the OLS and IV estimation results respectively[12]. Two findings stand out. First, the coefficient for market potential growth being positive and statistically significant suggests an improvement in market potential does contribute to city's income growth

[10]For example, Conley et al. (2003) concludes that knowledge transmission between people vanishes when they are 90-120 minutes away.

[11]Ideally, one would want to use the "market access" widely known in the trade literature (e.g. Redding and Venables 2004) as a more comprehensive measurement for the first mechanism. However, an empirical construction of market access requires assumptions on the structure of the underlying model as well as an estimation for the trade elasticity. The "market potential" on the other hand serves as a simpler, less model-dependent and yet remains informative alternative (Head and Mayer 2006).

[12]Double-LASSO still picks the initial logarithm of GDP per worker only.

which is consistent with the literature (e.g., Redding and Venables 2004). Second, even after controlling for market potential growth, the change in number of big cities within 1-hour radius variable remains significantly positive. This result points out that other than the improvement in market potential, there must be other channels through which the reduction in travel time to the big cities induced by HSR expansion contributes to prefectural economic development.

Another possible mechanism for agglomeration economies at the city cluster level is labor market pooling that dates back to Marshall (1890). The original idea is that geographic concentration of workers and firms increases the likelihood of good matches and protects against risks by providing more outside options. In my case, the labor market pooling works through the possibility that a reduction in inter-city travel time makes it easier for workers residing in different cities to commute to and work in the same city and hence leads to a denser labor market[13].

However, this channel could hardly be the main source for agglomeration economies in mainland China. Studies on major HSR system in other countries, e.g., Hayakawa et al. (2021) on Tokaido Shinkansen, have documented that HSR is rarely used by workers for daily commuting due to sparsity of stations and expensive prices. The HSR in mainland China also faces the problems of limited number of stations (usually one or two) within a city and relatively high ticket price compared to other transportation modes, and thus is hardly able to substantially change the commuting pattern. The unique institutional feature, *Hukou*, of China further limits the impact of HSR on commuting. *Hukou* ties each Chinese citizen to a prefectural city via a household registration status normally based on the citizen's birthplace. Though it does not preclude citizens from working outside their *Hukou* city, the *Hukou* system severely restricts the access of workers without *Hukou* of a city to the city's public goods such as children's education offered by public schools, health care, unemployment benefits and social security (Chan 2010)[14]. Since it is these institutional barriers that dominate the inter-city labor mobility frictions in China rather than transport costs (Fan 2019), HSR is not likely to contribute to the regional agglomeration economies through the pooling of labor markets across nearby cities.

The third mechanism is that a reduction in passenger travel time between cities makes face-to-face communication easier, thus facilitating knowledge spillovers[15]. Such agglomeration economies result when ideas are transferred between agents (irrespective of their market transactions) and the transfer is imperfect across space (Glaeser and Gottlieb 2009). The contributions of HSR to inter-city knowledge spillovers are well-established by a number of reduced-form empirical

[13]The idea that workers can separate between workplace and residence by commuting is more common among urban economics literature studying within-city distribution of economic activities. See, e.g., Monte et al. (2018) and Ahlfeldt et al. (2016).

[14]Over the past decades, *Hukou* has been relaxed through a series of reforms but remains restrictive especially in the big cities (You and Wu 2019).

[15]The idea-based agglomeration economies are also referred to as human capital externalities in labor and urban economics literature. See, e.g., Acemoglu and Angrist (2000), Moretti (2004), Ciccone and Peri (2006), and Fu (2007).

studies finding that HSR stimulates patent growth of connected cities (Yu et al. 2019; Bian et al. 2019; Ji and Yang 2020; Dong et al. 2020). However, the magnitude, spatial scope and general equilibrium effects of the knowledge spillover channel remain unanswered. In the next section, I construct a spatial general equilibrium model that encompasses the knowledge spillover channel as a driving force of the agglomeration economies. I then proceed to estimate the magnitude of the knowledge spillover force as well as its spatial decay rate using model-based econometric techniques.

1.3 Model

The model is a dynamic general equilibrium economic geography model based on Ahlfeldt et al. (2015) and Caliendo et al. (2019). It features intra- and international trade, frictional migration across domestic regions and dynamics in labor market. The distribution of economic activities is driven by both agglomeration forces (productivity spillovers) and dispersion forces (inelastic land supply). To model productivity spillovers, I embed city-level endogenous productivity which depends on the travel time-weighted sum of employment density in surrounding cities. I include dynamics in labor markets since first, data suggest the net inter-prefecture migration in China is huge and hence its labor market should not be taken as being at a steady state[16]. Second, dynamic labor markets are essential in studying the medium and long run general equilibrium effects when productivity spillovers across cities are interacting with frictional domestic migration.

The model has $N + 1$ regions representing China's N prefectures plus the Rest of World, indexed by $n, i \in \{1, \ldots, N + 1\}$. Each region n is endowed with a priced fixed factor (land), denoted by $\overline{H_n}$, that is used by workers for housing and by firms for production. Each region is also populated with an endogenous measure of L_n workers who can costly move across regions within China but not across countries. I assume workers consume in the same region where they work and they cannot borrow or save.

1.3.1 Workers

A worker o living in region n at time t derives her current period's utility from consumption over local final goods, $C^o_{n,t}$, and residential land use, $H^o_{n,t}$, adjusted by residential amenities, B_n, which capture exogenous common characteristics making a region more or less attractive place to live:

$$U(B_n, C^o_{n,t}, H^o_{n,t}) = ln(B_n(C^o_{n,t})^\alpha (H^o_{n,t})^{1-\alpha}) \tag{1.2}$$

[16]See Appendix for detailed discussion.

The worker faces the following budget constraint given $P_{n,t}$ as the consumption goods price, $r_{n,t}$ as the land rental price and $v^o_{n,t}$ as the nominal income:

$$P_{n,t}C^o_{n,t} + r_{n,t}H^o_{n,t} \leq v^o_{n,t} \tag{1.3}$$

The indirect utility for the worker becomes $u^o_{n,t} = ln(\frac{B_n v^o_{n,t}}{P^\alpha_{n,t} r^{1-\alpha}_{n,t}})$ and the average current period's indirect utility for a random worker in n can thus be represented as

$$u_{n,t} = ln(\frac{B_n v_{n,t}}{P^\alpha_{n,t} r^{1-\alpha}_{n,t}}) \tag{1.4}$$

I assume income of a worker consists of a wage income $w^o_{n,t}$ as well as a land income that is distributed evenly among the workers working there[17]: $v^o_{n,t} = w^o_{n,t} + \frac{r_{n,t}\overline{H_n}}{L_{n,t}}$.

1.3.2 Migration

The migration part follows a dynamic spatial discrete choice model based on Artuç, Chaudhuri, and McLaren (2010) and Caliendo et al. (2019). At the beginning of each period, workers observe the economic conditions in all labor markets as well as the realizations of their own idiosyncratic shocks. If they begin the period in a certain labor market, they inelastically supply one unit of labor and earn the corresponding nominal income. At the end of the period, workers have the option to reallocate, or formally, a worker o currently working in region n maximizes the following lifetime utility

$$ln(V^o_{n,t}) = U(B_n, C^o_{n,t}, H^o_{n,t}) + Max_i\{\beta E[ln(V^o_{i,t+1})] - ln(\kappa_{ni}) + \frac{1}{\nu}\epsilon^o_{i,t}\} \tag{1.5}$$

where κ_{ni} is the origin-destination specific utility loss associated with relocating. β is the discount factor and $\frac{1}{\nu}$ is a normalization factor. This utility formulation captures the idea that workers not only value the current-period utility but also the option value to move into any other market in the future.

Denote $V_{n,t} = E(V^o_{n,t})$ as the expected lifetime utility for a random worker in city n at time t. I assume workers are risk-neutral and have perfect foresight. Then we can have the rule of motion for expected lifetime utility as well as the fraction of workers that relocate from n to i at the end of period t as[18]

$$V^\nu_{n,t} = [exp(u_{n,t})]^\nu [\sum_i V^\nu_{i,t+1}\kappa^{-\nu}_{ni}]^\beta \tag{1.6}$$

[17]The household registration system in China, *Hukou*, limits the access of workers without *Hukou* of a city to the city's local land income as discussed in Tombe and Zhu (2019). I omit such restrictions for simplicity.

[18]See Appendix A.1 and A.2 for detailed derivation.

$$\mu_{ni,t} = \frac{V_{i,t+1}^{\beta\nu}\kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,t+1}^{\beta\nu}\kappa_{nm}^{-\nu}} \tag{1.7}$$

under the assumption that the idiosyncratic shock ϵ is i.i.d. over time and distributed Type-I extreme value with zero mean. Note that with the specification for $u_{n,t}$, Equation (1.6) can be further expanded to $V_{n,t}^{\nu} = [\frac{B_n v_{n,t}}{P_{n,t}^{\alpha} r_{n,t}^{1-\alpha}}]^{\nu}[\sum_i V_{i,t+1}^{\nu}\kappa_{ni}^{-\nu}]^{\beta}$. After accounting for migration costs, the higher lifetime utility a region can provide the more workers it will attract. The equilibrium condition characterizing how the distribution of labor across markets evolves over time becomes

$$L_{i,t+1} = \sum_{n=1}^{N} \mu_{ni,t} L_{n,t} \tag{1.8}$$

1.3.3 Production and Trade

The production part follows an Eaton and Kortum (2002) framework with endogenous productivity and input-output linkages. There is only one traded sector. Each region is able to produce every variety $\omega \in [0, 1]$ and has a perfectly competitive firm producing a region specific composite good $Y_{n,t}$ using the CES technology across a continuum of horizontally differentiated varieties $y_{n,t}(\omega)$:

$$Y_{n,t} = (\int_0^1 y_{n,t}(\omega)^{\frac{\sigma-1}{\sigma}} d\omega)^{\frac{\sigma}{\sigma-1}} \tag{1.9}$$

These varieties are produced by perfectly competitive firms using labor $L_{n,t}$, land $H_{n,t}$ and intermediate inputs $Z_{n,t}$. With the assumption of a Cobb-Douglas production function, the unit cost of production for a firm in region n producing variety ω with productivity $A_{n,t}\varphi$ is

$$c_{n,t}(\varphi) = \frac{1}{A_{n,t}\varphi} w_{n,t}^{\eta_1} r_{n,t}^{\eta_2} P_{n,t}^{\eta_3} \tag{1.10}$$

where $w_{n,t}$ is the wage, $r_{n,t}$ is the rental cost of land and $P_{n,t}$ is the price of purchased intermediate inputs which is the same as the price of the final good $Y_{n,t}$. η_1, η_2 and η_3 are the shares of costs spent on wage expense, land rental and input purchase respectively. Assume that the firm specific productivity is distributed Fréchet with CDF $F(\varphi) = e^{-\varphi^{-\theta}}$ where θ is the dispersion parameter.

$A_{n,t}$ is the regional productivity that is enjoyed by all firms producing in n at t. As in Ahlfeldt et al. (2015), I assume it to be composing of an exogenous production fundamental component, $\overline{A_n}$, and an endogenous production externality component, $\Upsilon_{n,t}$:

$$A_{n,t} = \overline{A_n}(\Upsilon_{n,t})^{\rho} \tag{1.11}$$

where ρ determines the magnitude of the production agglomeration force. Production externalities, $\Upsilon_{n,t}$, is further modeled as a travel-time, ι, weighted sum of employment density in other regions:

$$\Upsilon_{n,t} \equiv \sum_{s=1}^{N} e^{-\delta \iota_{sn}} \left(\frac{L_{s,t}}{\overline{H_s}}\right) \tag{1.12}$$

$\overline{A_n}$ captures all fundamentals (e.g. climatic conditions) rendering a region more or less productive that are not related to the employment density of its surrounding regions. $\Upsilon_{n,t}$ structures the productivity spillovers across regions. Specifically, the productivity spillovers from region s to region n decline with travel time, ι_{sn}, through an iceberg factor $e^{-\delta \iota_{sn}} \in (0, 1]$. δ determines the spatial decay rate of productivity spillovers. Traveling within each region is instantaneous: $\iota_{ss} = 0$ for all region s. To interpret the two productivity spillover parameters ρ and δ, a positive value of ρ indicates increasing returns to scale and agglomeration economies. The larger ρ is, the stronger the production agglomeration force. As for δ, if δ is 0, then productivity spillovers are perfect across geographic space. The larger δ is, the quicker productivity spillovers decay with travel time and hence the smaller productivity spillovers' spatial scope.

Trade faces the standard iceberg costs: in order for 1 unit of product to arrive in region n from region i, $\tau_{in} > 1$ units of products need to be produced and shipped. $\tau_{ii} = 1$ for all i. Assuming away arbitrage we have consumer price of variety ω for goods produced in region i and sold in region n as $p_{in,t}(\omega) = \tau_{in} c_{i,t}(\omega)$. Consumers source from the cheapest producer: $p_{n,t}(\omega) = min_i\{p_{in,t}(\omega)\}$.

To model international trade, I divide the regions in China into two mutually exclusive groups: port regions and inland regions. A port region a can trade directly with the ROW (indexed by c) and face the iceberg trade costs discussed above, τ_{ac}. An inland region b, on the other hand, needs to first ship its goods to the closest port region and then trade with the ROW: $\tau_{bc} = \tau_{ba}\tau_{ac}$.

1.3.4 Unbalanced Trade and Market Clearing

To accommodate the trade imbalance as is evident in the interregional IO data, I assume exogenous trade surpluses and deficits that are proportional to the region's total labor income. Specifically, let $S_{n,t}$ denote region n's trade surplus at time t, then $S_{n,t} = \chi_n w_{n,t} L_{n,t}$ and $\sum_{n=1}^{N+1} S_{n,t} = 0$. A trade surplus works as a capital outflow that shrinks the nominal income of the region: $v_{n,t} L_{n,t} = w_{n,t} L_{n,t} + r_{n,t} \overline{H_n} - S_{n,t}$.

Markets for final goods, labor and housing clear during each time period t. The final goods

market clearing condition implies

$$X_{n,t} = \alpha v_{n,t} L_{n,t} + \eta_3 \sum_{i=1}^{N+1} \pi_{in,t} X_{i,t} \tag{1.13}$$

where $X_{n,t}$ denotes the total expenditures by region n and $\pi_{in,t}$ denotes the fraction of region i's spending allocated to goods produced in region n. The first item on the right-hand side is total demand of the goods by local workers and the second item is total demand by both local and non-local producers as intermediate inputs.

Labor market clearing condition for region n implies

$$w_{n,t} L_{n,t} = \eta_1 \sum_{i=1}^{N+1} \pi_{in,t} X_{i,t} \tag{1.14}$$

and the housing market clearing implies

$$r_{n,t} \overline{H_n} = (1-\alpha) v_{n,t} L_{n,t} + \eta_2 \sum_{i=1}^{N+1} \pi_{in,t} X_{i,t} \tag{1.15}$$

where the first item on the right-hand side is total demand of land by workers for residential purposes and the second item is total demand by producers as industrial land.

1.3.5 Model Solution

Substituting $v_{n,t} L_{n,t} = w_{n,t} L_{n,t} + r_{n,t} \overline{H_n} - S_{n,t}$ into Equation (1.15) and combining the result with Cobb-Douglas production technologies yield the following representation for the total land income in region n:

$$\alpha r_{n,t} \overline{H_n} = (1 - \alpha + \frac{\eta_2}{\eta_1}) w_{n,t} L_{n,t} - (1-\alpha) \chi_n w_{n,t} L_{n,t} \tag{1.16}$$

This equation suggests that given the model setup we can compute equilibrium land rental price directly from equilibrium wage rate and labor distribution. Moreover, we can arrive at a concise proportional relationship between wage income w_n and total nominal income v_n:

$$v_{n,t} = \frac{1 + \frac{\eta_2}{\eta_1} - \chi_n}{\alpha} w_{n,t} \tag{1.17}$$

Competition implies that the price paid for a particular variety ω in region n is given by the minimum unit cost across regions after accounting for trade costs:

$$p_{n,t}(\omega) = min_i\{\tau_{in}c_{i,t}(\varphi_i(\omega))\} = min_i\{\frac{\tau_{in}}{A_{i,t}\varphi_i(\omega)}w_{i,t}^{\eta_1}r_{i,t}^{\eta_2}P_{i,t}^{\eta_3}\} \tag{1.18}$$

Given the Fréchet distribution, the aggregate price index in region n is:

$$P_{n,t} \propto [\sum_{i=1}^{N+1}(\tau_{in}w_{i,t}^{\eta_1}r_{i,t}^{\eta_2}P_{i,t}^{\eta_3})^{-\theta}A_{i,t}^{\theta}]^{-1/\theta} \tag{1.19}$$

Then the fraction of region n's spending allocated to goods produced in region i is

$$\pi_{ni,t} = \frac{(\tau_{in}w_{i,t}^{\eta_1}r_{i,t}^{\eta_2}P_{i,t}^{\eta_3})^{-\theta}A_{i,t}^{\theta}}{\sum_{m=1}^{N+1}(\tau_{mn}w_{m,t}^{\eta_1}r_{m,t}^{\eta_2}P_{m,t}^{\eta_3})^{-\theta}A_{m,t}^{\theta}} \tag{1.20}$$

1.3.6 Equilibrium

The distribution of labor $\{L_t\}$ characterizes the endogenous state of the economy during each time period. Time-invariant fundamentals of the economy include stock of land $\{\overline{H_n}\}$, exogenous regional productivity fundamentals $\{\overline{A_n}\}$, amenity fundamentals $\{B_n\}$, bilateral trade costs $\{\tau_{in}\}$, migration costs $\{\kappa_{in}\}$ and travel time $\{\iota_{in}\}$. The parameters assumed constant across all time include the final consumption expenditure share on goods (α), share of labor costs, land costs, intermediate goods costs in production inputs (η_1, η_2, η_3), trade elasticity (θ), discount factor (β), migration elasticity (ν), magnitude of production agglomeration force (ρ) and spatial decay rate of productivity spillovers (δ).

Denote $\overline{\Theta}$ as the time-invariant fundamentals. Then a *temporary equilibrium* is a vector of wages $w(L_t, \overline{\Theta})$ that satisfies the equilibrium conditions of the static subproblem outlined in (1.13)-(1.20). The temporary equilibrium is identical to the solution of a static new economic geography model. A *sequential competitive equilibrium* is a sequence of $\{L_t, \mu_t, V_t, w_t(L_t, \overline{\Theta})\}_{t=0}^{\infty}$ that solves equilibrium conditions (1.5) to (1.7) and the temporary equilibrium at each t given $(L_0, \overline{\Theta})$. μ_t, V_t, L_0 are the migration shares, lifetime utilities and initial labor distribution respectively. Finally, a *stationary equilibrium* is a sequential competitive equilibrium such that $\{L_t, \mu_t, V_t, w_t(L_t, \overline{\Theta})\}_{t=0}^{\infty}$ are constant for all t. In the stationary equilibrium, no endogenous variables change over time. Note that labor distribution staying unchanged in the stationary equilibrium simply means the labor inflow and outflow of each market exactly offset each other. It does not suggest that workers stop moving. In other words, stationary equilibrium could still have positive and even large gross labor flows between cities as long as the net labor flows remain zero.

1.4 Parameterization

To link the model to data, there are eight parameters to be determined $\{\alpha, \eta_1, \eta_2, \eta_3, \theta, \beta, \nu, \delta, \rho\}$. I calibrate the consumption expenditure share on goods α to China Statistical Yearbook (0.85 in 2010) and match the shares of various inputs in production, η_1, η_2 and η_3, to the 2010 inter-regional input output table (0.32, 0.05 and 0.63 respectively). I set the productivity dispersion parameter or trade elasticity θ to 4 following Simonovska and Waugh (2014) and the discount factor β for a 5-year period to 0.8. I estimate the migration elasticity ν, magnitude of production agglomeration force ρ and spatial decay rate of productivity spillovers δ based on the following methods.

1.4.1 Migration Elasticity ν

Unlike trade elasticity, there is not a commonly used value for the income elasticity of domestic migration in developing countries to the best of my knowledge. To estimate such a migration elasticity ν, I follow Artuç, Chaudhuri and McLaren (2010) (henceforth ACM) to derive an estimating equation from the structural model that relates migration flows in the current period to the spot utilities in the following period as well as future migration flows[19]:

$$ln(\frac{\mu_{ni,t}}{\mu_{nn,t}}) = Const_1 + \beta\nu(u_{i,t+1} - u_{n,t+1}) + \beta ln(\frac{\mu_{ni,t+1}}{\mu_{nn,t+1}}) + \epsilon_{t+1} \tag{1.21}$$

$\mu_{ni,t}$ is the share of workers in prefecture n migrating to prefecture i at the end of period t, $u_{i,t+1}$ is the spot utilities for working in prefecture i during period $t+1$, β is the predetermined discount factor and ν is the variable of interest.

There are three issues to be addressed before bringing Equation (1.21) to the data. First, the migration data only contain province-to-prefecture migration flow information whereas the two labor markets, i and n, in Equation (1.21) are supposed to be at the same aggregation level[20]. Simply treating a prefecture and a province as the destination and origin regions would require additional assumptions on how to aggregate prefecture-level variables to province-level. To avoid enforcing extra assumptions, I transform the estimation equation so that it compares the migration flows from a province to two prefectures, namely:

$$ln(\frac{\mu_{p_k \to i,t}}{\mu_{p_k \to n,t}}) = Const_2 + \beta\nu(u_{i,t+1} - u_{n,t+1}) + \beta ln(\frac{\mu_{p_k \to i,t+1}}{\mu_{p_k \to n,t+1}}) + \epsilon_{t+1} \tag{1.22}$$

where p_k denotes any province, and i and n are two distinct prefectures in the same province

[19] See the online appendix of ACM for derivation details.

[20] The migration data consist of the 2000, 2010 China National Population Census and the 2005, 2015 Population Micro Census (1% random sample survey). I assume migration shares for workers are the same as those for total population due to data availability.

other than p_k.

The second issue lies in the data availability of the spot utilities term $u_{i,t+1}$[21]. By Equation (1.4), $u_{i,t+1}$ is composed of a prefecture-specific amenity fundamentals term B_i, nominal income per worker $v_{i,t+1}$ and price index $\mathbb{P}_{i,t+1} = P_{i,t+1}^{\alpha} r_{i,t+1}^{1-\alpha}$. The goods market clearing condition assures that nominal income in the model and nominal GDP per worker are equivalent and the latter is observable from the data. However, there is no data counterpart for B_i. Also the CPI data only document CPI changes across years for each prefecture and are not useful for between-prefecture comparison in a given year. To overcome these data shortages, I modify the previous estimation equation by replacing the variables with their intertemporal changes as following[22]:

$$ln(\frac{\dot{\mu}_{p_k \to i,t}}{\dot{\mu}_{p_k \to n,t}}) = Const_3 + \beta\nu ln(\frac{\dot{v}_{i,t+1}/\dot{\mathbb{P}}_{i,t+1}}{\dot{v}_{n,t+1}/\dot{\mathbb{P}}_{n,t+1}}) + \beta ln(\frac{\dot{\mu}_{p_k \to i,t+1}}{\dot{\mu}_{p_k \to n,t+1}}) + \varpi_{t+1} \tag{1.23}$$

where $\dot{x}_t = x_t / x_{t-1}$. Taking intertemporal changes in the spot utilities cancels out the unobservable time-invariant amenity fundamentals and utilizes the CPI movements of each prefecture which are readily available from the data. Intuitively, migration flow changes contain information on expected values that depend on future real income changes and the changes in option value of migration across markets, while the latter are reflected by future migration flow changes. Equation (1.23) is my preferred estimation equation and will be used in the empirical analysis.

The third issue is that the residuals term, ϖ_{t+1}, contains the shock revealed in $t+1$ and hence is highly likely to be correlated with the regressors. I propose two instrumental variables, both of which are supposedly not correlated with the new information revealed in $t+1$. The first IV is lagged migration flow changes $\dot{\mu}_{p_k \to i,t-1}$. As in ACM, my model implies past values of endogenous variables are valid instruments. For the second IV, I use a Bartik-style expected income instrument based on national average income changes by sector weighted by each prefecture's employment distribution across sectors. In other words, $\varsigma_{i,t+1} = \sum_{j=1}^{3} l_{i,t}^{j} \dot{\bar{v}}_{t+1}^{j}$ instruments for prefecture i's income changes using its employment share of sector j in the starting year, $l_{i,t}^{j}$, and national average income changes in the three sectors, $\dot{\bar{v}}_{t+1}^{j}$[23]. Table 1.3 shows the intertemporal changes of migration shares between eight aggregate regions in China[24]. Table 1.4 reports the employment shares across the three sectors averaging over prefectures within each of the aggregate region.

After dealing with these issues, I implement the estimation strategy by selecting Year 2010 as

[21] In ACM's model, the spot utilities term is composed of wages only and hence there is no data availability issue.

[22] See Appendix A.3 for detailed derivation.

[23] I rely on the 2005 Population Micro Census for prefecture-level employment shares by detailed industry and aggregate them to the three sectors: primary sector (agriculture), secondary sector (construction, utilities and manufacturing) and tertiary sector (service). See Appendix for a list of industries by sectors.

[24] The eight regions of China are: Northeast, North Municipalities, North Coast, Central Coast, South Coast, Central, Northwest and Southwest. See Appendix for a list of provinces by region.

Table 1.3: Internal Migration Shares of China

Origin	Destination							
	North-east	North Muni	North Coast	Central Coast	South Coast	Central	North-west	South-west
2005-2010/2000-2005								
Northeast	0.99	1.58	1.31	1.68	2.46	1.84	2.82	2.23
North Municipalities	1.92	0.99	1.38	1.09	1.43	1.13	2.07	0.92
North Coast	1.83	1.90	0.99	2.00	2.29	2.28	3.05	1.77
Central Coast	2.12	1.39	1.32	0.99	2.12	1.23	1.95	1.14
South Coast	2.36	2.06	2.10	1.62	1.00	1.13	2.14	0.94
Central	2.01	1.80	1.61	1.56	2.48	0.95	2.03	1.82
Northwest	1.55	1.94	1.47	2.25	2.32	2.20	0.99	1.68
Southwest	2.30	1.42	1.10	1.45	2.91	1.78	1.77	0.95
2010-2015/2005-2010								
Northeast	1.00	1.09	0.74	0.96	0.53	1.50	1.01	1.50
North Municipalities	1.33	0.99	1.41	1.20	0.92	1.70	1.79	1.30
North Coast	0.75	1.13	1.00	0.89	0.43	1.44	0.98	1.53
Central Coast	0.65	1.18	0.99	1.00	0.63	1.37	1.03	1.98
South Coast	0.73	0.84	0.98	0.73	1.00	1.12	1.09	1.27
Central	0.84	1.20	1.19	0.72	0.27	1.05	0.98	1.17
Northwest	0.93	1.12	1.01	0.98	0.31	1.29	1.00	1.42
Southwest	1.44	1.48	1.39	0.69	0.27	1.29	0.72	1.06

Notes: Table shows the intertemporal changes of share of population in each of the origin region migrating to each of the destination region. Reported values are the ratios between current five-year migration shares and previous five-year migration shares. Data sources: 2010 National Population Census of China and 2005, 2015 Population Micro Census.

the base year ($t = Year_{2010}$) and five years as the time interval ($t + 1 = Year_{2015}$). I restrict the destination prefectures in the sample to be non-autonomous, leading to a nonrepetitive sample of 39,741 observations[25]. Each observation consists of an origin province and two destination prefectures.

I obtain a coefficient of 0.75 for $\beta\nu$, implying $\nu = 0.94$ (standard deviation 0.09). I use this number in my empirical analysis below. Though to the best of my knowledge there is no benchmark value for this five-year migration elasticity, this estimate is in line with the literature. ACM and Caliendo et al. (2019), for example, both study the inter-state inter-sector mobility in the United States with a dynamic labor market model and empirically employ lagged migration flows and wages as instruments. Their estimates based on the March CPS correspond to a counterpart of ν being 0.53 and 0.50 respectively at an annual frequency. My estimate at the five-year frequency is larger than theirs. Intuitively, lower frequencies lead to a larger elasticity of migration flows to

[25]Non-autonomous prefectures are prefectures consisting most of the Han Chinese, the majority ethnicity group in China. 304 out of the 334 prefectures in my data are non-autonomous.

Table 1.4: Employment Share across Sectors

	Primary Sector	Secondary Sector	Tertiary Sector
Region			
Northeast	0.20	0.42	0.38
North Municipalities	0.02	0.43	0.55
North Coast	0.13	0.55	0.32
Central Coast	0.10	0.53	0.37
South Coast	0.17	0.45	0.38
Central	0.18	0.46	0.36
Northwest	0.21	0.42	0.37
Southwest	0.26	0.38	0.36

Notes: Reported employment shares are the average across prefectures within each region as of 2005.

changes in income, thus larger ν. Tombe and Zhu (2019) is also comparable. They work on the 2000 and 2005 migration across provinces and sectors in China and obtain a migration elasticity corresponding to the ν in my model being 1.48. It is reasonable that they have a larger migration elasticity as they identify migrants as workers with a current residence location different from birthplace, whereas I work on a five-year migration frequency.

1.4.2 Joint Estimation of Productivity Spillover Parameters ρ and δ

A strong productivity spillover between cities could be due to a large regional agglomeration force (large ρ) or a less frictional spatial transfer (small δ), while my data are not sufficient to disentangle these two channels. Thus I follow a structural estimation strategy similar to the one in Ahlfeldt et al. (2015) to jointly estimate ρ and δ. The idea is to first derive a closed-form solution for the production fundamentals $\overline{A_n}$ with only observable data and model parameters, and then use it to develop moment conditions that exploit the exogenous variation in prefectural HSR linkage time.

Specifically, I start with the equilibrium trade share equation (1.20) and show that

$$ln(\frac{X_{nF,t}}{\widetilde{X_{F,t}}}) = -\theta ln(\frac{\tau_{nF}}{\widetilde{\tau_F}}) - \theta\eta_1 ln(\frac{w_{n,t}}{\widetilde{w_t}}) - \theta\eta_2 ln(\frac{r_{n,t}}{\widetilde{r_t}}) - \theta\eta_3 ln(\frac{P_{n,t}}{\widetilde{P_t}}) + \theta ln(\frac{A_{n,t}}{\widetilde{A_t}}) \tag{1.24}$$

where $X_{nF,t}$ is the export value of prefecture n in year t (F denotes ROW) and all the denominators with tilde are geometric mean of the respective variables (e.g. $\widetilde{X_{F,t}} = exp\{\frac{1}{N}\sum_{i=1}^{N} ln\widetilde{X_{iF,t}}\}$). Taking time differences (to difference out unobservable trade costs) and combining with the specification (1.11) that regional productivity $A_{n,t}$ is composed of production fundamentals $\overline{A_n}$

and production externalities $\Upsilon_{n,t}$, I derive the following closed-form solution for $\overline{A_n}$:

$$\Delta ln(\frac{\overline{A_{n,t}}}{\widetilde{A_t}}) = \underbrace{\eta_1 \Delta ln(\frac{w_{n,t}}{\widetilde{w}_t}) + \eta_2 \Delta ln(\frac{r_{n,t}}{\widetilde{r}_t}) + \eta_3 \Delta ln(\frac{P_{n,t}}{\widetilde{P}_t}) + \frac{1}{\theta}\Delta ln(\frac{X_{nF,t}}{\widetilde{X_{F,t}}})}_{\Delta ln(A_{n,t}/\widetilde{A_t})} - \rho \Delta ln(\frac{\Upsilon_{n,t}(\delta)}{\widetilde{\Upsilon_t(\delta)}}) \tag{1.25}$$

The first three terms on the right-hand side are factor prices changes and the fourth term is export value changes which are all observable from data[26]. The last term is the production externality changes where by specification (1.12) all components are observable except ρ and δ—the two parameters to be estimated. Therefore, equation (1.25) reveals that the changes in production fundamentals are one-to-one functions of observable data and model parameters[27]. This relationship is robust to time-invariant factors (by taking time differences) and year fixed effects that are common across all cities (by dividing by geometric means).

Next I build moment conditions which impose that cities do not have systematic changes in their production fundamentals after controlling for the exogenous HSR linkage time. In other words, for cities linked by HSR during the same period, the systematic changes in their regional productivities, $\Delta ln(A_{n,t}/\widetilde{A_t})$, are fully explained by the changes in their production externalities, $\Delta ln(\Upsilon_{n,t}/\widetilde{\Upsilon_t})$. I capture the exogenous HSR linkage time by dividing the cities into three groups: connected to the HSR network by 2010, connected between 2010 and 2015, and not connected by 2015. I thus generate three moment conditions, each of which imposes the group mean of the production fundamental changes is close to zero. Or formally, the moment conditions are

$$\mathbb{E}[\mathbb{I}_k \times \Delta ln(\overline{A}_{n,t}/\widetilde{\overline{A}}_t)] = 0 \tag{1.26}$$

where $\mathbb{I}_k : k \in \{1, 2, 3\}$ are the indicators for the three city groups. I then implement the Generalized Method of Moments (GMM) to jointly estimate ρ and δ.

Except jointly pinning down two parameters, the GMM strategy here is essentially similar to an IV estimation that estimates the productivity spillover parameters by regressing regional productivity changes on production externality changes while instrumenting with HSR linkage time. The production fundamental changes are acting as the structural residuals. The estimation equation could be written as

$$\Delta ln(A_{n,t}/\widetilde{A_t}) = \rho \Delta ln(\Upsilon_{n,t}(\delta)/\widetilde{\Upsilon_t(\delta)}) + \epsilon \tag{1.27}$$

with $\epsilon = \Delta ln(\overline{A_{n,t}}/\widetilde{\overline{A_t}})$ and the instrument variable being $\mathbb{I}_k$. The exclusion restriction for the

[26] Empirically, I use time difference of the corresponding variables between 2010 and 2015.

[27] In the model, production fundamentals $\overline{A_n}$ do not change across time. Here I allow the empirical counterparts of $\overline{A_n}$ to fluctuate between 2010 and 2015 and treat the changes as structural residuals to be minimized.

Table 1.5: Calibrated Model Parameters

Parameter	Value	Description
Calibrated Independently		
α	0.85	Share of goods in consumption
(η_1, η_2, η_3)	(0.32, 0.05, 0.63)	Share of factor inputs in production
θ	4	Trade elasticity
β	0.8	Discount factor
Calibrated in Equilibrium		
ν	0.94	Migration elasticity
ρ	0.056	Magnitude of production agglomeration
δ	0.058	spatial decay rate of productivity spillover
κ_{ni}	Table 1.6	Migration costs
τ_{ni}	Table 1.6	Trade costs

IV estimation, which in this case plays the same role as the moment conditions in the GMM estimation, is that the HSR linkages affect regional productivities only through their impact on production externalities. Or HSR linkages are uncorrelated with production fundamental changes. As the HSR network is planned and administered by the Chinese central government, it is plausible that the HSR linkages are exogenous to the local cities.

With the above GMM estimation strategy, I find substantial and statistically significant production agglomeration force with an estimated $\rho = 0.056$ (standard deviation 0.006). The estimated spatial decay rate of productivity spillover between cities is $\delta = 0.058$ (standard deviation 0.012)[28], suggesting for cities that are 60 minutes away 97% of the production externalities would decay on the road and for cities that are 120 minutes away the productivity spillovers are negligible. My estimate of the elasticity of productivity with respect to city size ($\rho = 0.056$) is within the range of 0.02 to 0.10 generally reported in the literature, e.g., Rosenthal and Strange (2004), Melo et al. (2009), Combes et al. (2012) and Ahlfeldt et al. (2015). My estimate of spatial decay rate of productivity spillovers ($\delta = 0.058$) is also consistent with the literature studying geographical scope of knowledge transmission, e.g., Conley et al. (2003) find that knowledge transmission between people vanishes when they are 90-120 minutes away.

1.5 Inferring Migration and Trade Costs

In this section, I first estimate the inter-prefecture domestic migration costs using the 2010 National Population Census of China. The 2010 National Population Census asks participants for

[28]The standard deviation calculations for ρ and δ do not account for spatial or serial correlation and hence are likely under-biased. I conduct a sensitivity analysis on the value of δ in Section 1.7.2, but leave a more robust estimation for these two productivity spillover parameters for future research.

the prefecture they currently reside in and the province they lived in five years ago. In other words, it only documents the province-to-prefecture migration flows. To infer the prefecture-to-prefecture migration frictions needed for the quantitative analysis, I use a nested nonlinear least square procedure to jointly estimate migration costs and prefecture-specific lifetime utility based on the strategies in Fan (2019). Similarly, the 2010 Interregional Input-Output table used as trade data is at provincial level while the quantitative analysis requires prefectural-level trade information. To solve this issue, I jointly infer the inter-prefectural trade costs and prefecture-specific production unit costs.

1.5.1 Migration Costs

I first specify the migration cost of moving from prefecture n to prefecture i as

$$ln(\kappa_{ni}) = \sum_{k=1}^{4} \lambda_k I_k + \lambda_5 dist_{ni} + \lambda_6 Cdist_{ni} + residual \tag{1.28}$$

where $I_1 - I_4$ are dummy variables. I_1 indicates if n and i belong to different prefectures within the same province. I_2 indicates if they belong to different provinces within the same large region. I_3 indicates if they belong to different large regions. I_4 indicates if they belong to adjacent provinces (provinces sharing a border). $dist_{ni}$ is the great-circle distance and $Cdist_{ni}$ is the historical cultural distance between the two prefectures. $Cdist_{ni}$ is smaller if they have similar compositions of ethnic minorities[29]. Migration costs of stayers are normalized to be 0: $ln(\kappa_{nn}) = 0$.

I then use a nested nonlinear least square procedure to jointly estimate Equation (1.28) and recover the location-specific lifetime utilities in 2010, $\{V_{i,2010}\}$. The procedure consists of two loops. In the inner loop, I choose $\{V_{i,2010}\}$ so that given migration frictions $\{\kappa_{ni}\}$ and migration elasticity ν the labor distribution of each prefecture matches the data. Intuitively, the higher lifetime utilities a prefecture provides the more workers it attracts. In the outer loop, I choose $\{\lambda\}$ to minimize the deviations of province-to-prefecture migration flows in the model from the 2010 Population Census data.

Specifically, the inner loop takes a guess of $\{\kappa_{ni}\}$ and solves for $\{V_{i,2010}\}$ based on

$$L_{i,2010}^{data} = \sum_{n=1}^{N} \frac{V_{i,2010}^{\beta\nu} \kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,2010}^{\beta\nu} \kappa_{nm}^{-\nu}} L_{n,2005}^{data} \tag{1.29}$$

which can be derived from Equation (1.6) and (1.7). The migration elasticity ν is determined in Section 1.4.1 and the discount factor β is preset to 0.8. Proposition 1 assures the feasibility of this

[29]The historical cultural distance is constructed as $1 - corr(C_n, C_i)$, where C_n is a vector representing the ethnic composition of prefecture n in the 1990 National Population Census. I use Fan (2019)'s cultural distance data.

step by proving the existence and uniqueness of $\{V_{i,2010}\}$ (up to a normalization).

Proposition 1. *Given migration costs $\{\kappa_{ni}\}$, migration elasticity ν and discount factor β, there exists a unique set of $\{V_{i,2010}\}$ (up to a normalization) such that the model-predicted labor distribution matches the data, i.e. Equation (1.29) holds.*

Proof. See Appendix. □

The outer loop aggregates the migration origins to the provincial level and compares it with the data. Or formally, it minimizes the following loss function

$$min_{\{\lambda\}} \sum_{p_k \in P, n} (ln(\sum_{i \in p_k} \mu_{in,2005} L^{data}_{i,2005}) - ln(\mu^{data}_{p_k \to n,2005} L^{data}_{p_k,2005}))^2$$

where $p_k \in P$ denotes the individual province in China, i and n denote the prefectures, $\mu_{in,2005} = \frac{V^{\beta\nu}_{i,2010}\kappa^{-\nu}_{ni}}{\sum_{m=1}^{N} V^{\beta\nu}_{m,2010}\kappa^{-\nu}_{nm}}$ is the migration share calculated from the inner loop.

1.5.2 Trade Cost

The trade cost estimation strategy is similar to the one used for migration cost estimation. I first specify the iceberg domestic trade cost of shipping a good from prefecture n to i with the following log-linear function form:

$$ln(\tau_{ni}) = \sum_{k=1}^{4} \gamma_k I_k + \gamma_5 dist_{ni} + residual \tag{1.30}$$

where $I_1 - I_4$ and $dist_{ni}$ are the same variables as in the migration cost estimation equation. Trade costs of within-prefecture trade are normalized to be 0: $ln(\tau_{nn}) = 0$. All trade between Chinese prefectures and the ROW need to go through one of the port prefectures[30]. I specify the trade cost between a Chinese inland prefecture and the ROW as the sum of two components: the trade cost between that prefecture and its nearest port prefecture, and an export-friction parameter ζ that captures tariff and non-tariff barriers to international trade. Or formally, $ln(\tau_{inland,ROW}) = ln(\tau_{inland,port}) + \zeta$.

To quantify $\{\tau_{ni}\}$, I again use a nonlinear least square procedure that consists of two loops. In the inner loop, I choose location-specific production unit costs, $\{c_{i,2010}\}$, so that the final good markets clear and the expenditures of each prefecture consist with the data. Intuitively, the smaller unit costs firms in a prefecture incur the more total output they would have. In the outer loop, I

[30] I identify 39 port prefectures according to Export Ports and Routes of China (Zhang 2005). Refer to Appendix for the list.

choose $\{\gamma_1 - \gamma_5, \zeta\}$ to minimize the deviations of model-predicted inter-provincial trade flows and prefectural international trade from their data counterpart.

Specifically, the inner loop takes a guess of $\{\tau_{ni}\}$ and solves for $\{c_{i,2010}\}$ based on

$$X_{n,2010}^{data} = \sum_{i=1}^{N+1} \frac{c_{n,2010}^{-\theta}\tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} c_{m,2010}^{-\theta}\tau_{mi}^{-\theta}} X_{i,2010}^{data} - S_{n,2010}^{data} \tag{1.31}$$

where $X_{n,2010}$ is the total expenditure of prefecture n in year 2010 that can be backed out from the GDP data[31]. $S_{n,2010}$ is n's trade surplus given by data and θ is the trade elasticity predetermined to be 4. Proposition 2 supports the feasibility of this step by proving the existence and uniqueness of $\{c_{i,2010}\}$ (up to a normalization).

Proposition 2. *Given trade costs $\{\tau_{ni}\}$, trade surplus $\{S_{n,2010}\}$ and trade elasticity θ, there exists a unique set of $\{c_{i,2010}\}$ (up to a normalization) such that the final good markets clear and the model-predicted expenditures of all regions match the data, i.e. Equation (1.31) holds.*

Proof. See Appendix. □

The outer loop aggregates the inter-prefectural trade to inter-provincial trade and then approaches it as well as the prefectural import and export to the data. Or formally, it minimizes the following loss function

$$\begin{aligned} min_{\{\gamma\},\zeta} \sum_{p_k,p_j \in P} (ln(\sum_{i\in p_k, n\in p_j} \pi_{in,2010} X_{i,2010}^{data}) - ln(X_{p_j\to p_k,2010}^{data}))^2 \\ + \sum_i (ln(\pi_{iROW,2010} X_{i,2010}^{data}) - ln(X_{ROW\to i,2010}^{data}))^2 \\ + \sum_i (ln(\pi_{ROWi,2010} X_{ROW,2010}^{data}) - ln(X_{i\to ROW,2010}^{data}))^2 \end{aligned}$$

where $p_k, p_j \in P$ denote the individual provinces in China. i and n denote the prefectures in p_k and p_j respectively. $\pi_{in,2010} = \frac{c_{n,2010}^{-\theta}\tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} c_{m,2010}^{-\theta}\tau_{mi}^{-\theta}}$ is the trade share calculated from the inner loop. The inter-provincial trade $X_{p_j\to p_k,2010}$, prefectural import $X_{ROW\to i,2010}$ and prefectural export $X_{i\to ROW,2010}$ are from the data.

1.5.3 Estimation Results

Table 1.6 reports the migration cost and trade cost estimation results. As shown in the first column, all coefficients in the migration cost estimation model are statistically significant. The coefficients being positive except the dummy capturing moving between cities in adjacent provinces suggests

[31] Cobb-Douglas production function together with the zero-profit condition gives $X_{n,t} = \frac{GDP_{n,t}}{\eta_1+\eta_2}$.

that institutional, geographic and cultural barriers all impede the domestic labor mobility in China. Migrating an additional 1000 kilometers adds the migration costs by 80 log points. Increasing the cultural distance from 25^{th} to 75^{th} percentile leads to another 26 log points raise in migration costs given the interquartile range of culture distance being 0.5. The institutional barrier for moving between provinces, either within the same region (712 log points) or across regions (692 log points), is much stronger than moving within province (31 log points). Such a large migration friction for crossing a provincial border is expected due to the existence of *Hukou*. Compared to moving between provinces in the same region, moving across regions is more affected by the geographic barrier due to longer distance but slightly less affected by the institutional barrier. However, the difference in the institutional barrier is so small that it is dominated by the frictions for crossing provincial border. This again is expected as the *Hukou* policy is administered and implemented at the provincial level rather than at the regional level.

My migration cost estimates, especially for the institutional barriers, are larger than those of Fan (2019) who also estimates inter-prefecture migration frictions in China using a similar nested nonlinear least square procedure. The main reason to the differences is that my migration elasticity estimate is smaller than Fan (0.94 compared to 4). As migration elasticity and bilateral migration costs jointly determine the migration frictions, a smaller migration elasticity naturally leads to larger migration cost estimates.

The second column of Table 1.6 presents trade cost estimates. The coefficients being statistically significant and positive except the dummy capturing trade between cities in adjacent provinces suggest that institutional and geographic barriers raise the domestic trade costs in China. The institutional barriers for trade across provinces or regions, 92 and 94 log points respectively, are much larger than that for trade between cities within a province, 15 log points. The geographic distance increases trade costs by 18 log points with each additional 1000 kilometers. These trade friction coefficients are in line with the literature studying interregional trade within China (e.g., Fan 2019). Compared to trade barrier estimates for the U.S. (e.g., Crafts and Klein 2014), the larger institutional barrier estimates here suggest that China as a developing country still faces outstanding barriers to trade flows at provincial border.

1.6 Solution Algorithm

The entire dynamic sequential competitive equilibrium is characterized by the following set of equations:

$$L_{i,t} = \sum_{n=1}^{N} \frac{V_{i,t}^{\beta\nu} \kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,t}^{\beta\nu} \kappa_{nm}^{-\nu}} L_{n,t-1} \tag{1.32}$$

Table 1.6: Estimates of Migration Costs and Trade Costs

	Migration	Trade
I_1 (Different Prefectures, Same Province)	0.31	0.15
	(0.13)	(0.08)
I_2 (Different Provinces, Same Region)	7.12	0.92
	(0.13)	(0.08)
I_3 (Different Regions)	6.92	0.94
	(0.09)	(0.05)
I_4 (Adjacent Provinces)	-0.33	-0.15
	(0.08)	(0.05)
I_5 (Great-circle Distance)	0.80	0.18
	(0.05)	(0.03)
I_6 (Culture Distance)	0.53	
	(0.11)	
ζ (Export Friction)		1.13
		(0.04)
Observations	9990	1566
R^2	0.54	0.58

Notes: Standard errors in parentheses. *Great-circle Distance* is measured in 1000 km. *Cultural Distance* is measured as one minus the correlation in historical ethnic minority shares between cities. For trade cost estimation, R^2 is 0.69 when matching the inter-provincial IO data and 0.43 when matching the prefectural international trade data.

$$V_{n,t}^{\nu} = (\frac{1+\eta_2/\eta_1-\chi_n}{\alpha})^{\nu}(w_{n,t}\xi_{n,t})^{\nu}[\sum_{i}^{N} V_{i,t+1}^{\nu}\kappa_{ni}^{-\nu}]^{\beta} \tag{1.33}$$

$$w_{n,t}L_{n,t} = \sum_{i=1}^{N+1} \frac{c_{n,t}^{-\theta}\tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} c_{m,t}^{-\theta}\tau_{mi}^{-\theta}}(1-\eta_1\chi_i)w_{i,t}L_{i,t} \tag{1.34}$$

$$c_{i,t} = \frac{w_{i,t}^{\eta_1} r_{i,t}^{\eta_2} P_{i,t}^{\eta_3}}{A_{i,t}} \tag{1.35}$$

$$A_{i,t} = \overline{A_i}[\sum_{s=1}^{N} e^{-\delta\iota_{si}}(\frac{L_{s,t}}{\overline{H_s}})]^{\rho} \tag{1.36}$$

$$\alpha r_{n,t}\overline{H_n} = (1-\alpha+\frac{\eta_2}{\eta_1})w_{n,t}L_{n,t} - (1-\alpha)\chi_n w_{n,t}L_{n,t} \tag{1.37}$$

$$\xi_{n,t} = \frac{B_n}{P_{n,t}^{\alpha} r_{n,t}^{1-\alpha}} \tag{1.38}$$

$$P_{n,t} \propto [\sum_{i=1}^{N+1} (\tau_{in} w_{i,t}^{\eta_1} r_{i,t}^{\eta_2} P_{i,t}^{\eta_3})^{-\theta} A_{i,t}^{\theta}]^{-1/\theta} \tag{1.39}$$

Only the first two equations involve intertemporal changes in the state variables and hence are the key to capture the dynamic feature of the labor market. The remaining equations concern solely with the temporary equilibrium at each time period t.

Variables and parameters determined outside this solution algorithm are elasticities $(\nu, \theta, \rho, \delta)$, share of goods in consumption and shares of various inputs in production $(\alpha, \eta_1, \eta_2, \eta_3)$, discount factor (β), land area $(\overline{H_i})$, ratio of trade surplus to wage income (χ_i), bilateral migration frictions, trade frictions and travel time $(\kappa_{ni}, \tau_{ni}, \iota_{ni})$, as well as the initial distribution of state variables $(L_{i,0}, V_{i,0}, w_{i,0}, c_{i,0})$. Assuming these variables and parameters are known and the economy takes $T \geq 0$ periods to converge to the steady state, we can jointly solve the equilibrium paths for the endogenous variables of interest $(\{L_{i,t}\}_{t=1}^T, \{w_{i,t}\}_{t=1}^T, \{A_{i,t}\}_{t=1}^T)$ as well as identify production fundamentals and amenities $(\overline{A_i}, B_i)$ with the following steps (see Appendix for details). Note that in my empirical analysis I calibrate $t = 0$ to the 2010 China economy and set each time interval to be five years, nevertheless, the solution algorithm presented here is not specific to my context and could well fit other economies and time frequencies.

I. Solve $\overline{A_i}$ using initial observables. Land market clearing condition ensures that the initial rental price $r_{i,0}$ is identified once wage and labor distribution are known (Eq (1.37)). The zero profit condition (Eq (1.34-1.35)) together with the aggregate price index decomposition (Eq (1.39)) pins down the initial price level $P_{i,0}$ up to a normalization, which can then be plugged back into Eq (1.35) to get initial regional productivities $A_{i,0}$. By Eq (1.36), $\overline{A_i}$ can be identified by dividing regional productivities by the productivity agglomeration force.

II. Solve steady state equilibrium. First make a guess for the region-specific time-invariant amenities $B_i^{(1)}$[32]. The "(1)" superscript in this section indicates the variable is conditional on the first iteration of B. We will deal with the identification of B_i in Step IV.

i) Guess steady state wage distribution $w_i^{*(1)}$ [33] and find the steady state labor distribution $L_i^{*(1)}$ that agrees with it by guessing and updating $L_i^{*(1)}$ in a nonlinear fixed point problem[34].

[32]Normalize $B_i^{(1)}$ as the dynamic labor market equation is homogeneous of degree 0 in B.

[33]Normalize $w_i^{*(1)}$ as the labor demand equation is homogeneous of degree 1 in w and the dynamic labor market equation is homogeneous of degree 0 in w.

[34]Solution algorithm for the nonlinear fixed point problem works as following: i) Guess steady state labor distribution $L_i^{*(1)}$. ii) Use Eq (1.36)-(1.37) to calculate the productivities $\widehat{A_i^*}$ and rental price $\widehat{r_i^*}$ that consist with the guessed wage and labor distribution. iii) Plug $\{w_i^{*(1)}, \widehat{A_i^*}, \widehat{r_i^*}\}$ into Eq (1.39) and solve a nonlinear fixed point problem for the price level for final goods $\widehat{P_i^*}$ up to a normalization. iv) Use Eq (1.38) to calculate the price-adjusted amenities

ii) Based on the computed labor distribution, calculate the wage needed to satisfy the labor demand condition (Eq (1.34)). Check if it is close to the wage guess from Step i). If not, update the wage guess and repeat Step i)-ii) until convergence. Denote the resulting steady state variable values as $\{w_{i,ss}^{(1)}, L_{i,ss}^{(1)}, V_{i,ss}^{(1)}\}$.

III. Solve for the dynamic transition. The model suggests that given $B_i^{(1)}$, the labor distribution should start from $L_{i,0}$ and converge to $L_{i,ss}^{(1)}$ after a sufficiently long T periods ($L_{i,t=0}^{(1)} = L_{i,0}, L_{i,t=T}^{(1)} = L_{i,ss}^{(1)}$).

i) Guess the full transition path of labor distribution $\{L_{i,t}^{(1)}\}_{t=1}^{T-1}$ and find the wage path $\{w_{i,t}^{(1)}\}_{t=1}^{T-1}$ that consists with it by guessing and updating $\{w_{i,t}^{(1)}\}_{t=1}^{T-1}$ in a nonlinear fixed point problem. Compute the price-adjusted amenities $\{\xi_{i,t}^{(1)}\}_{t=1}^{T-1}$ based on the labor and wage path.

ii) For each $0 < t \leq T-1$, use $w_{i,t}^{(1)}, \xi_{i,t}^{(1)}, V_{i,t+1}^{(1)}$ and Eq (1.33) to solve backwards for $V_{i,t}^{(1)}$[35]. Then use $V_{i,t+1}^{(1)}, L_{i,t+1}^{(1)}$ and the dynamic labor market equation (Eq (1.32)) to solve backwards for $L_{i,t}^{(1)}$. This delivers a new transition path for labor distribution. Compare the new path with the guess in Step i). Check if these two paths are close, if not, update the guess and repeat Step i) and ii) until convergence.

IV. Solve the amenity fundamentals B_i. Step II and III equip us with an algorithm to solve for the full dynamic path of $\{\{L_{i,t}^{(1)}\}_{t=1}^{T}, \{w_{i,t}^{(1)}\}_{t=1}^{T}, \{V_{i,t}^{(1)}\}_{t=1}^{T}\}$ given any guess for amenity fundamentals $B_i^{(1)}$. To identify B_i, we approach the initial lifetime utilities suggested by the above dynamic transition path $V_{i,0}^{(1)}$ to the values calibrated to the observed data $V_{i,0}$. Specifically, combining Eq (1.33) and (1.38) gives

$$B_i = \frac{\alpha}{1+\eta_2/\eta_1}\left(\frac{V_{n,0}^{\nu}}{[\sum_i^N V_{i,1}^{\nu}\kappa_{ni}^{-\nu}]^{\beta}}\right)^{1/\nu} / \frac{w_{i,0}}{P_{i,0}^{\alpha} r_{i,0}^{1-\alpha}} \tag{1.40}$$

All variables on the right-hand side of Eq (1.40) are from the data or calibrated directly to the data except $V_{i,1}$. Intuitively, after controlling for real income, migrants are more attracted to the cities with better amenities or higher expected option values of relocating in the future. The final step is to replace $V_{i,1}$ with $V_{i,1}^{(1)}$ computed in Step III and obtain a new value for B_i. Check if it is close to the initial guess $B_i^{(1)}$. If not, update the guess for B and repeat Step II-IV until convergence.

$\widehat{\xi_i^*}$. v) Plug $\{w_i^{*(1)}, \widehat{\xi_i^*}\}$ into Eq (1.33) and solve a nonlinear fixed point problem for the lifetime utility $\widehat{V_i^*}$ up to a normalization. vi) Use $\widehat{V_i^*}$ and compute the right-hand side of Eq (1.32) as $\widehat{L_i^{*(1)}}$. Check if $\widehat{L_i^{*(1)}}$ is close to the guess from Step i). If not, update the guess and repeat Step ii) to vi) until convergence. Refer to Appendix for more details.

[35]Note that we start from $t = T-1$ where $V_{i,t+1}^{(1)} = V_{i,ss}^{(1)}$ and $L_{i,t+1}^{(1)} = L_{i,ss}^{(1)}$ are known from Step II.

1.7 Counterfactual Analysis

In this section, I undertake two counterfactuals to explore quantitatively the role of productivity spillovers across cities in the distribution of economic activities in mainland China. The first exercise aims to quantify the distributional impacts of the HSR network completed before 2015 on mainland China's economy in the medium and long run. The second exercise assumes a counterfactual world where productivity spillovers decay 30% slower ($\delta_{new} = 0.7 * \delta$) and examines how labor distribution (L), regional productivities (A) and real income (v/P) would respond.

Before conducting these two exercises, I first characterize a baseline economy to 2010 mainland China following the solution algorithm outlined in Section 1.6. The baseline economy assumes all the model parameters $\{\alpha, \eta_1, \eta_2, \eta_3, \theta, \beta, \nu, \delta, \rho\}$, production fundamentals and residential amenities $\{\{\overline{A_i}\}, \{B_i\}\}$ as well as the migration and trade costs, $\{\{\kappa_{ni}\}, \{\tau_{ni}\}\}$, to take the 2010 value and stay constant thereafter, and the inter-city passenger travel time $\{\iota_{ni}\}$ to take the 2015 value and stay constant thereafter. The initial year is 2010 ($t = 0$ indicates $Year_{2010}$) and each time interval is five years (e.g. $t = 1$ indicates $Year_{2015}$).

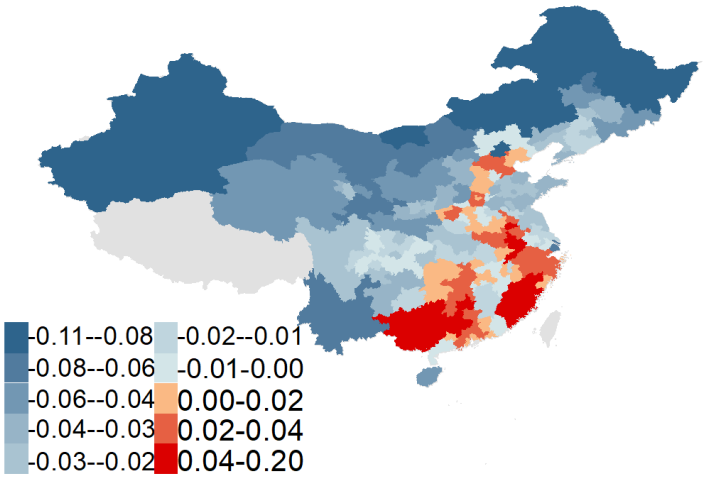
1.7.1 Distributional Impacts of HSR

The first quantitative exercise studies the impacts of HSR on labor distribution, regional productivities and real income in mainland China by answering the following question: what would have happened differently across China if there were no HSR? Specifically, I simulate a counterfactual economy where all the exogenous variables and initial endogenous variables are identical to the baseline economy except that the inter-city passenger travel time stays forever at the pre-HSR level. I then solve for both the long run equilibrium (steady state) and the transition path of such a no-HSR economy.

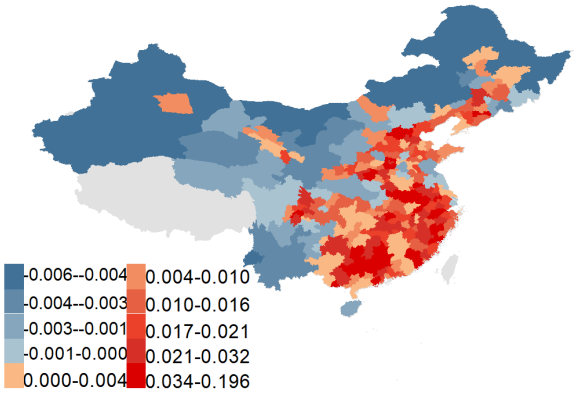
Figure 1.3 presents the long run differences in labor, regional productivity and real income by comparing the baseline economy (with HSR) to the counterfactual economy (without HSR) for the 333 Chinese prefectural cities. At the aggregate level, HSR reshuffles the labor distribution of mainland China by affecting the location choice of 1.33% of the total workforce or 10.11 million workers. It also enlarges the between-city inequality in regional productivities and real income by 1.94% and 3.16% respectively based on interquartile range (IQR) calculation.

In particular, Figure 1.3(a) shows that HSR incentivizes workers in the northeastern and northwestern regions to migrate to and settle in the southern, central and southeastern regions. For instance, the three northeastern provinces (Heilongjiang, Jilin and Liaoning) have a mean value of -0.05 across their 36 prefectures, indicating 5% of their workforce will relocate elsewhere due to HSR. Guangxi province located along the southern coast, on the other hand, gains the most among

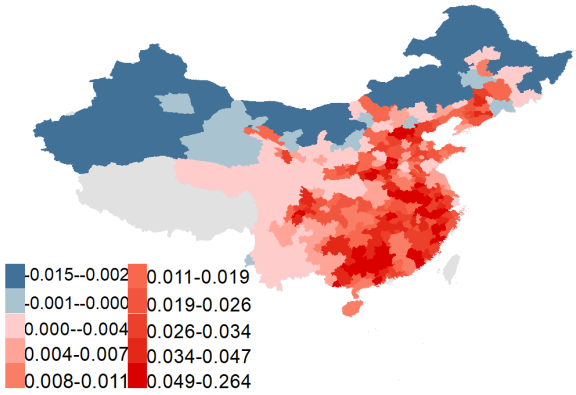
Figure 1.3: Distributional Impacts of HSR



(a) Labor



(b) Regional Productivity

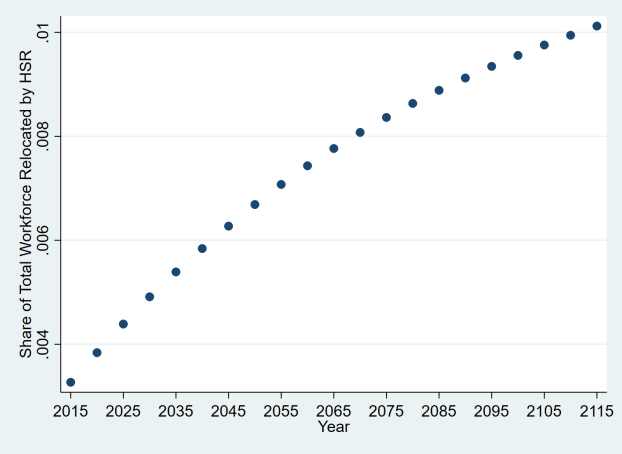


(c) Real Income

Notes: The figures present the long run differences in labor (Panel a), regional productivity (Panel b) and real income (Panel c) by comparing the baseline economy to the counterfactual no-HSR economy for the 333 prefectural cities. Red shaded areas are with positive values and blue shaded areas are with negative values. Gray shaded areas are excluded from the analysis due to limited data. The 10 legends correspond to the deciles (with the decile closest to zero adjusted to zero to separate positive and negative values).

all provinces with its 14 prefectures on average creating 8.59% more jobs due to HSR. The labor relocation pattern generally matches the distributional impacts of HSR on regional productivities and real income per worker as shown in Figure 1.3(b) and 1.3(c) respectively. In the southern, central and southeastern regions, HSR substantially strengthens the productivity spillovers across cities as both prefectural cities and HSR routes are densely located. Such agglomeration effects assure higher regional productivities which lead to higher income and hence attract migrants.

Figure 1.4: Impacts of HSR on Labor Relocation



Note: The figure presents the impact of HSR on labor relocation measured as the number of workers with different location choices between the baseline economy (with HSR) and the counterfactual economy (without HSR) over total workforce.

The western and northern regions, on the contrary, tend to have cities being so far away from each other that the inter-city passenger travel time remains considerable even after HSR. The productivity spillovers across cities there are thus much less significant.

Figure 1.4 illustrates the dynamic impacts of HSR on labor relocation by plotting the share of total workforce choosing different prefectural cities between the baseline economy (with HSR) and the counterfactual economy (without HSR). My quantitative results indicate that HSR is expected to change the location choice of 0.49% of total workforce (3.73 million workers) in 20 years and 0.74% of total workforce (5.64 million workers) in 50 years.

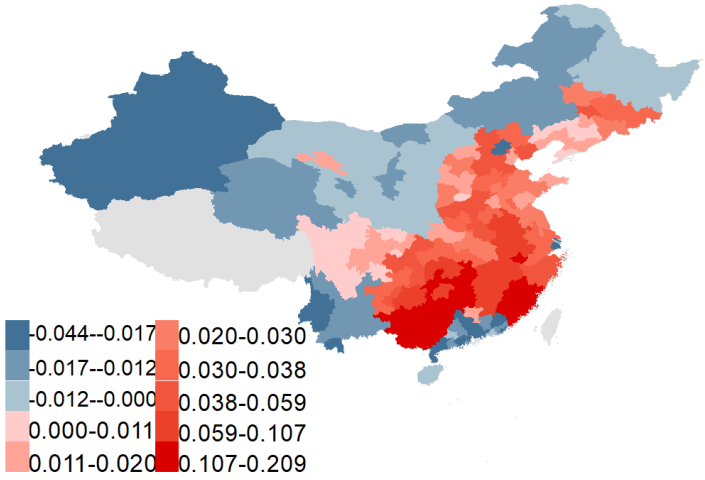
1.7.2 Sensitivity Analysis on the Strength of Productivity Spillovers

In the second quantitative exercise, I examine the influence of the strength of productivity spillovers on distribution of economic activities by constructing a counterfactual world where the productivity spillovers across cities are stronger than those estimated in Section 1.4.2. Specifically, I reduce the spatial decay rate of productivity spillovers, δ, by 30%, recompute the associated *stationary equilibrium* based on Section 1.6 and compare the resulting economy to the baseline economy[36].

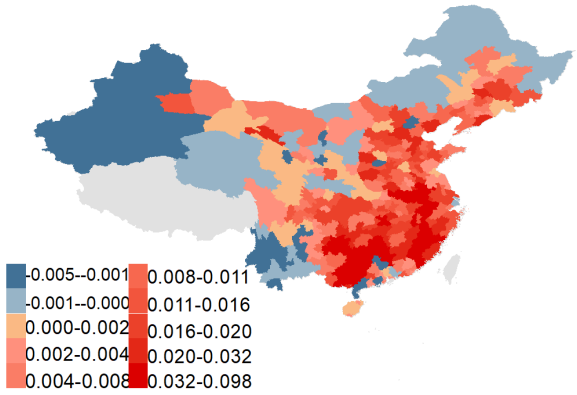
The results indicate that at the aggregate level, a 30% decrease in spatial decay rate of productivity spillovers affects the location choice of 1.43% of the total workforce or 10.85 million workers. It also enlarges the between-city inequality in regional productivities and real income by 2.49% and 1.58% respectively based on interquartile range (IQR) calculation. Figure 1.5 presents the prefectural-level differences between the counterfactual economy ($\delta_{new} = 0.041$) and the baseline

[36]With $\delta_{new} = 0.041$, cities that are 60 minutes away have 91% of the production externalities between them decay on the road and cities that are 120 minutes away remain to have negligible productivity spillovers.

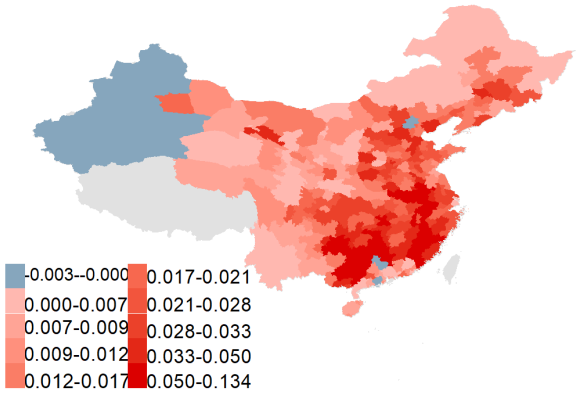
Figure 1.5: Slower Spatial Decay Rate of Productivity Spillovers



(a) Labor



(b) Regional Productivity



(c) Real Income

Notes: The figures present the long run differences in labor (Panel a), regional productivity (Panel b) and real income (Panel c) by comparing the counterfactual economy where spatial decay rate of productivity spillovers is reduced by 30% to the baseline economy for the 333 prefectural cities. Red shaded areas are with positive values and blue shaded areas are with negative values. Gray shaded areas are excluded from the analysis due to limited data. The 10 legends correspond to the deciles (with the decile closest to zero adjusted to zero to separate positive and negative values).

economy ($\delta = 0.058$). Stronger productivity spillovers again benefit the southern, central and southeastern cities more than the northeastern and northwestern cities. This finding is expected as cities are more densely located in the southern regions and hence the productivity spillovers across cities are amplified—similar reasoning as the first counterfactual exercise on the distributional impacts of HSR.

1.8 Conclusion

This paper establishes that productivity spillovers across cities are non-negligible by using the rapid expansion of high-speed rail (HSR) in China as a natural experiment. HSR shortens inter-city passenger travel time, makes face-to-face communication easier and thus facilitates knowledge spillovers. With a dynamic spatial general equilibrium model, I quantify the distributional impacts of the HSR network completed in mainland China before 2015 and find that HSR will affect the location choice of 1.33% of the total workforce in the long run. It benefits southern and southeastern regions where both cities and HSR routes are densely located substantially more than the northern or western regions in terms of labor inflow, regional productivities and real income.

It is worth pointing out the following limitations when interpreting the quantitative results. First, the analysis does not account for the impacts of HSR on migration frictions with the argument that the migration frictions in mainland China are largely driven by its migration policy, *Hukou*, rather than transportation costs. However, as discussed by several studies on the economic geography of China (e.g. Fan 2019 and You and Wu 2020), the Chinese government is easing the *Hukou* policy especially in small- and middle-size cities during recent years. The actual role that HSR plays in migration is likely more important than assumed in this project. Second, this paper assumes HSR has no impacts on trade costs as it does not allow freight transportation. This is admittedly a strong assumption as HSR facilitates information transmission and is likely lowering trade costs for goods and services. Therefore, this paper's quantitative results regarding the contributions of HSR on mainland China are recommended to be taken as a lower bound.

There are some other interesting and important topics about the influence of HSR on the distribution of economic activities. For example, HSR might have heterogeneous impacts across different sectors or different types of workers. Also a more comprehensive understanding of a large-scale infrastructure project as HSR requires a thorough cost-benefit analysis. I abstract from these discussions in this paper and leave them to future research.

Bibliography

[1] Artuç, E., Chaudhuri, S. and McLaren, J. 2010. "Trade Shocks and Labor Adjustment: A Structural Empirical Approach". *American Economic Review*, 100(3), 1008–1045.

[2] Au, C. and Henderson, J.V. 2006. "Are Chinese Cities Too Small?". *The Review of Economic Studies*, 73(3), 549–576.

[3] Ahlfeldt, G.M. and Feddersen, A. 2018. "From Periphery to Core: Measuring Agglomeration Effects Using High-speed Rail". *Journal of Economic Geography*, 18(2), 355-390.

[4] Ahlfeldt, G.M., Redding, S.J., Sturm D.M. and Wolf, N. 2015. "The Economics of Density: Evidence From the Berlin Wall". *Econometrica*, 83(6), 2127-2189.

[5] Allen, T. and Arkolakis, C. 2014. "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 129 (3), 1085–1140.

[6] Artuç, E., Chaudhuri, S. and McLaren, J. 2010. "Trade Shocks and Labor Adjustment: A Structural Empirical Approach". *American Economic Review*, 100(3), 1008-45.

[7] Banerjee, A., Duflo, E. and Qian, Q. 2020. "On the road: Access to transportation infrastructure and economic growth in China", *Journal of Development Economics*, 145, 102442.

[8] Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M.A. and Zhang, Q. 2017. "Roads, Railroads and Decentralization of Chinese Cities". *Review of Economics and Statistics*, 99(3), 435-448.

[9] Baum-Snow, N., Henderson, J.V., Turner, M.A. and Zhang, Q. 2019. "Does Investment in National Highways Help or Hurt Hinterland City Growth?". *Journal of Urban Economics*, 115: 103124.

[10] Bernard, A.B., Moxnes, A. and Saito Y.U. 2019. "Production Networks, Geography, and Firm Performance". *Journal of Political Economy*, 127(2), 639-688.

[11] Bian, Y., W, L. and Bai, J. 2019. "Does High-speed Rail Improve Regional Innovation in China?", *Journal of Financial Research*, 468(6), 132-149.

[12] Caliendo, L., Dvorkin, M. and Parro, F. 2019. "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock". *Econometrica*, 87(3), 741-835.

[13] Chan, K.W. 2010. "The Household Registration System and Migrant Labor in China: Notes on a Debate. *Population and Development Review*, 36(2), 357-364.

[14] Charnoz, P., Lelarge, C. and Trevien, C. 2018. "Communication Costs and the Internal Organisation of Multi-plant Businesses: Evidence from the Impact of the French High-speed Rail". *The Economic Journal*, 128(610), 949-994.

[15] Combes, P., Duranton, G., Gobillon, L., Puga, D. and Roux, S. 2012. "The Productivity Advantages of Large Cities: Distinguishing Agglomeration from Firm Selection". *Econometrica*, 80(6), 2543–2594.

[16] Conley, T., Flyer, F., and Tsiang, G. 2003. "Spillovers from local market human capital and the spatial distribution of productivity in Malaysia". *Advances in Economic Analysis and Policy*, 3(1), 1-47.

[17] Crafts, N. and Klein, A. 2015. "Geography and intra-national home bias: U.S. domestic trade in 1949 and 2007". *Journal of Economic Geography*, 15(3), 477–497.

[18] Desmet, K. and Rossi-Hansberg, E. 2013. "Urban Accounting and Welfare". *American Economic Review*, 103(6), 2296-2327.

[19] Donaldson, D. and Hornbeck, R. 2016. "Railroads and American Economic Growth: a 'Market Access' Approach". *Quarterly Journal of Economics*, 131(2), 799-858.

[20] Dong, X., Zheng, S. and Kahn, M. 2020. "The role of transportation speed in facilitating high skilled teamwork across cities". *Journal of Urban Economics*, 115: 103212

[21] Eaton, J., Kramarz, F. and Kortum, S. 2019. "Firm-to-Firm Trade: Exports, Imports, and the Labor Market". *Society for Economic Dynamics*, Meeting Paper.

[22] Faber, B. 2014. "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System". *Review of Economic Studies*, 81(3), 1046-70.

[23] Fan, J. 2019. "Internal geography, labor mobility, and the distributional impacts of trade". *American Economic Journal: Macroeconomics*, 11(3), 252-88.

[24] Fujita, M., Krugman, P. and Venables, A.J. 1999."The Spatial Economy: Cities, Regions, and International Trade". *The MIT Press*, Cambridge, MA.

[25] Glaeser, E.L. and Gottlieb, J.D. 2009. "The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States", *Journal of Economic Literature*, 47(4), 983-1028.

[26] 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, 315–348.

[27] Hayakawa, K., Koster, H., Tabuchi, T. and Thisse, J.. 2021. "High-speed Rail and the Spatial Distribution of Economic Activity: Evidence from Japan's Shinkansen", *Research Institute of Economy, Trade and Industry (RIETI)*, Discussion papers 21003.

[28] Head, K. and Mayer, T. 2006. "Regional wage and employment responses to market potential in the EU", *Regional Science and Urban Economics*, 36, 573–594.

[29] Ji, Y. and Yang, Q. 2019. "Can the High-Speed Rail Service Promote Enterprise Innovation? A Study Based on Quasi-natural Experiments", *The Journal of World Economy*, 2, 147-166.

[30] Li, Y. 2016. "China High Speed Railways and Stations (2016)". https://doi.org/10.7910/DVN/JIISNB, Harvard Dataverse, V1

[31] Lin, Y. 2017. "Travel costs and urban specialization patterns: Evidence from China's high speed railway system". *Journal of Urban Economics*, 98, 98-123.

[32] Lucas, R.E. and Rossi-Hansberg, E. 2002. "On the internal structure of cities". *Econometrica*, 70(4), 1445–76

[33] Marshall, A. 1890. "Principles of economics", Macmillan, London, UK.

[34] Melo, P., Graham, D. and Noland, R. 2009. "A Meta-analysis of Estimates of Urban Agglomeration Economies". *Regional Science and Urban Economics*, 39(3), 332-342.

[35] Michaels, G., Rauch, F. and Redding S.J. 2012. "Urbanization and Structural Transformation". *The Quarterly Journal of Economics*, 127(2), 535–586.

[36] International Union of Railways. 2018. "High Speed Rail: Fast Track to Sustainable Mobility".

[37] Isard, W. 1951. "Interregional and Regional Input-Output Analysis: A Model of a Space-Economy". *The Review of Economics and Statistics*, 33(4), 318-328.

[38] Monte, F., Redding, S.J. and Rossi-Hansberg, E. 2018. "Commuting, Migration, and Local Employment Elasticities". *American Economic Review*, 108(12), 3855-3890.

[39] Krugman, P. 1991. "Increasing Returns and Economic Geography". *Journal of Political Economy*, 99(3), 483-99.

[40] Ollivier, G., Sondhi, J. and Zhou, N. 2014. "High-Speed Railways in China: A Look at Construction Costs". *China Transport Topics No. 9. World Bank, Beijing*.

[41] Ottaviano, Gianmarco I. P. 2008. "Infrastructure and economic geography: An overview of theory and evidence", EIB Papers, ISSN 0257-7755, European Investment Bank (EIB), Luxembourg, 13(2), 8-35.

[42] Redding, S.J. 2016. "Goods Trade, Factor Mobility and Welfare". *Journal of International Economics*, 101, 148-167.

[43] Redding, S.J. and Venables, A.J. 2004. "Economic geography and international inequality", *Journal of International Economics*, 62, 53-82.

[44] Redding, S.J. and Rossi-Hansberg, E. 2017. "Quantitative Spatial Economics". *Annual Review of Economics*, 9, 21-58.

[45] Rosenthal, S. and Strange, W.C. 2004. "Evidence on the Nature and Sources of Agglomeration Economies". *Handbook of Regional and Urban Economics*, 4, 2119-2171.

[46] Simonovska, I. and Waugh, M.E. 2014. "The Elasticity of Trade: Estimates and Evidence". *Journal of International Economics*, 92(1), 34-50.

[47] Sotelo, S. 2020. "Domestic Trade Frictions and Agriculture". *Journal of Political Economy*, 128(7), 2690-2738.

[48] State Council of the People's Republic of China. 2004. "The Medium- and Long-Term Railway Network Plan".

[49] Tian, L. and Yu, Y. 2021. "Geographic Spillovers and Firm Exports: Evidence from China". Working Paper.

[50] Tombe, T. and Zhu, X. 2019. "Trade, Migration, and Productivity: A Quantitative Analysis of China". *American Economic Review*, 109(5), pp. 1843-72

[51] Wu, W. and You, W. 2020. "The Welfare Implications of Internal Migration Restrictions: Evidence from China". Working paper.

[52] Xu, M. 2018. "Riding on the New Silk Road: Quantifying the Welfare Gains from High-speed Railways". Working paper.

[53] Yu, Y., Zhuang, H., Liu, D. and Fu, Y. 2019. "Does the Opening of High-Speed Rail Accelerate the Spillover of Technological Innovation? Evidence from 230 Prefecture-level Cities in China". *Journal of Finance and Economics*, 45(11), 20-31.

[54] Zhang, T. 2005. "Export Ports and Routes of China". *University of International Business and Economic Press*.

1.9 Appendix

1.9.1 Theoretical Appendix

Derivation of Dynamic Labor Market Equation (Eq (1.6))

The lifetime utility for a worker o working in market n at the end of period t is

$$ln(V^o_{n,t}) = Max_i\ U(B_n, C^o_{n,t}, H^o_{n,t}) + \beta E(ln(V^o_{i,t+1})) - ln(\kappa_{ni}) + \frac{1}{\nu}\epsilon^o_{i,t}$$

The homogeneous consumption preference assumption assures that $C^o_{n,t}$ and $H^o_{n,t}$ are the same for all workers in the same market and hence we can replace $U(B_n, C^o_{n,t}, H^o_{n,t})$ with $U_{n,t}$. The expected lifetime utility of a random worker in n at t then becomes

$$ln(V_{n,t}) = E[ln(V^o_{n,t})] = E[Max_i\ U_{n,t} + \beta ln(V_{i,t+1}) - ln(\kappa_{ni}) + \frac{1}{\nu}\epsilon^o_{i,t}]$$

To find $ln(V_{n,t})$, I assume that the idiosyncratic shock ϵ is i.i.d. over time and distributed based on the Gumbel distribution, specifically with a CDF $F(\epsilon) = exp(-(exp(-\epsilon - \overline{\gamma}))$ where $\overline{\gamma} = \int_{-\infty}^{\infty} xexp(-x - exp(-x))dx$ is the Euler–Mascheroni constant. Let $X_{ni,t} = U_{n,t} + \beta ln(V_{i,t+1}) - ln(\kappa_{ni})$, then

$$\begin{aligned}
Pr[Max_i\ X_{ni,t} + \frac{1}{\nu}\epsilon^o_{i,t} <= x] &= \prod_i Pr[X_{ni,t} + \frac{1}{\nu}\epsilon^o_{i,t} <= x] \\
&= exp\{\sum_i lnPr[X_{ni,t} + \frac{1}{\nu}\epsilon^o_{i,t} <= x]\} \\
&= exp\{\sum_i lnPr[\epsilon^o_{i,t} <= \nu(x - X_{ni,t})]\} \\
&\overset{\text{cdf for } \epsilon}{=} exp\{\sum_i -exp[\nu(X_{ni,t} - x)]\} \\
&= exp\{-exp[-\frac{x - \frac{1}{\nu}ln\sum_i exp(\nu X_{ni,t})}{1/\nu}]\}
\end{aligned}$$

The fourth equation uses the definition of CDF for Gumbel distribution and drops the constant term. The last equation is identical to the CDF of a Gumbel distribution with a location parameter $\frac{1}{\nu}ln\sum_i exp(\nu X_{ni,t})$ and a scale parameter $\frac{1}{\nu}$. The expectation of it can be written as $\frac{1}{\nu}ln\sum_i exp(\nu X_{ni,t}) + \frac{\overline{\gamma}}{\nu}$. Substitute the computed expectation into the expected lifetime utility

equation to get

$$ln(V_{n,t}) = \frac{1}{\nu} ln \sum_i exp[\nu U_{i,t} + \beta\nu ln(V_{i,t+1}) - \nu ln(\kappa_{ni})]$$

From the model setup, we know $ln(V_{n,t}) = U_{n,t} + \beta ln(V_{n,t+1})$ and hence

$$\begin{aligned} ln(V_{n,t}) = U_{n,t} + \beta ln(V_{n,t+1}) &= U_{n,t} + \beta(\frac{1}{\nu} ln \sum_i exp[\nu U_{i,t+1} + \beta\nu ln(V_{i,t+2}) - \nu ln(\kappa_{ni})] \\ &= U_{n,t} + \beta(\frac{1}{\nu} ln \sum_i exp[\nu ln(V_{i,t+1}) - \nu ln(\kappa_{ni}))] \end{aligned}$$

Rearrange, take exponentials and replace $U_{n,t}$ with $u_{n,t}$ we get

$$V_{n,t}^{\nu} = [exp(u_{n,t})]^{\nu} [\sum_i V_{i,t+1}^{\nu} \kappa_{ni}^{-\nu}]^{\beta}$$

With the specification $u_{n,t} = ln(\frac{B_n v_{n,t}}{P_{n,t}^{\alpha} r_{n,t}^{1-\alpha}})$, the above equation can be further expanded as

$$V_{n,t}^{\nu} = [\frac{B_n v_{n,t}}{P_{n,t}^{\alpha} r_{n,t}^{1-\alpha}}]^{\nu} [\sum_i V_{i,t+1}^{\nu} \kappa_{ni}^{-\nu}]^{\beta}$$

Derivation of Migration Share Equation (Eq (1.7))

Share of workers in region n that migrate to region i at the end of period t is equal to the probability that the expected lifetime utility of moving to region i exceeds that of any other region. Or formally,

$$\begin{aligned} \mu_{ni,t} &= Pr\{E[\beta ln(V_{i,t+1}^o) - ln(\kappa_{ni}) + \frac{1}{\nu}\epsilon_{i,t}^o] \geq E(max_{m \neq i} \beta ln(V_{m,t+1}^o) - ln(\kappa_{nm}) + \frac{1}{\nu}\epsilon_{m,t}^o)\} \\ &= Pr\{\beta ln(V_{i,t+1}) - ln(\kappa_{ni}) + \frac{1}{\nu}\epsilon_{i,t} \geq max_{m \neq i} \beta ln(V_{m,t+1}) - ln(\kappa_{nm}) + \frac{1}{\nu}\epsilon_{m,t}\} \end{aligned}$$

By assuming the idiosyncratic shocks are distributed Gumbel as in previous section, we get

$$\mu_{ni,t} = \int_{-\infty}^{\infty} f(\epsilon_{i,t}) \prod_{m \neq i} F\{\beta\nu[ln(V_{i,t+1}) - ln(V_{m,t+1})] - \nu[ln(\kappa_{ni}) - ln(\kappa_{nm})] + \epsilon_{i,t}\} d\epsilon_{i,t}$$

Let $\lambda_{i,t} = log \sum_m exp[-\beta\nu(lnV_{i,t+1} - lnV_{m,t+1}) - \nu[ln(\kappa_{ni}) - ln(\kappa_{nm})]$ and $\widehat{y}_{i,t} = \epsilon_{i,t} + \overline{\gamma} - \lambda_{i,t}$ where $\overline{\gamma}$ is again the Euler–Mascheroni constant. Applying the definition for Gumbel distribution,

the above equation can then be written as

$$\begin{aligned}\mu_{ni,t} &= \int_{-\infty}^{\infty} exp(-\epsilon_{i,t} - \overline{\gamma}) exp[-exp(-\epsilon_{i,t} - \overline{\gamma}) exp(\lambda_{i,t})] d\epsilon_{i,t} \\ &= \int_{-\infty}^{\infty} exp(-\widehat{y}_{i,t} - \lambda_{i,t}) exp[-exp(-\widehat{y}_{i,t} - \lambda_{i,t} + \lambda_{i,t})] d\widehat{y}_{i,t} \\ &= exp(-\lambda_{i,t}) \int_{-\infty}^{\infty} exp(-\widehat{y}_{i,t} - exp(-\widehat{y}_{i,t})) d\widehat{y}_{i,t}\end{aligned}$$

The integral term in the last equation equals to 1 since it fits the standard Gumbel distribution with $\widehat{y}_{i,t}$ being the random variable. Plugging back $\lambda_{i,t}$, we obtain

$$\begin{aligned}\mu_{ni,t} &= exp(-log \sum_{m} exp[-\beta\nu(lnV_{i,t+1} - lnV_{m,t+1}) - \nu[ln(\kappa_{ni}) - ln(\kappa_{nm})]) \\ &= -\sum_{m} exp[ln(\frac{V_{i,t+1}^{\beta\nu} \kappa_{ni}^{-\nu}}{V_{m,t+1}^{\beta\nu} \kappa_{nm}^{-\nu}})] \\ &= \frac{V_{i,t+1}^{\beta\nu} \kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,t+1}^{\beta\nu} \kappa_{nm}^{-\nu}}\end{aligned}$$

Derivation of Migration Elasticity Estimation Equation (Eq (1.23))

Following ACM, we can derive an estimation equation that relates migration flows in the current period to the spot utilities in the following period as well as future migration flows:

$$ln(\frac{\mu_{ni,t}}{\mu_{nn,t}}) = Const_1 + \beta\nu(u_{i,t+1} - u_{n,t+1}) + \beta ln(\frac{\mu_{ni,t+1}}{\mu_{nn,t+1}}) + \epsilon_{t+1}$$

Since the data only contain province-to-prefecture migration information, we transform the above equation so that it compares the migration flows from a province to two distinct prefectures:

$$ln(\frac{\mu_{p_k \to i,t}}{\mu_{p_k \to n,t}}) = Const_2 + \beta\nu(u_{i,t+1} - u_{n,t+1}) + \beta ln(\frac{\mu_{p_k \to i,t+1}}{\mu_{p_k \to n,t+1}}) + \epsilon_{t+1}$$

where p_k denotes any province, and i and n denote distinct prefectures. With the specification for $u_{i,t+1}$, we can further express the equation as

$$ln(\frac{\mu_{p_k \to i,t}}{\mu_{p_k \to n,t}}) = Const_2 + \beta\nu ln(\frac{\frac{B_i v_{i,t+1}}{\mathbb{P}_{i,t+1}}}{\frac{B_n v_{n,t+1}}{\mathbb{P}_{n,t+1}}}) + \beta ln(\frac{\mu_{p_k \to i,t+1}}{\mu_{p_k \to n,t+1}}) + \epsilon_{t+1}$$

where $\mathbb{P}_{i,t+1} = P_{i,t+1}^{\alpha} r_{i,t+1}^{1-\alpha}$. Similarly, we have the relationship holds for the previous period

$$ln(\frac{\mu_{p_k \to i,t-1}}{\mu_{p_k \to n,t-1}}) = Const_{2'} + \beta\nu ln(\frac{\frac{B_i v_{i,t}}{\mathbb{P}_{i,t}}}{\frac{B_n v_{n,t}}{\mathbb{P}_{n,t}}}) + \beta ln(\frac{\mu_{p_k \to i,t}}{\mu_{p_k \to n,t}}) + \epsilon_t$$

Subtracting these two equations gives

$$ln(\frac{\mu_{p_k \to i,t}/\mu_{p_k \to i,t-1}}{\mu_{p_k \to n,t}/\mu_{p_k \to n,t-1}}) = Const_3 + \beta\nu ln(\frac{\frac{B_i v_{i,t+1}}{\mathbb{P}_{i,t+1}}/\frac{B_i v_{i,t}}{\mathbb{P}_{i,t}}}{\frac{B_n v_{i,t+1}}{\mathbb{P}_{i,t}+1}\frac{B_n v_{n,t}}{\mathbb{P}_{n,t}}}) + \beta ln(\frac{\mu_{p_k \to i,t+1}/\mu_{p_k \to i,t}}{\mu_{p_k \to n,t+1}\mu_{p_k \to n,t}}) + \varpi_{t+1}$$

Or,

$$ln(\frac{\dot{\mu}_{p_k \to i,t}}{\dot{\mu}_{p_k \to n,t}}) = Const_3 + \beta\nu ln(\frac{\dot{v}_{i,t+1}/\dot{\mathbb{P}}_{i,t+1}}{\dot{v}_{n,t+1}/\dot{\mathbb{P}}_{n,t+1}}) + \beta ln(\frac{\dot{\mu}_{p_k \to i,t+1}}{\dot{\mu}_{p_k \to n,t+1}}) + \varpi_{t+1}$$

where $\dot{x}_t = x_t/x_{t-1}$.

Proof of Propositions

Proposition 1. *Given migration costs $\{\kappa_{ni}\}$, migration elasticity ν and discount factor β, there exists a unique set of $\{V_{i,2010}\}$ (up to a normalization) such that the model-predicted labor distribution matches the data, i.e. $L_{i,2010}^{data} = \sum_{n=1}^{N} \frac{V_{i,2010}^{\beta\nu}\kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,2010}^{\beta\nu}\kappa_{nm}^{-\nu}} L_{n,2005}^{data}$ holds.*

Proof. By Eq (6) and (7) we know that

$$L_{i,t+1} = \sum_{n=1}^{N} \mu_{ni,t} L_{n,t}$$

where $\mu_{ni,t} = \frac{V_{i,t+1}^{\beta\nu}\kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,t+1}^{\beta\nu}\kappa_{nm}^{-\nu}}$. Setting $t+1 = 2010$ and five-year as the time interval, we can tell $L_{i,t+1}$ and $L_{i,t}$ from the data. Migration costs $\{\kappa_{ni}\}$, migration elasticity ν and discount factor β are also predetermined and hence the only unknowns in the equation is $\{V_{i,2010}\}$.

Define the worker deficit in each labor market i as $D_i(V_{i,t+1}) = L_{i,t+1} - \sum_{n=1}^{N} \mu_{ni,t} L_{n,t}$. To prove Proposition 1, I first show the following four conditions are met:

1. $\{D_i(V_{i,t+1})\}$ is continuous.
2. $\{D_i(V_{i,t+1})\}$ is homogeneous of degree 0.
3. $\sum_{i=1}^{N} D_i(V_{i,t+1}) = 0, \forall V_{i,t+1} \in \mathbb{R}_+^N$.
4. $\{D_i(V_{i,t+1})\}$ exhibits gross substitute property.

Condition 1 is satisfied by construction. Condition 2 is easily satisfied by noticing μ is homogeneous

of degree 0 in V. Condition 3 can be shown by noticing $\sum_{i=1}^{N} \mu_{ni,t} = 1, \forall n$ and $\sum_{i=1}^{N} D_i(V_{i,t+1}) = \sum_{i=1}^{N} L_{i,t+1} - \sum_{i=1}^{N}\sum_{n=1}^{N} \mu_{ni,t} L_{n,t} = \sum_{i=1}^{N} L_{i,t+1} - \sum_{n=1}^{N} L_{n,t} = 0$. To show Condition 4, I calculate the derivatives:

$$\frac{\partial D_i(V_{i,t+1})}{\partial V_{i,t+1}} = -\sum_{n=1}^{N} \frac{\partial \mu_{ni,t}}{\partial V_{i,t+1}} L_{n,t} = -\sum_{n=1}^{N} \frac{\beta\nu V_{i,t+1}^{\beta\nu-1} \kappa_{ni}^{-\nu} \sum_{m\neq i} V_{m,2010}^{\beta\nu} \kappa_{nm}^{-\nu}}{(\sum_{m=1}^{N} V_{m,t+1}^{\beta\nu} \kappa_{nm}^{-\nu})^2} L_{n,t} < 0$$

$$\frac{\partial D_i(V_{i,t+1})}{\partial V_{m,t+1}} = -\sum_{n=1}^{N} \frac{\partial \mu_{ni,t}}{\partial V_{m,t+1}} L_{n,t} = -\sum_{n=1}^{N} \frac{-\beta\nu V_{i,t+1}^{\beta\nu} \kappa_{ni}^{-\nu} V_{m,2010}^{\beta\nu} \kappa_{nm}^{-\nu}}{(\sum_{m=1}^{N} V_{m,t+1}^{\beta\nu} \kappa_{nm}^{-\nu})^2} L_{n,t} > 0$$

The next steps of the proof follow Michaels, Redding and Rauch (2012) and Ahlfeldt et al. (2015). Condition 1 and 2 guarantee the existence of a solution. By Condition 2, we can normalize $\{V_{i,t+1}\}$ to the simplex $\{V_{i,t+1} \in \mathbb{R}_+, \forall i : \sum_{i=1}^{N} V_{i,t+1} = 1\}$. Construct $D_i(V_{i,t+1})^+ = max\{0, D_i(V_{i,t+1})\}$ and $f(V_{i,t+1}) = \frac{V_{i,t+1} + D_i^+(V_{i,t+1})}{\sum_i (V_{i,t+1} + D_i^+(V_{i,t+1}))}$, then f serves as a continuous function mapping the unit simplex onto itself. The existence of a solution to $V_{i,t+1} = f(V_{i,t+1})$ then follows from the Brouwer's fixed point theorem. Condition 3 and 4 guarantee the uniqueness of the solution (refer to Ahlfeldt et al. (2015) for detailed discussion).

□

Proposition 2. *Given trade costs $\{\tau_{ni}\}$, trade surplus $\{S_{n,2010}\}$ and trade elasticity θ, there exists a unique set of $\{c_{i,2010}\}$ (up to a normalization) such that the final good markets clear and the model-predicted expenditures of all regions match the data, i.e. $X_{n,2010} = \sum_{i=1}^{N+1} \frac{c_{n,2010}^{-\theta} \tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} c_{m,2010}^{-\theta} \tau_{mi}^{-\theta}} X_{i,2010} - S_{n,2010}^{data}$ holds.*

Proof. Similar to the proof for Proposition 1, while here we prove the existence and uniqueness of $\{c_{i,2010}\}$. □

1.9.2 Empirical Appendix

Net Domestic Migration in China

Table 1.7: Domestic Migration in China

	mean	25%	50%	75%	max
Prefecture-level					
Intra-province Migrant Stock (millions)	0.18	0.04	0.09	0.18	2.50
Intra-province Migrant Share	0.03	0.02	0.03	0.06	0.22
Inter-province Migrant Stock (millions)	0.19	0.02	0.04	0.10	4.30
Inter-province Migrant Share	0.02	0.00	0.01	0.03	0.26
Province-level (in absolute value)					
Net Inter-province Migrant Flows (millions)	1.8	0.5	1.38	2.61	8.67
Net Inter-province Migrant Share	0.04	0.02	0.03	0.05	0.16
United States (in absolute value)					
Net Inter-state Migrant Share	0.005	0.001	0.002	0.005	0.06

Notes: Migrants are defined as citizens with a different residence prefecture in China (county in U.S.) from five years ago. Data sources: 2015 Population Micro Census and the 2011-2015 American Community Survey (bottom row).

A large scale of domestic labor flows has been an outstanding feature of the Chinese economy over the past decades. Table 1.7 presents the total number of inter- and intra-provincial migrants in China between 2010 and 2015 as well as their shares of total population[37]. In 2015, a Chinese prefecture with median level of domestic migration would see 92 thousand of its population lived in another prefecture within the same province and another 43.5 thousand lived in a different province five years ago. In terms of shares of total population, a Chinese city in 2015 would on average have 5% of its population moved from another prefecture during the past five years.

Another feature with the domestic migration patterns of China is its significant imbalance in migration flows. The middle panel of Table 1.7 reports the net province-to-province migration flows, which are aggregated from the province-to-prefecture migration flow information in the population census. Net migration flows here are defined as the absolute difference between gross migration inflows and outflows of a province during the past five years. In 2015, provinces in China on average faced a number of net migrants as large as 4% of their total population, and the share was 16% for the most mobile province. As for comparison, the U.S. concurrent share of net five-year inter-state migrants in population was 0.5% averaging across states[38]. This observation

[37] As the publicly available 2015 Population Micro Census data do not contain individual-level employment status, migrants defined here include both workers and non-workers. This project makes the (admittedly strong) assumption that workers have the same migration pattern as the population.

[38] Author's calculation based on the 2011-2015 county-to-county ACS migration flows data.

acts as a strong signal that the labor market in China is not at steady state where labor gross inflows and outflows balance each other, and hence motivates a model with dynamic labor markets.

Segmentation

Table 1.8: List of Provinces by Region

Region	*Provinces*
Northeast (36)	Heilongjiang (13), Jilin (9), Liaoning (14)
North Municipalities (2)	Beijing (1), Tianjin (1)
North Coast (28)	Hebei (11), Shandong (17)
Central Coast (25)	Jiangsu(13), Shanghai (1), Zhejiang (11)
South Coast (33)	Fujian (9), Guangdong (21), Hainan (3)
Central (84)	Shanxi (11), Henan (18), Anhui (16), Hubei (14), Hunan (14), Jiangxi (11)
Northwest (64)	Neimenggu (12), Shannxi (10), Gansu (14), Qianghai (8), Ningxia (5), Xinjiang (15)
Southwest (61)	Sichuan (21), Chongqing (1), Yunnan (16) Guizhou (9), Guangxi (14)

Notes: Numbers in parentheses indicate the amount of prefectures within the corresponding geographic division.

Table 1.9: List of Industries by Sector

Sector	*Industry*
Primary	Farming, Forestry, Animal Husbandry and Fishery
Secondary	Mining and Quarrying
	Manufacturing
	Production and Supply of Electricity Gas and Water
	Construction
Tertiary	Geological Prospecting and Water Conservancy
	Transport, Storage, Post & Telecommunication Services
	Wholesale and Retail Trade & Catering Services
	Finance and Insurance
	Real Estate
	Social Services
	Health Care, Sports & Social Welfare
	Education, Culture and Arts, Radio, Film and Television
	Scientific Research and Polytechnic Services
	Government Agencies, Party Agencies and Social Organizations
	Other Services

Notes: This sector division is based on the 2005 Population Micro Census.

Table 1.10: List of Port Prefectures by Province

Province	*Prefecture*
Tianjin (1)	Tianjin
Hebei (3)	Tangshan, Qinhuangdao, Cangzhou
Liaoning (4)	Dalian, Dandong, Yingkou, Huludao
Shanghai (1)	Shanghai
Jiangsu (4)	Suzhou, Nantong, Lianyungang, Yancheng
Zhejiang (5)	Hangzhou, Ningbo, Wenzhou, Zhoushan, Taizhou
Fujian (3)	Fuzhou, Xiamen, Quanzhou
Shandong (4)	Qingdao, Yantai, Weihai, Rizhao
Guangdong (8)	Guangzhou, Shenzhen, Zhuhai, Shantou, Jiangmen, Zhanjiang, Maoming, Zhongshan
Guangxi (3)	Beihai, Fangchenggang, Qinzhou
Hainan (3)	Haikou, Sanya, Shengzhixia

Notes: Numbers in parentheses indicate the amount of port prefectures within the corresponding province.

Construction of Least Cost Path Spanning Trees

This section describes construction of the least cost path and Euclidean spanning tree hypothetical HSR networks depicted in Figure 1.2b. Here I just sketch the key steps given the procedure follows Faber (2014) closely. I define 34 nodal cities with the 30 provincial capital cities and 15 sub-provincial municipalities (11 of the 15 sub-provincial municipalities are also provincial capital cities). These politically important and economically prosperous cities are targeted by the Chinese planners when placing the HSR network ("The Medium- and Long-Term Railway Network Plan (2004)"). Both hypothetical HSR networks aim to answer the question of which HSR routes the Chinese central government would have built if their only objective had been to connect all targeted nodal cities while minimizing the total construction cost.

For the least cost path spanning tree network (Figure 1.2b) construction, I first define a linear construction cost function[39]:

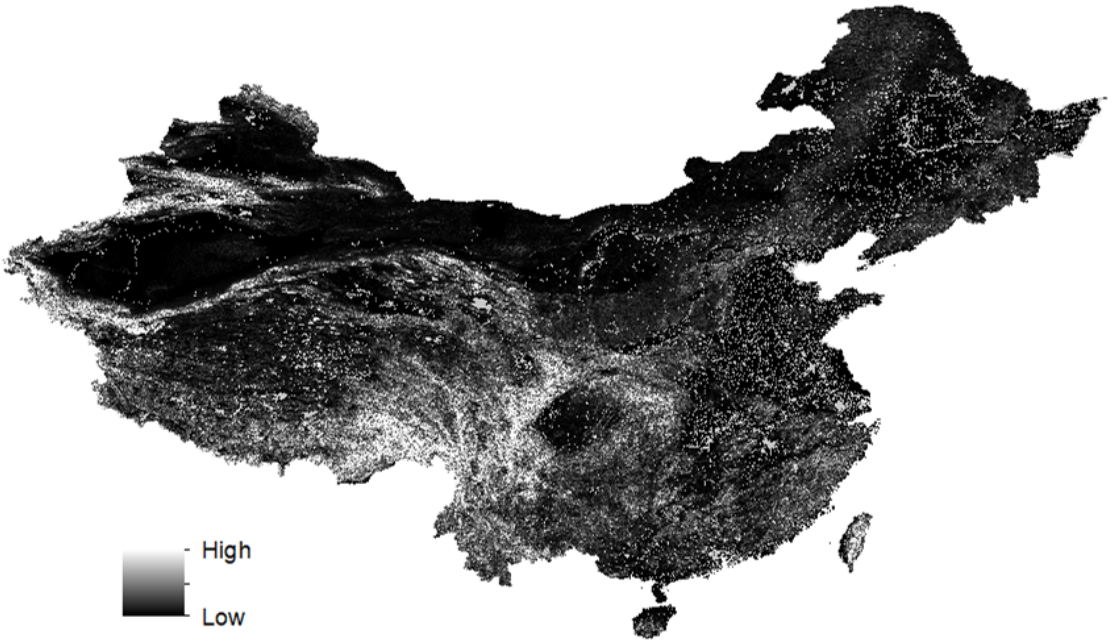
$$c_n = 1 + slope_n + 20 * Developed_n + 30 * Water_n + 30 * Wetland_n$$

where c_n is the incurred construction cost if HSR passes a pixel of land n, $slope_n$ is the average slope gradient of land n, $Developed_n, Water_n$ and $Wetland_n$ are dummy variables indicating whether n is covered by artificial structures, water body and wetland respectively. The cost specification implies that low construction costs are associated with shorter and flatter routes that avoid artificial structures, water bodies or wetlands. I use 1-km resolution raster data of the elevation and land use for mainland China in 2005 from the Resource and Environment Science Data Center, Chinese Academy of Sciences. Figure 1.6 depicts the resulting construction cost surface with darker regions indicating lower construction costs.

I then proceed to construct the least cost HSR paths between each of the 561 (34*33/2) possible bilateral pairs of targeted nodal cities. After extracting the 561 individual least cost HSR paths, the final step is to apply Kruskal's minimum spanning tree algorithm to find the single hypothetical HSR network that connects all nodal cities while minimizing aggregate construction costs.

[39]I adapt the simple linear cost function form from the highway construction literature (Faber 2014) and select the cost coefficients based on the World Bank report on HSR construction costs in China (Ollivier et al. 2014). As robustness checks, I choose alternative values (15, 25, 50) for the cost coefficients and the resulting HSR construction cost surfaces are very similar.

Figure 1.6: Construction Cost Raster



1.9.3 Robustness Checks for Reduced-form Analysis

Define top-50 cities based on 2010 GDP level

Table 1.11: Robustness Check I for Impact of HSR Construction on Prefectural Economic Development

	2010-2017				2000-2007			
Dependent variable	OLS	OLS	LCP IV	LCP IV	OLS	OLS	LCP IV	LCP IV
$\Delta ln(rGDP/worker)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ BigCityAccess	0.047***	0.045***	0.134**	0.106**	-0.005	-0.006	0.063	0.056
	(0.014)	(0.016)	(0.068)	(0.052)	(0.018)	(0.019)	(0.072)	(0.071)
lngdppw10	-0.194***	-0.198***	-0.220***	-0.216***	0.171***	0.170***	0.151***	0.152***
	(0.045)	(0.044)	(0.047)	(0.046)	(0.045)	(0.044)	(0.045)	(0.046)
Δ lnMP		4.586**		4.491**		1.151		1.054
		(2.221)		(1.869)		(1.420)		(1.277)
First stage F-Stat			16.90	14.17			16.90	14.17
Obs	333	333	333	333	333	333	333	333
R^2	0.87	0.88			0.94	0.94		

Notes: All regressions include province fixed effects. $\Delta BigCityAccess$ is the change in number of top 50 cities (rank based on 2010 GDP level) within 1-hour radius. LCP refers to the least cost path spanning tree network. lngdppw10 is log of GDP per worker in 2010. ΔlnMP is log change of market access between 2010 and 2017. Standard errors are clustered at the province level. ***1%, **5%, and *10% significance levels.

Count number of top-50 cities within 2-hour radius

Table 1.12: Robustness Check II for Impact of HSR Construction on Prefectural Economic Development

	2010-2017				2000-2007			
Dependent variable	OLS	OLS	LCP IV	LCP IV	OLS	OLS	LCP IV	LCP IV
$\Delta ln(rGDP/worker)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ BigCityAccess	0.011*	0.009	0.056*	0.047	-0.004	-0.004	-0.002	-0.004
	(0.005)	(0.005)	(0.034)	(0.032)	(0.007)	(0.006)	(0.021)	(0.021)
lngdppw10	-0.184***	-0.188***	-0.199***	-0.201***	0.171***	0.170***	0.170***	0.170***
	(0.045)	(0.045)	(0.049)	(0.050)	(0.046)	(0.045)	(0.043)	(0.043)
ΔlnMP		4.579*		4.237**		1.180		1.177
		(2.339)		(1.982)		(1.409)		(1.314)
First stage F-Stat			3.26	3.10			3.26	3.10
Obs	333	333	333	333	333	333	333	333
R^2	0.87	0.88			0.94	0.94		

Notes: All regressions include province fixed effects. $\Delta BigCityAccess$ is the change in number of top 50 cities (rank based on 2010 employment density) within 2-hour radius. LCP refers to the least cost path spanning tree network. lngdppw10 is log of GDP per worker in 2010. ΔlnMP is log change of market access between 2010 and 2017. Standard errors are clustered at the province level. ***1%, **5%, and *10% significance levels.

Regress GDP growth on full set of initial prefecture-level control variables

Table 1.13: Regressions of Income Growth on Initial Economic Conditions

Dependent variable	OLS	OLS
$\Delta ln(rGDP/worker)$	(1)	(2)
lngdp10	0.230	0.356
	(0.206)	(0.226)
lngdppw10	-0.448**	-0.576**
	(0.188)	(0.219)
lnpop10	-0.175	-0.307
	(0.172)	(0.196)
lnempden10	0.014	0.021
	(0.019)	(0.019)
lntotimpexp10	0.027**	0.028**
	0.011	(0.012)
lninvest10	0.062	0.067*
	(0.041)	(0.038)
lnindout10	-0.068**	-0.073***
	(0.027)	(0.025)
lngovtrev10	0.010	0.009
	(0.047)	(0.048)
agsh10	0.003	-0.032
	(0.192)	(0.186)
collegesh10	0.244	0.242
	(0.583)	(0.616)
lnMP10		4.805***
		(1.677)
Obs	333	333
R^2	0.89	0.90

Notes: Dependent variable is log change of real GDP per worker between 2010 and 2017. All explanatory variables take 2010 value. lngdp10 is log of GDP. lngdppw10 is log of GDP per worker. lnpop10 is log of population. lnempden10 is log of employment density. lntotimpexp10 is log of total import and export value. lninvest10 is log of investment in fixed assets. lnindout10 is log of industrial output. lngovtrev10 is log of government revenue. agsh10 is share of agriculture employment in total prefecture employment. collegesh10 is share of college-educated workers. lnMP10 is log of market potential. Both regressions include province fixed effects. Standard errors are clustered at the province level. ***1%, **5%, and *10% significance levels.

Add more prefecture-level control variables

Table 1.14: Robustness Check III for Impact of HSR Construction on Prefectural Economic Development

	2010-2017				2000-2007			
Dependent variable	OLS	OLS	LCP IV	LCP IV	OLS	OLS	LCP IV	LCP IV
$\Delta ln(rGDP/worker)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ BigCityAccess	0.037**	0.043***	0.103***	0.093	0.027	0.027	0.047	0.046
	(0.017)	(0.015)	(0.039)	(0.035)	(0.026)	(0.027)	(0.046)	(0.045)
lngdppw10	-0.288***	-0.291***	-0.298***	-0.299***	0.210***	0.209***	0.207***	0.207***
	(0.058)	(0.059)	(0.058)	(0.058)	(0.047)	(0.047)	(0.045)	(0.045)
lntotimpexp10	0.035***	0.038***	0.034***	0.037***	-0.006	-0.006	-0.007	-0.006
	(0.012)	(0.012)	(0.011)	(0.011)	(0.013)	(0.012)	(0.012)	(0.011)
lnindout10	0.003	0.001	-0.003	-0.003	-0.046**	-0.046**	-0.048**	-0.048**
	(0.024)	(0.022)	(0.022)	(0.019)	(0.022)	(0.022)	(0.021)	(0.021)
lnMP10	0.033	0.021	0.037*	0.023	0.103***	0.102***	0.105***	0.103***
	(0.023)	(0.022)	(0.022)	(0.020)	(0.022)	(0.023)	(0.021)	(0.022)
ΔlnMP		4.385**		4.646***		0.392		0.490
		(1.909)		(1.655)		(0.953)		(0.897)
First stage F-Stat			19.18	20.79			19.18	20.79
Obs	333	333	333	333	333	333	333	333
R^2	0.89	0.89			0.94	0.94		

Notes: All regressions include province fixed effects. $\Delta BigCityAccess$ is the change in number of top 50 cities (rank based on 2010 employment density) within 1-hour radius. LCP refers to the least cost path spanning tree network. lngdppw10 is log of GDP per worker in 2010. lntotimpexp10 is log of total import and export value in 2010. lnindout10 is log of industrial output in 2010. lnMP10 is log of market potential in 2010. ΔlnMP is log change of market access between 2010 and 2017. Standard errors are clustered at the province level. ***1%, **5%, and *10% significance levels.

1.9.4 Quantification Appendix

Detailed Solution Algorithm

The entire dynamic sequential competitive equilibrium is characterized by the following set of equations:

$$L_{i,t} = \sum_{n=1}^{N} \frac{V_{i,t}^{\beta\nu} \kappa_{ni}^{-\nu}}{\sum_{m=1}^{N} V_{m,t}^{\beta\nu} \kappa_{nm}^{-\nu}} L_{n,t-1} \tag{1.41}$$

$$V_{n,t}^{\nu} = (\frac{1+\eta_2/\eta_1-\chi_n}{\alpha})^{\nu} (w_{n,t}\xi_{n,t})^{\nu} [\sum_{i}^{N} V_{i,t+1}^{\nu} \kappa_{ni}^{-\nu}]^{\beta} \tag{1.42}$$

$$w_{n,t}L_{n,t} = \sum_{i=1}^{N+1} \frac{c_{n,t}^{-\theta}\tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} c_{m,t}^{-\theta}\tau_{mi}^{-\theta}} (1-\eta_1\chi_i) w_{i,t}L_{i,t} \tag{1.43}$$

$$c_{i,t} = \frac{w_{i,t}^{\eta_1} r_{i,t}^{\eta_2} P_{i,t}^{\eta_3}}{A_{i,t}} \tag{1.44}$$

$$A_{i,t} = \overline{A_i} [\sum_{s=1}^{N} e^{-\delta\iota_{si}} (\frac{L_{s,t}}{\overline{H_s}})]^{\rho} \tag{1.45}$$

$$\alpha r_{n,t}\overline{H_n} = (1-\alpha+\frac{\eta_2}{\eta_1}) w_{n,t}L_{n,t} - (1-\alpha)\chi_n w_{n,t}L_{n,t} \tag{1.46}$$

$$\xi_{n,t} = \frac{B_n}{P_{n,t}^{\alpha} r_{n,t}^{1-\alpha}} \tag{1.47}$$

$$P_{n,t} \propto [\sum_{i=1}^{N+1} (\tau_{in} w_{i,t}^{\eta_1} r_{i,t}^{\eta_2} P_{i,t}^{\eta_3})^{-\theta} A_{i,t}^{\theta}]^{-1/\theta} \tag{1.48}$$

Note that only the first two equations involve intertemporal changes in the state variables and hence are the key to capture the dynamic feature of the labor market. The remaining equations concern solely with the temporary equilibrium at each time period t.

Variables and parameters determined outside this solution algorithm are elasticities $(\nu, \theta, \delta, \rho)$, consumption share and input shares $(\alpha, \eta_1, \eta_2, \eta_3)$, discount factor (β), land area $(\overline{H_i})$, ratio of trade surplus to wage income (χ_i), bilateral migration frictions, trade frictions and travel time $(\kappa_{ni}, \tau_{ni}, \iota_{ni})$, as well as the initial distribution of state variables $(L_{i,0}, V_{i,0}, w_{i,0}, c_{i,0})$. Assuming these variables and parameters are known and the economy takes $T \geq 0$ periods to converge to the steady state, we can jointly solve the equilibrium paths for the endogenous variables of interest $(\{L_{i,t}\}_{t=1}^{T}, \{w_{i,t}\}_{t=1}^{T}, \{A_{i,t}\}_{t=1}^{T})$ as well as identify production and amenity fundamentals

$(\overline{A_i}, \overline{B_i})$ as following:

A. Production Fundamentals $\overline{A_i}$

Productivity of any region i at any time t, $A_{i,t}$, can be decomposed into a region-specific time-invariant part, $\overline{A_i}$, as well as a time-variant part that represents the production agglomeration force the region is experiencing in that period. $\overline{A_i}$ can be backed out by fitting the initial distribution of state variables into the algorithm below:

I. Plug $w_{i,0}$ and $L_{i,0}$ into Eq (1.46) to solve for the initial rental price $r_{i,0}$.

II. Use Eq (1.44) and (1.48) to solve for the initial price level $P_{i,0}$ up to a normalization. A normalization for P is needed since the equation set characterizing the sequential competitive equilibrium is homogeneous of degree 0 in P.

III. Use Eq (1.44) again to solve for initial regional productivities $A_{i,0}$.

IV. Plug $A_{i,0}$ into Eq (1.45) to get $\overline{A_i}$.

B. Steady State

The steady state is defined as when all the endogenous variables stop changing. That in steady state labor distribution stays unchanged simply means the labor inflow and outflow of each region exactly offset each other. It does not suggest that workers stop moving. In other words, in steady state the gross labor flows of each region could still be positive and even large, while the net labor flows of them are zero. A strategy to solve for the steady state takes the following steps:

I. Make a guess for region-specific time-invariant amenities $B_i^{(1)}$ and normalize. We are allowed to do the normalization as the dynamic labor market equation is homogeneous of degree 0 in B.

II. Make a guess for steady state wage distribution $w_i^{*(1)}$ and normalize[40]. The normalization is innocuous since the labor demand equation is homogeneous of degree 1 in w and the dynamic labor market equation is homogeneous of degree 0 in w.

III. Find the steady state labor distribution $L_i^{*(1)}$ that agrees with the guessed wage distribution.

i) Make a guess for steady state labor distribution $L_i^{*(1)}$. [41]

ii) Use Eq (1.45) and (1.46) to calculate the productivities $\widehat{A_i^*}$ and rental price $\widehat{r_i^*}$ that are consistent with the guessed wage and labor distribution.

iii) Plug $\{w_i^{*(1)}, \widehat{A_i^*}, \widehat{r_i^*}\}$ into Eq (1.48) and solve a nonlinear fixed point problem for the price level for final goods $\widehat{P_i^*}$ up to a normalization. [42]

[40]The "(1)" superscript means the variable is conditional on the first iteration of B, $B_i^{(1)}$.

[41]Note that total number of workers in China and in ROW should both be fixed: $\sum_{i=1}^{N} L_i^{*(1)} = \sum_{i=1}^{N} L_{i,0}$ and $L_{ROW}^{*(1)} = L_{ROW,0}^{*}$.

[42]Solution algorithm for the nonlinear fixed point problem here works as following: (1) Make a guess for P_i^* and normalize. (2) Compute the right-hand side of Eq (1.48) and normalize it as in Step (1). (3) Check if the result from

iv) Use Eq (1.47) to calculate the price-adjusted amenities $\widehat{\xi_i^*}$.

v) Plug $\{w_i^{*(1)}, \widehat{\xi_i^*}\}$ into Eq (1.42) and solve a nonlinear fixed point problem for the lifetime utility $\widehat{V_i^*}$ up to a normalization.[43]

vi) Use $\widehat{V_i^*}$ and compute the right-hand side of Eq (1.41) as $\widehat{L_i^{*(1)}}$. Check if $\widehat{L_i^{*(1)}}$ is close to the guess from Step i). If not, update the guess and repeat Step ii) to vi) until convergence.

IV. Denote the resulting steady state distributions of endogenous variables calculated from Step III as $\{L_i^*(w_i^{*(1)}), A_i^*(w_i^{*(1)}), r_i^*(w_i^{*(1)}), P_i^*(w_i^{*(1)})\}$. Use Eq (1.43) and (1.44) to calculate the wage needed to satisfy the labor demand condition[44]:

$$w_{n,update}^* = (\sum_{i=1}^{N+1} \frac{{c_n^*}^{-\theta}\tau_{ni}^{-\theta}}{\sum_{m=1}^{N+1} {c_m^*}^{-\theta}\tau_{mi}^{-\theta}} w_{i,guess}^* L_i^* - \chi_n w_{n,guess}^* L_n^*)/L_n^*$$

$$c_i^* = \frac{{w_{i,guess}^*}^{\eta_1} {r_i^*}^{\eta_2} {P_i^*}^{\eta_3}}{A_i^*}$$

V. Check if $w_{i,update}^{*(1)}$ is in equilibrium with the wage guess from Step II. If not, update the wage guess and repeat Step III and IV until convergence.

At the end of this section, we have solved the steady state equilibrium given the guess for region-specific time-invariant amenities $B_i^{(1)}$. Denote the steady state variable values as $\{w_{i,ss}^{(1)}, L_{i,ss}^{(1)}, V_{i,ss}^{(1)}\}$. Next section we work towards solving the full sequential competitive equilibrium given the steady state economy and the amenity fundamentals.

C. Dynamic Transition

The model suggests that given $B_i^{(1)}$, the labor distribution should start from $L_{i,0}$ ($L_{i,t=0}^{(1)} = L_{i,0}$) and converge to $L_{i,ss}^{(1)}$ after a sufficiently long T periods ($L_{i,t=T}^{(1)} = L_{i,ss}^{(1)}$):

I. Make a guess for the full transition path of labor distribution $\{L_{i,t}^{(1)}\}_{t=1}^{T-1}$.

II. Use Eq (1.45) to solve for the transition path of regional aggregate productivities $\{A_{i,t}^{(1)}\}_{t=1}^{T-1}$ based on $\{L_{i,t}^{(1)}\}_{t=1}^{T-1}$.

III. Find the path of wage distribution $\{w_{i,t}^{(1)}\}_{t=1}^{T-1}$ that consists with the labor transition path.

i) For each period t, make a guess for the wage as $w_{i,t}^{(1)}$ and normalize.

Step (2) is in equilibrium with the guess from Step (1). If not, update the guess and repeat Step (2) and (3) until convergence.

[43] Solution algorithm for the nonlinear fixed point problem here works as following: (1) Make a guess for V_i^* and normalize. (2) Compute the right-hand side of Equation (1.42) as $\widehat{V_i^*}^{\nu}$, convert it back to $\widehat{V_i^*}$ and normalize as in Step (1). Note that in steady state $V_{i,t} = V_{i,t+1} = V_i^*$. (3) Check if the result from Step (2) is in equilibrium with the guess from Step (1). If not, update the guess and repeat Step (2) and (3) until convergence.

[44] Condition on $w_i^{*(1)}$ and the "(1)" superscripts are omitted in the following two expressions for clarity.

ii) Plug $w_{i,t}^{(1)}$ and $L_{i,t}^{(1)}$ into Equation (1.46) to get rental price $r_{i,t}^{(1)}$.

iii) Fit $r_{i,t}^{(1)}$ and $L_{i,t}^{(1)}$ into the nonlinear fixed point problem elaborated in B.III.iii) to calculate the distribution of final goods price $P_{i,t}^{(1)}$.

iv) Calculate the wage distribution implied by the labor demand condition based on Equation (1.43) and (1.44)[45]. Check if the result is close to the initial guess. If not, update the wage guess and repeat Step ii) through iv) until convergence. Denote the final result as $\{w_{i,t}^{(1)}\}_{t=1}^{T-1}$.

IV. Given $\{L_{i,t}^{(1)}\}_{t=1}^{T-1}$ and $\{w_{i,t}^{(1)}\}_{t=1}^{T-1}$, calculate price-adjusted regional amenities $\{\xi_{i,t}^{(1)}\}_{t=1}^{T-1}$ according to Equation (1.46)-(1.48)[46].

V. For each $0 < t \leq T-1$, use $w_{i,t}^{(1)}, \xi_{i,t}^{(1)}, V_{i,t+1}^{(1)}$ and Eq (1.42) to solve backwards for $V_{i,t}^{(1)}$[47].

VI. For each $0 < t \leq T-1$, use $V_{i,t+1}^{(1)}, L_{i,t+1}^{(1)}$ and Eq (1.41) to solve backwards for $L_{i,t}^{(1)}$. This delivers a new transition path for labor distribution. Compare the new path with the guess in Step I. Check if these two paths are close, if not, update the guess and repeat Step II through VI until convergence.

D. Amenity Fundamentals B_i

Step B and Step C equip us with an algorithm to solve for the full dynamic path of $\{L_{i,t}^{(1)}, w_{i,t}^{(1)}, V_{i,t}^{(1)}\}_{t=1}^{T}$ given a guess for amenity fundamentals $B_i^{(1)}$. This section identifies B_i by approaching the initial lifetime utilities suggested by the above dynamic transition path $V_{i,0}^{(1)}$ to the values calibrated to the observed data $V_{i,0}$.

Specifically, combining Eq (1.42) and (1.47) gives

$$B_i = \frac{\alpha}{1+\eta_2/\eta_1}\left(\frac{V_{n,0}^{\nu}}{[\sum_i^N V_{i,1}^{\nu}\kappa_{ni}^{-\nu}]^{\beta}}\right)^{1/\nu} \Big/ \frac{w_{i,0}}{P_{i,0}^{\alpha} r_{i,0}^{1-\alpha}} \tag{1.49}$$

All variables on the right-hand side of Eq (1.49) are from the data or calibrated directly to the data except $V_{i,1}$. The final piece of the solution algorithm works as following: obtain an updated value for B using

$$B_i^{(2)} = \frac{\alpha}{1+\eta_2/\eta_1}\left(\frac{V_{n,0}^{\nu}}{[\sum_i^N V_{i,1}^{(1)\nu}\kappa_{ni}^{-\nu}]^{\beta}}\right)^{1/\nu} \Big/ \frac{w_{i,0}}{P_{i,0}^{\alpha} r_{i,0}^{1-\alpha}}$$

where $V_{i,1}^{(1)}$ is computed in Step C based on $B_i^{(1)}$. Check if $B_i^{(2)}$ is close to the initial guess $B_i^{(1)}$. If not, update the guess for B and repeat Step B-C until convergence.

[45] As in Step B.IV.

[46] As in Step B.III.ii) through B.III.iv).

[47] Note that we start from $t = T-1$ where $V_{i,t+1}^{(1)} = V_{i,ss}^{(1)}$ and $L_{i,t+1}^{(1)} = L_{i,ss}^{(1)}$ are known from Step B.

Chapter 2

Does Comparative Advantage Predict Future Growth?

Co-authored with Dominick Bartelme

2.1 Introduction

In this paper, we quantitatively examine whether a country's disaggregated export-revealed comparative advantage structure can be used to predict its GDP growth. The idea that the sectoral composition of a country's export basket affects its income has always been at the core of international economics. As early as in the 19^{th} century, David Ricardo in his book, *On the Principles of Political Economy and Taxation*, articulated that to maximize income countries should engage in international trade by exporting according to their comparative advantage. During recent years along with numerous theoretical work trying to pin down the exact mechanisms at play, economists are also exploring various empirical measures of the comparative advantage structure and trying to find a robust way to apply it when forecasting income growth.

One of the most influential empirical attempts during the past two decades is Hausmann, Hwang and Rodrik (2007), which has been widely acknowledged and applied by both economists and policy makers due to its strong predictive power and straight-forward interpretation (e.g., Berg, Ostry, and Zettelmeyer 2012; Hallak and Schott 2011). Specifically, Hausmann and his coauthors construct an aggregate index, *EXPY*, measuring the productivity level associated with a country's export basket. They analyze the changes of *EXPY* over time for some of the largest exporting countries, discuss potential fundamental determinants of cross-country *EXPY* variation and finally predict economic growth with it. Though such an aggregate index is desirable for its easiness to implement, we notice that the functional form used to construct *EXPY* is arbitrary and

lacks a disciplined theoretical foundation. Also the aggregation of sectoral data into a single index inevitably loses a lot of information reflecting detailed comparative advantage structure.

We start our analysis by empirically replicating the work of Hausmann et al. (2007) and testing the predictive power of *EXPY* over GDP growth. We construct *EXPY* using trade data from UN COMTRADE and real income changes data from the Penn World Table 9.0. The final dataset contains trade and socioeconomic information for 127 countries in 268 sectors between 1965 and 2015. To evaluate the prediction performance, we follow a common practice in the forecasting literature (Hyndman and Athanasopoulos 2021) to separate the available data into two mutually exclusive sections, training and test data. We implement two separation methods: i) hold-out randomly 10% of all country-sector-period observations as the test data and ii) hold-out the last period as the test data. In both scenarios we use the training data only to fit the prediction models and the test data only to evaluate their prediction performance. We refer to the predictions given by the prediction model when fitting the training dataset as in-sample results and those given by the prediction model when fitting the test dataset as out-of-sample results. The out-of-sample results are generally the focus of a prediction task as they provide a more reliable indication of how well the model performs on new data.

After these data preparations, we run the same linear regressions as in Hausmann et al. (2007) but find that the predictive power of *EXPY* over GDP growth only exists in in-sample results. For out-of-sample GDP growth forecasts which is one of the key questions of the economic growth studies, the Diebold-Mariano test results suggest we cannot reject the null hypothesis that *EXPY* has no statistically significant predictive power after controlling for the initial GDP level. We propose two possible explanations for this observation. First, the observable export data may encompass more information than just the comparative advantage structure of the exporting country. For instance, they also speak to the demand changes of trade partners that are external to the home country as pointed out by Bartelme, Lan and Levchenko (2021)[1]. To filter the comparative advantage structure out from the export data, we control for cross-country variations in external demand by using their external firm market access (FMA).

Second, an aggregate index as *EXPY* may miss certain fundamental information contained only in the sector-level export data. We need to directly investigate the linkage between the disaggregated sector-level export-revealed comparative advantage structure and GDP growth. A primary challenge with using data at the sectoral level is that the high dimensionality (hundreds of explanatory variables) often leads to nonnegligible econometric issues such as multicollinearity to common regression methods including Ordinary Least Squares (OLS). To deal with this problem, we employ machine learning techniques—Random Forests in particular. Random Forests is considered

[1]We do not aim to theoretically decompose export structure in this paper. Readers interested in this subject may refer to Bartelme, Lan and Levchenko (2021) for detailed discussions.

as one of the most powerful and interpretable machine learning methods. It has been applied and tested across numerous natural and social science fields (e.g. ecology: Prasad, Iverson and Liaw 2006, chemistry: Svetnik et al. 2003; political science: McAlexander and Mentch 2020). In economics, the popularity of Random Forests is also rising and economists are actively examining their econometric properties—e.g. Athey and Wager (2018) have proven the asymptotic normality for random forests under some mild conditions.

Equipped with both classic econometric regression model (OLS) and machine learning regression model (Random Forests), we continue our empirical analysis by comparing the out-of-sample prediction performance of *EXPY*, FMA-adjusted *EXPY* as well as the sector-level FMA-adjusted export basket that represents the disaggregated comparative advantage structure. We find that the disaggregated sector-level comparative advantage structure outperforms the others in both five-year and ten-year panels, as well as for both hold-out randomly (cross validation) and hold-out last period (one-period ahead forecast) scenarios. Its outperformance over *EXPY* stands even after we control for the initial GDP level which is a strong predictor of GDP growth by itself. As for robustness checks, we add in human capital index and region dummies as additional control variables. These two macro variables are widely-acknowledged GDP growth predictors (Sala-i-Martin et al. 2004). The predictive power of the disaggregated comparative advantage structure becomes not as strong as before, though remains in some prediction scenarios especially the one-period ahead forecasts. We conclude that the sector-level export-revealed comparative advantage structure serves as a better GDP growth predictor than *EXPY*, although its predictive power is still limited when controlling for a few standard macro variables.

In terms of literature contribution, this paper belongs to the large literature studying the linkage between trade patterns and income growth, e.g., Matsuyama (1992), Grossman and Helpman (1993), Hidalgo et al. (2007), Hausmann and Hidalgo (2011), Jarreau and Poncet (2012), and Hausmann et al. (2014). The most closely related one is a recent working paper by Bartelme, Lan and Levchenko (2021), which examines the impacts of foreign sectoral demand and supply shocks on income growth. Our paper differs from theirs in that we focus on applying machine learning techniques to nonparametrically uncover the predictive power of sector-level comparative advantage structure rather than an econometric estimation of shock elasticities.

This paper also contributes to the growing literature in applied macro and international economics that employs machine learning techniques to forecast macro variables. For example, Tkacz (2001) and Nakamura (2005) are early attempts to apply neural networks to forecast Canadian GDP growth and inflation respectively. Cook and Smalter Hall (2017) use deep learning to forecast unemployment. Sermpinis et al. (2014) use Support Vector Machine regressions to forecast inflation and unemployment. Döpke et al. (2015) and Ng (2014) aim to use random forests and boosting techniques to predict recessions. Machine learning has been used extensively in the

fields of statistics and computer science, but far less in economics. Athey (2018), Mullainathan and Spiess (2017) and Coulombe et al. (2020) provide excellent discussions about the potential contributions of machine learning to empirical economic research.

The rest of this paper is structured as following. Section 2.2 briefly reviews Hausmann et al.'s construction of *EXPY*. Section 2.3.1 replicates Hausmann et al.'s empirics with our own data and discusses the in-sample significance of *EXPY*. Section 2.3.2 tests the predictive power of *EXPY* for out-of-sample GDP growth forecasts. Section 2.4 more generally investigates the predictive power of export-revealed comparative advantage structure. Specifically, in Section 2.4.1, we modify *EXPY* by controlling for countries' external firm market access and examine the performance of this modified *EXPY* in out-of-sample prediction tasks. In Section 2.4.2, we focus on the disaggregated sector-level comparative advantage structure and evaluate its predictive power with machine learning techniques. Section 2.5 runs robustness checks by controlling for a few additional standard macro variables such as human capital and region dummies. Section 2.6 concludes.

2.2 *EXPY*

Hausmann, Hwang and Rodrik (2007), henceforth HHR, examine the proposition that the export basket of a country may have important implications for its economic growth. Empirically, they construct an index of the "income level of a country's exports" or *EXPY* and show that this aggregate index is capable of capturing the predictive power of export-revealed comparative advantage structure for GDP growth.

HHR first define an index $PRODY$ for each product k that represents the income level associated with it. Formally, $PRODY$ is constructed with the weighted average of the per capita GDPs of countries exporting that product:

$$PRODY_k = \sum_j \frac{(x_{jk}/X_j)}{\sum_{j'}(x_{j'k}/X_{j'})} Y_j \tag{2.1}$$

x_{jk}, X_j and Y_j denote country j's export of product k, total export and per capita GDP respectively. The weights are normalized share of product k in country j's export basket. From there, HHR define a country i's *EXPY* index as the weighted average of *PRODY* corresponding to its export structure:

$$EXPY_i = \sum_k \left(\frac{x_{ik}}{X_i}\right) PRODY_k \tag{2.2}$$

They then show the explanation power of $EXPY$ on GDP growth using various regression specifications in the empirical section and argue that $EXPY$ serves as a robust predictor for GDP growth.

2.3 Replication and Extension of HHR

2.3.1 In-sample Performance

We replicate HHR's empirics with our own data to validate the in-sample predictive power of *EXPY* over income growth. We use the UN Comtrade Database for the sector-level trade data and pick the time range to be between 1965 and 2015. To mitigate the influence of outliers, we use the 3-year average trade volume for every 5 years and concord 786 4-digit Standard International Trade Classification (SITC) products to sectors. We end up with a sample of 127 countries and 268 sectors. Data for real GDP per capita and human capital come from Penn World Table version 9.0.

Table 2.1 presents the results of our in-sample replication exercise. We adopt the same panel data regression specifications as in HHR: regressing average per-capita GDP growth for the next 5 and 10 years on *EXPY* while controlling for initial per-capita GDP level and human capital. The first and third columns are the OLS regression results using our data, while the second and fourth columns are the original OLS regression results taken from HHR[2]. All regressions include period fixed effects. In their paper, HHR also perform IV, two-way fixed effects (with both period and country dummies) and GMM regressions to show the robustness of the statistical significance of *EXPY*. Given our OLS results are sufficient to validate the in-sample statistical significance of *EXPY*, we omit the results with the other regression methods[3].

Our estimates of the coefficients on *EXPY* are statistically significant with a positive sign in both 5-year and 10-year panels, confirming HHR's argument regarding the in-sample predictive power of *EXPY* over income growth. These estimates suggest that a 10% increase in *EXPY* raises per capita GDP by 0.12-0.13 percentage points. Scales of our estimates are smaller, but comparable to those of HHR's. The other explanatory variables, initial per capita GDP level and human capital, are also found to have strong predictive power on GDP growth as expected.

2.3.2 Out-of-sample Performance

This section tests the predictive power of *EXPY* over GDP growth in out-of-sample prediction tasks. Out-of-sample evaluations are necessary as in-sample predictions are often subject to overfitting—corresponding too closely to the particular choice of training data and hence failing to fit additional data or predict future observations reliably. We take the initial GDP level as a baseline predictor and compare the prediction performance of using it alone versus using it together with *EXPY*. If *EXPY* has a significant out-of-sample predictive power as claimed by HHR,

[2]Column (1) and Column (5) from Table 9 in Hausmann et al. (2007).

[3]Compared to OLS, IV and GMM regressions give larger coefficients for $EXPY$ while the statistical significance levels are about the same.

Table 2.1: Panel growth regressions, OLS

	5-year panel		10-year panel	
	Our data	HHR 2007	Our data	HHR 2007
log initial GDP/cap	-.012	-.012	-.014	-.013
	(10.25)**	(4.39)**	(9.96)**	(4.42)**
log initial *EXPY*	**.012**	**.029**	**.013**	**.029**
	(4.94)**	**(5.38)****	**(4.51)****	**(5.22)****
log human capital	.037	.007	.042	.008
	(10.48)**	(3.27)**	(9.82)**	(3.75)**
Constant	.002	-.115	.000	-.108
	(2.29)**	(4.08)**	(0.18)	(3.68)**
Observations	1065	604	523	299

Note: Robust t-statistics in parentheses. *Significant at 10% level. **Significant at 5% level. All equations include period dummies.

we should see that adding *EXPY* as an additional predictor improves the predictive accuracy.

To prepare the out-of-sample data exercises, we divide our data into a training dataset and a test dataset. We use only the training dataset to run the regressions (or fit machine learning models in later sections) and use only the test dataset to evaluate the out-of-sample prediction performance. Such a division generally mitigates the overfitting concern as the test dataset is untouched during the fitting process (Hyndman and Athanasopoulos 2021). We do the division in two ways. One is to randomly select 90% of all the country-sector-period observations as the training dataset and keep the remaining as the test dataset. To deal with the issue that a specific choice of training data set may generate biased predictions, we repeat this selection 100 times and report the average values. This kind of resampling procedure is commonly known as cross validation in statistics and computer science. The other division way is to simply hold out the last period—similar to the standard one-period ahead forecasting task[4].

Another empirical challenge is that we need a systematic and disciplined way to compare the predictive accuracy given by two different sets of forecasts. A common approach is to select the forecast that has the smaller error measurement based on a loss function such as the Root Mean Square Error (RMSE). However, this type of approach does not speak to whether the difference between the two sets of forecasts is statistically significant, or it is simply due to a specific choice of data values in the sample. To address this issue, we propose to complement the RMSE comparison with the Harvey-Leybourne-Newbold Diebold-Mariano (HLN-DM) statistical test. This test is a finite-sample modified version given by Harvey, Leybourne and Newbold (1998) on the general Diebold and Mariano (1995) test. Specifically, the HLN-DM test works as following:

[4]We demeaned every variable by year to account for the period fixed effects.

Define the forecast errors as

$$e_{it} = \hat{y}_{it} - y_t \tag{2.3}$$

where $\hat{y}_{it}, i = 1, 2$ are the two forecast sets and y_t is the true value. Both $\hat{y}_{it}$ and y_t are vectors. Let $g(.)$ be the loss function and denote the loss differential between these two forecast sets by

$$d_t = g(e_{1t}) - g(e_{2t}) \tag{2.4}$$

The null hypothesis is that the two forecast sets have equal accuracy. With the observation that $\hat{y}_{1t}$ and $\hat{y}_{2t}$ have equal accuracy if and only if the expectation of the loss differential is zero, we can write the null hypothesis formally as

$$H_o : E(d_t) = 0, \forall t \tag{2.5}$$

Diebold and Mariano (1995) then derive a test statistic accordingly. For one-period ahead forecasts as in our case, the DM test statistic could be written as

$$DM = \bar{d} / \sqrt{\frac{\sum_{t=1}^{T} (d_t - \bar{d})^2}{T^2}} \tag{2.6}$$

where $\bar{d}$ is the sample mean of the loss differential and T is the number of observations in the test dataset. Noticing that the DM test seems to reject too often when having a finite small sample, Harvey, Leybourne, and Newbold (1997) improve on it by developing a bias correction to the DM test statistic and comparing the resulting statistic with the Student-t distribution instead of the standard normal. Again when dealing with one-period ahead forecasts as in our case, the HLN-DM test statistic is simply $\sqrt{\frac{T-1}{T}} DM$.

Table 2.2 summarizes our out-of-sample prediction results. We again conduct the exercise for both 5-year and 10-year panels. The first two rows show the out-of-sample root mean square errors (RMSE) associated with these two sets of predictors. The last row presents the HLN-DM test statistics which follow the Student-t distribution asymptotically. For hold-out randomly cases, reported values are the mean and standard errors of the RMSEs across 100 simulations.

In exercises (i), (iii) and (iv), hold-out randomly of the 5-year panel, hold-out randomly or last period of the 10-year panel, predicting with initial GDP and *EXPY* generates slightly smaller RMSEs than predicting with initial GDP only. However, the HLN-DM test results suggest that these two forecast sets are not statistically significantly different from each other. This finding can also be confirmed by noticing that the differences in mean RMSEs (< 0.01) are smaller than their respective standard errors (0.02-0.03). While for case (ii), hold-out last period of the 5-year panel, with a higher RMSE and HLN-DM test stat larger than 1.645, we are 90% confident that predicting

Table 2.2: Out-of-sample Performance

	5-year panel		10-year panel	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
Initial GDP				
RMSE	0.1609	0.1220	0.2527	0.2176
	(0.02)		(0.03)	
Initial GDP+*EXPY*				
RMSE	0.1597	0.1257	0.2497	0.2163
	(0.02)		(0.03)	
HLN-DM	0.60	1.95*	0.78	0.34

Note: Hold-out randomly cases: reported values are the mean and standard errors across 100 simulations. HLN-DM test stats: * significant at 10% level. **significant at 5% level.

with initial GDP and *EXPY* together is less accurate than predicting with initial GDP only.

In conclusion, adding HHR's *EXPY* as an additional predictor to the initial GDP level does not statistically significantly improve the forecast over GDP growth in any of the four data exercises. We cannot validate the out-of-sample predictive power of *EXPY* on GDP growth with our data.

2.4 Export-revealed CA Structure

In this section, we more generally investigate the relationship between export-revealed comparative advantage structure and GDP growth. Specifically, we aim to understand the failure of *EXPY* in out-of-sample prediction tasks by examining the following two hypotheses. First, the observable export data contain more information than just the comparative advantage structure of the home country, and that other information is diluting the predictive power of comparative advantage structure. Second, to build an aggregate index as *EXPY* we have to enforce a functional form when bringing the sector-level data together. It might be that the sector-level information lost during the aggregation procedure have a strong linkage with GDP growth.

2.4.1 Adjusted *EXPY*

To address the first hypothesis, we start with an adjustment on *PRODY*—an index quantifying the productivity level of each exporting product and serving as the building block of *EXPY*. HHR define *PRODY* of a product k as the weighted average of per capita GDPs of countries exporting k. They set the weights as the normalized export shares of product k in each country's export

basket. Or formally,

$$PRODY_k \equiv \sum_j \frac{\left(\frac{X_{jk}}{\sum_{k'} X_{jk'}}\right)}{\sum_{j'}\left(\frac{X_{j'k}}{\sum_{k'} X_{j'k'}}\right)} Y_j \tag{2.7}$$

However, the export share of product k in the home country's export basket also speaks to the demand changes of its trade partners that are external to the home country[5]. Thus to reveal the comparative advantage structure of the home country from its observable export data, we need to control for cross-country variations in external demand. Empirically we propose to use the country-sector specific firm market access (FMA) to capture the external foreign demand shocks. Our formal definition of the adjusted *PRODY* for product k is

$$Adj\ PRODY_k \equiv \sum_j \frac{\left(\frac{X_{jk}/FMA_{jk}}{\sum_{k'} X_{jk'}/FMA_{jk'}}\right)}{\sum_{j'}\left(\frac{X_{j'k}/FMA_{j'k}}{\sum_{k'} X_{j'k'}/FMA_{j'k'}}\right)} Y_j \tag{2.8}$$

where X_{jk} and Y_j represent country j's export of product k and per capita GDP level respectively.

We follow an identical way to HHR when constructing the adjusted *EXPY* from the adjusted *PRODY*. Substituting the *PRODY* term in HHR's *EXPY* formulation with our adjusted *PRODY* gives

$$Adj\ EXPY_i \equiv \sum_k \left(\frac{X_{ik}}{\sum_i X_{ik}}\right) Adj\ PRODY_k \tag{2.9}$$

Adj EXPY is our preferred aggregate index representing a country's comparative advantage structure. In the empirical section we will use both *EXPY* and *Adj EXPY* and test to see their prediction performance.

2.4.2 Sector-level CA Structure

Controlling for FMA targets the first hypothesis. To test the second hypothesis, we relax the parametric functional form needed for constructing an aggregate index as *EXPY*. In other words, we predict GDP growth directly with the disaggregated, sectoral, and FMA-adjusted export shares. We name such a country-sector specific index of comparative advantage structure as *ICA*

$$ICA_{ik} \equiv \frac{X_{ik}/FMA_{ik}}{\sum_{k'} X_{ik'}/FMA_{ik'}} \tag{2.10}$$

An empirical challenge with using sector-level *ICA* directly as explanatory variables is that its high dimensionality (268 sectors) brings a number of econometric issues such as multicollinearity

[5]See Appendix for a detailed discussion as well as a toy model.

to common regression methods including Ordinary Least Squares (OLS). To alleviate this concern, we employ nonparametic machine learning techniques—Random Forests in particular[6].

Random Forests algorithm builds upon the idea of decision tree (Quinlan 1993) which recursively partitions the training dataset into small boxes and then makes the prediction for a new observation based on the box it falls into. Observing that the decision trees are often affected heavily by the specific choice of training dataset (overfitting), Breiman (1996) proposes a bagging algorithm that obtains bootstrap samples by sampling with replacement from the training dataset and performs a decision tree analysis on each sample. Nevertheless, bagging does not entirely resolve the overfitting issue as there might be strong predictors that are shared by all trees in a bagging algorithm. These strong predictors will make all the decision trees look very similar and hence increase the correlation among them. Noticing this problem, Breiman (1999) further improves bagging with the Random Forests algorithm, which is essentially a bagging but considers only a random subset of explanatory variables instead of all variables at each split[7]. Since then, Random Forests has become one of the most commonly used machine learning algorithms known for its simplicity and generality (Athey and Wager 2018).

2.4.3 Prediction Results

With *Adj EXPY* and *ICA*, we are finally ready to perform the out-of-sample prediction exercises same as in Section 2.3.2 and test our two hypotheses. We again add in predicting with initial GDP level only as the baseline and predicting with HHR's original definition of *EXPY* for comparison. Table 2.3 and Table 2.4 present the out-of-sample root mean square errors (RMSE) and HLN-DM test stats respectively. For each predictor set, we run both OLS and Random Forest regressions. We carry the forecasts given by the regression method with smaller RMSE forward to the HLN-DM tests.

We first evaluate the out-of-sample predictive performance of *Adj EXPY*. Compared to the baseline, adding *Adj EXPY* as an additional predictor generates smaller RMSE in both hold-out randomly cases, (i) and (iii), as well as in the hold-out last period for the 10-year panel case, (iv). It generates slightly larger RMSE than the baseline in the hold-out last period for the 5-year panel case, (ii). Nevertheless, the differences between these two predictor sets are not significant in any of the four cases, as the HLN-DM test stats are all below 1.645. We are not able to conclude that *Adj EXPY* has statistically significant out-of-sample predictive power on GDP growth.

To check whether controlling for external foreign demand shocks helps with the predictive performance, we compare the out-of-sample predictive power of *Adj EXPY* and HHR's original

[6]We present results using other machine learning techniques such as Support Vector Machine and Gaussian Process Regression in the appendix.

[7]See Appendix for a detailed introduction to Random Forests.

definition of *EXPY*, labelled (2) and (3) in the result tables. As shown in Table 2.3, *Adj EXPY* generates smaller out-of-sample RMSEs than *EXPY* in the hold-out last period cases (ii) and (iv) but generates larger RMSEs than *EXPY* in the hold-out randomly cases (i) and (iii). Nevertheless, the HLN-DM test stats, 0.20, 0.56, 0.29 and 1.16 for the four cases respectively, again indicate we cannot reject the null hypothesis that these two predictor sets have same predictive accuracy in any case. The *Adj EXPY* does not outperform HHR's *EXPY* significantly. Our first hypothesis that the information besides the comparative advantage structure hidden in the observable export data is impeding the predictive performance is not sufficient[8].

To examine the second hypothesis that disaggregated sector-level comparative advantage structure is better at predicting GDP growth than an aggregate index, we turn our attention to the prediction results given by *ICA*, labelled (4) in Table 2.3 and Table 2.4. With about 270 explanatory variables, OLS becomes not capable of providing accurate predictions as can be seen from its high RMSEs. The usage of machine learning techniques such as Random Forests is necessary.

ICA generates the smallest out-of-sample RMSEs compared with the other three predictor sets across all four cases. If focusing on the RMSE results given by Random Forests only, the superior predictive performance of *ICA* is even more outstanding—suggesting its predictive power is mainly driven by the disaggregated sector-level explanatory variable rather than a specific choice of the regression method[9]. Together with the HLN-DM test stats between (1) the baseline and (4) *ICA*, we are 95% confident that *ICA* has significant out-of-sample predictive power for GDP growth in all cases except the hold-out last period in the 5-year panel. Based on the HLN-DM test stats between predictor sets (2) and (4), we are also 90% confident that *ICA* significantly outperforms HHR's original definition of *EXPY* except in the hold-out randomly case of the 5-year panel.

In conclusion, the disaggregated sector-level *ICA* has significant out-of-sample predictive power on GDP growth. Our second hypothesis that an aggregate index misses some fundamental linkage between the sectoral comparative advantage structure and GDP growth holds. In the next section, we conduct robustness checks to further examine whether the predictive power of *ICA* is strong and universal.

[8]It is worth pointing out *Adj EXPY*'s lack of predictive power does not entirely nullify our first hypothesis. One explanation could be controlling for foreign demand shocks is not enough to filter the information contained in the export data that is orthogonal to the CA structure. We leave a further discussion of this issue to future researchers.

[9]See Appendix for RMSE results using several other machine learning techniques. *ICA* generally outperforms the other predictor sets.

Table 2.3: RMSE Report, Controlling for Initial GDP

	5-year panel		10-year panel	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
(1) Initial GDP				
OLS	0.1609†	0.1220†	0.2527†	0.2176†
	(0.02)		(0.03)	
RF	0.1678	0.1445	0.2568	0.2529
	(0.02)		(0.03)	
(2) Initial GDP+*EXPY*				
OLS	0.1597	0.1257†	0.2497	0.2163†
	(0.02)		(0.03)	
RF	0.1570†	0.1339	0.2397†	0.2313
	(0.02)		(0.03)	
(3) Initial GDP+*Adj EXPY*				
OLS	0.1589†	0.1248†	0.2486	0.2122†
	(0.02)		(0.03)	
RF	0.1597	0.1360	0.2426†	0.2299
	(0.02)		(0.03)	
(4) Initial GDP+*ICA*				
OLS	0.2181	0.2125	0.3462	0.3589
	(0.03)		(0.06)	
RF	0.1496†	0.1191†	0.2188†	0.2028†
	(0.02)		(0.03)	

Note: All regressions include period dummies. Hold-out randomly cases: reported values are the mean and standard errors (in parentheses) across 100 simulations. † indicates the smaller RMSE given by the two regression methods.

Table 2.4: HLN-DM Test Stats, Controlling for Initial GDP

	Hold-out Randomly			Hold-out Last Period		
	(2)	(3)	(4)	(2)	(3)	(4)
5-year panel						
(1) Initial GDP	0.49	0.69	2.03**	1.95*	1.30	1.00
(2) Initial GDP+*EXPY*		0.20	1.23		0.56	1.91*
(3) Initial GDP+*Adj_EXPY*			1.50			1.51
(4) Initial GDP+*ICA*						
10-year panel						
(1) Initial GDP	0.97	0.73	3.03**	0.34	1.21	2.19**
(2) Initial GDP+*EXPY*		0.29	1.81*		1.16	1.87*
(3) Initial GDP+*Adj_EXPY*			2.02**			1.29
(4) Initial GDP+*ICA*						

Note: Hold-out randomly cases: reported values are the mean across 100 simulations. HLN-DM test stats: * significant at 10% level. **significant at 5% level.

2.5 Robustness Checks

We test the out-of-sample predictive power of HHR's original definition of *EXPY*, *Adj EXPY* and *ICA* while controlling for a few additional standard macro variables. Besides the initial GDP level as done in previous sections, we first add in human capital and then change to region dummies. Human capital is the same control variable used in Hausmann et al. (2007). Region dummies, including East Asia, Latin America, Africa and Europe, are widely used predictors that have high correlation with GDP growth (Sala-i-Martin et al. 2004). The argument is that if a predictor still holds significant out-of-sample predictive power after controlling for these variables, then there must be some unique linkage shared by that predictor and GDP growth while not captured by the other standard macro predictors. We would be confident in claiming that predictor helps with the predictive accuracy if people add it into their GDP growth predictor set.

Top half of Table 2.5 and Table 2.6 respectively present the out-of-sample RMSE and HLN-DM test results when we control for initial GDP level and human capital[10]. Baseline here is predicting with initial GDP and human capital. Both *EXPY* and *Adj EXPY* fail to add statistically significant out-of-sample predictive power in any of the four cases—reinforcing our findings in Section 2.4. Moreover, *EXPY* and *Adj EXPY* are significantly worse than the baseline in the hold-out last period with 5-year panel case, as shown by their higher RMSEs and the HLN-DM test stats larger than

[10]We calculate the Human Capital (HC) variable as log of the country-specific human capital data from Penn World Table 9.0.

1.645. For *ICA*, though generating the smallest RMSEs across all four cases, it only significantly improves the predictive power over baseline in the hold-out randomly with the 5-year panel case. In sum, this robustness check exercise again suggests that our second hypothesis is more likely to hold than our first hypothesis. However, we note that the out-of-sample predictive power of the disaggregated sector-level comparative advantage structure drops after adding human capital as another control variable.

Bottom half of Table 2.5 and Table 2.7 respectively present the out-of-sample RMSE and HLN-DM test results when we control for initial GDP level and region dummies. Baseline here is predicting with initial GDP and the four region dummies. *EXPY* fails to have significant out-of-sample predictive power in any of the four cases as before. *Adj EXPY*, on the other hand, has significant predictive power in the hold-out last period with the 10-year panel case. *ICA* significantly outperforms the baseline, *EXPY* as well as the *Adj EXPY* in both the hold-out last period cases, (i) and (iii), while not significantly in the two hold-out randomly cases. This robustness check exercise confirms that our second hypothesis is more probable than the first hypothesis. The predictive power of the disaggregated sector-level comparative advantage structure remains significant in out-of-sample prediction tasks, though less strong, after controlling for several additional standard macro variables.

Table 2.5: RMSE Reports for Robustness Checks

	5-year panel		10-year panel	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
(1)Initial GDP+HC				
OLS	0.1512(0.02)†	0.1167†	0.2264(0.02)†	0.1961†
RF	0.1561(0.02)	0.1266	0.2284(0.02)	0.2210
(2)Initial GDP+HC+*EXPY*				
OLS	0.1502(0.02)†	0.1203†	0.2229(0.02)	0.1947†
RF	0.1523(0.02)	0.1216	0.2194(0.02)†	0.2059
(3)Initial GDP+HC+*Adj_EXPY*				
OLS	0.1502(0.02)†	0.1204†	0.2240(0.02)	0.1948†
RF	0.1519(0.02)	0.1220	0.2205(0.02)†	0.2055
(4)Initial GDP+HC+*ICA*				
OLS	0.2158(0.04)	0.1970	0.3279(0.05)	0.3135
RF	0.1472(0.02)†	0.1133†	0.2087(0.02)†	0.1837†
(5)Initial GDP+Regions				
OLS	0.1517(0.02)	0.1268†	0.2292(0.03)	0.2170†
RF	0.1503(0.02)†	0.1290	0.2224(0.03)†	0.2243
(6)Initial GDP+Regions+*EXPY*				
OLS	0.1506(0.02)	0.1284	0.2262(0.03)	0.2120†
RF	0.1489(0.02)†	0.1273†	0.2189(0.03)†	0.2174
(7)Initial GDP+Regions+*Adj_EXPY*				
OLS	0.1507(0.02)	0.1270†	0.2272(0.03)	0.2120†
RF	0.1499(0.02)†	0.1297	0.2197(0.03)†	0.2202
(8)Initial GDP+Regions+*ICA*				
OLS	0.2149(0.04)	0.2102	0.3356(0.07)	0.3569
RF	0.1485(0.02)†	0.1192†	0.2134(0.03)†	0.1954†

Note: All regressions include period dummies. Hold-out randomly cases: reported values are the mean and standard errors (in parentheses) across 100 simulations. † indicates the smaller RMSE given by the two regression methods.

Table 2.6: HLN-DM Test Stats, Controlling for Initial GDP and Human Capital

	Hold-out Randomly			Hold-out Last Period		
	(2)	(3)	(4)	(2)	(3)	(4)
5-year panel						
(1) Initial GDP+HC	0.49	0.45	0.75	1.79*	2.18**	1.01
(2) Initial GDP+HC+*EXPY*		0.17	0.53		0.09	1.98**
(3) Initial GDP+HC+*Adj_EXPY*			0.61			2.01**
(4) Initial GDP+HC+*ICA*						
10-year panel						
(1) Initial GDP+HC	0.91	0.74	1.89*	0.35	0.36	1.49
(2) Initial GDP+HC+*EXPY*		0.25	1.09		0.05	1.21
(3) Initial GDP+HC+*Adj_EXPY*			1.12			1.26
(4) Initial GDP+HC+*ICA*						

Note: Hold-out randomly cases: reported values are the mean across 100 simulations. HLN-DM test stats: * significant at 10% level. **significant at 5% level.

Table 2.7: HLN-DM Test Stats, Controlling for Initial GDP and Region Dummies

	Hold-out Randomly			Hold-out Last Period		
	(6)	(7)	(8)	(6)	(7)	(8)
5-year panel						
(5) Initial GDP+Regions	0.85	0.33	0.50	0.92	0.18	2.17**
(6) Initial GDP+Regions+*EXPY*		0.41	0.19		1.06	2.37**
(7) Initial GDP+Regions+*Adj_EXPY*			0.39			2.08**
(8) Initial GDP+Regions+*ICA*						
10-year panel						
(5) Initial GDP+Regions	1.31	1.01	1.26	1.46	2.11**	3.47**
(6) Initial GDP+Regions+*EXPY*		0.34	0.79		0.01	2.36**
(7) Initial GDP+Regions+*Adj_EXPY*			0.88			2.56**
(8) Initial GDP+Regions+*ICA*						

Note: Hold-out randomly cases: reported values are the mean across 100 simulations. HLN-DM test stats: * significant at 10% level. **significant at 5% level.

2.6 Conclusion

In this paper we show that for out-of-sample GDP growth forecasts, an aggregate index summarizing the comparative advantage structure of a country such as Hausmann et al. (2007)'s *EXPY* does not have a statistically significant predictive power. We also demonstrate by using nonparametric machine learning techniques that the disaggregated sector-level comparative advantage structure better predicts GDP growth, though its predictive power is limited when controlling for a few additional standard macro variables. A more comprehensive understanding of the linkage between comparative advantage and GDP growth requires both a theoretical breakthrough, and a more robust empirical method to tell comparative advantage structure from the observable trade data.

Bibliography

[1] Athey, S. 2018. “The Impact of Machine Learning on Economics”. *The Economics of Artificial Intelligence: An Agenda* (forthcoming)

[2] Athey, S. and Wager, S. 2018. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. *Journal of the American Statistical Association*, 113(523), 1228-1242.

[3] Bartelme, D., Lan, T. and Levchenko, A., working paper. “Trade Specialization and Medium-Term Growth”.

[4] Berg, A., Ostry, J.D. and Zettelmeyer, J. 2012. “What makes growth sustained?” *Journal of Development Economics*, 98(2), 149-156.

[5] Breiman, L. 1996. “Bagging Predictors”, *Machine Learning*, 24, 123-140.

[6] Breiman, L. 2001. “Random Forests”, *Machine Learning*, 45(1), 5-32.

[7] Cook, T. and Smalter Hall, A. 2017. "Macroeconomic Indicator Forecasting with Deep Neural Networks".

[8] Coulombe, P.G., Leroux, M., Stevanovic, D. and Surprenant, S. 2020. "How is Machine Learning Useful for Macroeconomic Forecasting?"

[9] Diebold, F.X. and Mariano, R.S., 1995. “Comparing Predictive Accuracy”,*Journal of Business and Economic Statistics*, 13(3), 253-263.

[10] Diebold, F. 2015. “Comparing Predictive Accuracy, Twenty Years Later: a Personal Perspective on the Use and Abuse of Diebold-Mariano Tests”, *Journal of Business and Economic Statistics*, 33(1), 1.

[11] Döpke, J., Fritsche, U., and Pierdzioch, C. 2017. "Predicting recessions with boosted regression trees", *International Journal of Forecasting*, 33(4), 745–759.

[12] Grossman, G.M. and Helpman, E., 1993. "Endogenous Innovation in the Theory of Growth", *Journal of Economic Perspectives*, 8(1), 23-44.

[13] Hallak, J.C. and Schott, P.K. 2011. "Estimating Cross-Country Differences in Product Quality", *The Quarterly Journal of Economics*, 126(1), 417-474.

[14] Harvey, D.I., Leybourne, S.J. and Newbold, P. 1998. "Tests for Forecast Encompassing", *Journal of Business and Economic Statistics*, 16(2), 254-259.

[15] Hastie, T., Tibshirani, R. and Friedman, J.H. 2009. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer Science Business Media.

[16] Hausmann, R., Hwang, J. and Rodrik, D., 2007. "What You Export Matters", *Journal of Economic Growth*, 12, 1–25.

[17] Hausmann, R. and Hidalgo, C., 2011. "The Network Structure of Economic Output", *Journal of Economic Growth*, 16(4), 309-342

[18] Hyndman, R.J. and Athanasopoulos, G. 2021. "Forecasting: Principles and Practice (3rd ed)", *OTexts*.

[19] Hausmann, R., Hidalgo, C., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., and Yildirim, M.A., 2014. "The atlas of economic complexity: Mapping paths to prosperity" *MIT Press*.

[20] Hidalgo, C., Klinger, B., Barabasi, A., and Hausmann, R. 2007. "The Product Space Conditions the Development of Nations", *Science*, 371, 482–487.

[21] Jarreau, J. and Poncet,S., 2012. "Export sophistication and economic growth: Evidence from China", *Journal of Development Economics*, 97(2), 281-292.

[22] Matsuyama, K., 1992. "Agricultural Productivity, Comparative Advantage, and Economic Growth,", *Journal of Economic Theory*, 58(2), 317-334

[23] McAlexander, R.J., and Mentch, L. 2020. "Predictive inference with random forests: A new perspective on classical analyses", *Research and Politics*, 7(1), 205316802090548.

[24] Mullainathan, S., and Spiess, J.. 2017. "Machine Learning: An Applied Econometric Approach.", *Journal of Economic Perspectives*, 31(2), 87-106.

[25] Nakamura, E. 2005. "Inflation forecasting using a neural network", *Economics Letters*, 86(3), 373–378.

[26] Ng, S. 2014. "Viewpoint: Boosting recessions", *Canadian Journal of Economics*, 47(1), 1–34.

[27] Prasad, A.M., Iverson, L.R. and Liaw, A. 2006. "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction", *Ecosystems*, 9(2), 181-199.

[28] Quinlan, J.R. 1993. "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers Inc., San Francisco, California.

[29] Sala-I-Martin, X., Doppelhofer, G. and Miller, R.I. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach". *American Economic Review*, 94(4), 813-835

[30] Sermpinis, G., Stasinakis, C., Theofilatos, K., and Karathanasopoulos, A. 2014. "Inflation and unemployment forecasting with genetic support vector regression", *Journal of Forecasting*, 33(6), 471–487.

[31] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P. and Feuston, B.P. 2003. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling". *Journal of Chemical Information and Computer Sciences*, 43(6), 1947-1958.

[32] Tkacz, G. 2001. "Neural Network Forecasting of Canadian GDP Growth". *International Journal of Forecasting*, 17, 57-69.

2.7 Appendix

A. Toy Model

In this section we present a toy model to illustrate how we reveal the unobserved comparative advantage structure of a country from its observable sectoral export data. We show the theoretical foundation of the external firm market access (FMA), which we then empirically use as a controller for the cross-country variations in external demand. We suggest readers interested in a more thorough discussion of this topic to refer to Bartelme, Lan and Levchenko (2021).

Assume an Armington framework with N countries and K sectors. Labor is the only production factor. Workers are perfectly mobile across sectors but immobile across countries. Consumption preferences follow constant elasticity of substitution. Then as in the standard gravity formulation, we can write the total exports of the home country H in sector k, X_{Hk}, with a function of the unit production cost, c_{Hk}, and an external firm market access term, FMA_{Hk}:

$$X_{Hk} = \sum_{n \in N} p_{Hnk} q_{Hnk} = c_{Hk}^{1-\sigma} \sum_{n \in N} \tau_{Hnk}^{1-\sigma} \frac{E_{nk}}{P_{nk}^{1-\sigma}} = c_{Hk}^{1-\sigma} FMA_{Hk} \tag{2.11}$$

N includes all foreign export destinations. p_{Hnk} is the price a customer in country n pays for the good k shipped from country H. $\tau_{Hnk} > 1$ is iceberg trade cost. E_{nk} indicates country n's total expenditure on good k as final consumption or intermediate production materials. We use k to denote a sector or a good interchangeably[11]. $P_{nk}^{1-\sigma} = \sum_{j \in N \cup H} \left(c_{jk} \tau_{Hjk}\right)^{1-\sigma}$ is sectoral price index. $\sigma > 1$ is the Armington elasticity of substitution. As we can see from the derivation, FMA_{Hk} is linked to the demand changes of foreign trade partners but is exogenous to the home country.

We further express the unit production cost c_{Hk} as the ratio of the labor cost of the home country, w_H, to the productivity of the home country in sector k, T_{Hk}.

$$c_{Hk} = \frac{w_H}{T_{Hk}} \tag{2.12}$$

By productivity T_{Hk}, we mean how many units of good k a representative worker in country H can produce. Labor is perfectly mobile across sectors and hence the wage is constant across all sectors within a country. Given these assumptions, we can further express T_{Hk} with a function of country H's wage, sectoral export and FMA:

$$T_{Hk}^{\sigma-1} = w_H^{\sigma-1} \left(\frac{X_{Hk}}{FMA_{Hk}} \right) \tag{2.13}$$

[11] This is innocuous since by Armington assumption each sector has one unique good.

In our empirical section, we use a normalized country-sector specific productivity, named index of comparative advantage (ICA), to represent the sector-level export-revealed comparative advantage structure:

$$ICA_{Hk} = \frac{T_{Hk}^{\sigma-1}}{\sum_{k\prime} T_{Hk\prime}^{\sigma-1}} = \frac{X_{Hk}/FMA_{Hk}}{\sum_{k\prime} X_{Hk\prime}/FMA_{Hk\prime}} \tag{2.14}$$

To interpret the second equal sign, if the external demand conditions for a country remain constant (FMA stays fixed), an increase in the productivity of sector k relative to other sectors is expected to raise the share of sector k in the country's export basket. Empirically, we construct the *ICA* index for each country-sector-period pair.

B. Random Forest

Random forest builds upon the idea of "decision tree". A decision tree, as formalized by Quinlan (1993), is a tree-like flowchart consisting of a root (the starting point, which is usually the whole training dataset[12]), branches (splitting rules) and leaves (splitting outcomes). It starts from the root, asks a question about one independent variable to split the training dataset into two groups and then repeats this procedure for every nonterminal leaf (leaves that have not met the stopping rules).

A commonly used splitting rule is to go over each explanatory variable and pick the one that gives the smallest sum of squared prediction errors:

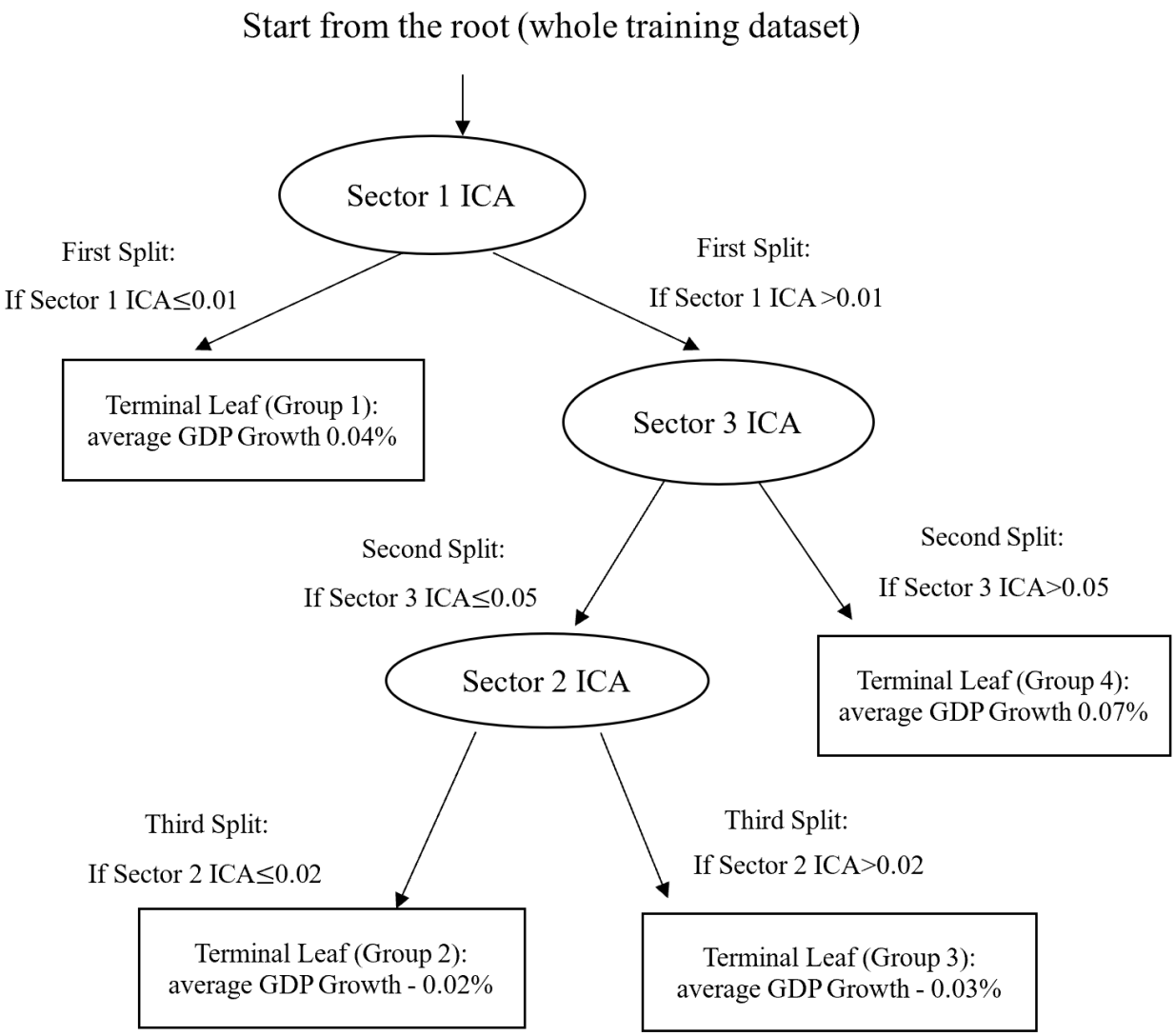
$$min\ SSE = \sum_{i \in S_1} (y_i - \bar{y_1})^2 + \sum_{i \in S_2} (y_i - \bar{y_2})^2 , \tag{2.15}$$

where i refers to the observation, y_i is the true outcome, S_1 and S_2 are the two observation groups split according to the splitting rules and $\bar{y_1}$ and $\bar{y_2}$ are the average outcomes of the observations in each group respectively. The stopping rules depend on the minimum leaf size (number of observations on a leaf) specified by the programmers. A split will not happen if either one of the two divided groups is left with fewer observations than the minimum leaf size benchmark after the split. Otherwise, the split process keeps running. In the end, a decision tree recursively partitions the entire training dataset into small leaves with pre-specified leaf sizes. It then makes the prediction for a new observation (from the test dataset) based on the leaf it falls onto—normally, the simple average of the outcomes of all other observations belonging to the same leaf.

Figure 2.1 illustrates the mechanism of a decision tree. After training, the decision tree splits the training dataset first according to Sector 1 *ICA* value, then according to Sector 3 *ICA* value and lastly according to Sector 2 *ICA* value. If comes in a new observation from the test dataset whose Sector 1 *ICA* is larger than 0.01, Sector 3 *ICA* smaller than 0.05 and Sector 2 *ICA* smaller than 0.02, then it will be placed onto the leaf associated with Group 2. If all other observations on this leaf have an average value of -0.02% for GDP growth, the outcome variable, then the decision tree will predict the new observation's GDP growth to be -0.02%.

[12]In applied statistics and computer sciences, a dataset is usually divided into a training dataset, which is used to fit the parameters of the models, and a test dataset, which is used to evaluate out-of-sample prediction performance. A typical exercise takes three steps. First, run regressions using only the training dataset while leaving the test dataset untouched. Second, plug the values of the explanatory variables from the test dataset into the fitted models from the first stage to make predictions. Third, compare the predicted values with the true values of dependent variable in the test dataset to assess the predictive accuracy.

Figure 2.1: Decision Tree Illustration



Decision trees often face the problem of overfitting (Hastie et al. 2009). Their prediction results tend to be affected heavily by the specific choice of training dataset. To deal with this problem, Breiman (1996) proposes a bagging algorithm that obtains bootstrap samples by sampling with replacement from the training dataset and performs a decision tree analysis on each sample. The final decision is made by the simple average of prediction results across these trees. Bagging outperforms a single decision tree since it results in a much lower variance while not affecting the bias (Breiman 1996).

Nevertheless, bagging does not entirely resolve the overfitting issue as there might be strong predictors that are shared by all trees in a bagging algorithm. These strong predictors will make all the decision trees look very similar and hence increase the correlation among them. Noticing this problem, Breiman (1999) further improves bagging with the Random Forest algorithm, which is essentially a bagging but considers only a random subset of explanatory variables instead of all variables at each split. Since then Random forests have been considered one of the most competitive regression methods and applied and tested among various fields.

In terms of statistical properties, Athey and Wager (2018) validate the asymptotic normality for random forests under some mild conditions.

C. Results Using other Machine Learning Techniques

SVM(linear): Support Vector Machine (default)

SVM(Gaussian): Support Vector Machine with a Gaussian kernel function

GPR: Gaussian Process Regression

Table 2.8: RMSE Report, Controlling for Initial GDP

	5-year panel		10-year panel	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
(1) Initial GDP				
OLS	0.1609(0.02)	0.1220	0.2527(0.03)	0.2176
RF	0.1678(0.02)	0.1445	0.2568(0.03)	0.2529
SVM(linear)	0.1614(0.02)	0.1250	0.2533(0.03)	0.2293
SVM(Gaussian)	0.1585(0.02)	0.1294	0.2457(0.03)	0.2287
GPR	0.1564(0.02)	0.1275	0.2409(0.03)	0.2169
(2) Initial GDP+*EXPY*				
OLS	0.1597(0.02)	0.1257	0.2497(0.03)	0.2163
RF	0.1570(0.02)	0.1339	0.2397(0.03)	0.2313
SVM(linear)	0.1603(0.02)	0.1257	0.2505(0.03)	0.2230
SVM(Gaussian)	0.1560(0.02)	0.1307	0.2406(0.03)	0.2247
GPR	0.1555(0.02)	0.1317	0.2367(0.03)	0.2191
(3) Initial GDP+*Adj EXPY*				
OLS	0.1589(0.02)	0.1248	0.2486(0.03)	0.2122
RF	0.1597(0.02)	0.1360	0.2426(0.03)	0.2299
SVM(linear)	0.1594(0.02)	0.1260	0.2493(0.03)	0.2185
SVM(Gaussian)	0.1564(0.02)	0.1332	0.2407(0.03)	0.2264
GPR	0.1562(0.02)	0.1338	0.2387(0.03)	0.2193
(4) Initial GDP+*ICA*				
OLS	0.2181(0.03)	0.2125	0.3462(0.06)	0.3589
RF	0.1496(0.02)	0.1191	0.2188(0.03)	0.2028
SVM(linear)	0.1594(0.02)	0.1204	0.2465(0.03)	0.2210
SVM(Gaussian)	0.1581(0.02)	0.1228	0.2393(0.03)	0.2104
GPR	0.1617(0.02)	0.1234	0.2528(0.03)	0.2215

Note: All regressions include period dummies. Hold-out randomly cases: reported values are the mean and standard errors (in parentheses) across 100 simulations.

Table 2.9: RMSE Report, Controlling for Initial GDP and Human Capital

	5-year panel		10-year panel	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
(1)Initial GDP+HC				
OLS	0.1512(0.02)	0.1167	0.2264(0.02)	0.1961
RF	0.1561(0.02)	0.1266	0.2284(0.02)	0.2210
SVM(linear)	0.1512(0.02)	0.1157	0.2267(0.02)	0.1952
SVM(Gaussian)	0.1528(0.02)	0.1210	0.2283(0.02)	0.2069
GPR	0.1515(0.02)	0.1198	0.2227(0.02)	0.1947
(2)Initial GDP+HC+*EXPY*				
OLS	0.1502(0.02)	0.1203	0.2229(0.02)	0.1947
RF	0.1523(0.02)	0.1216	0.2194(0.02)	0.2059
SVM(linear)	0.1502(0.02)	0.1186	0.2234(0.02)	0.1911
SVM(Gaussian)	0.1522(0.02)	0.1230	0.2249(0.02)	0.2001
GPR	0.1514(0.02)	0.1235	0.2206(0.02)	0.1966
(3)Initial GDP+HC+*Adj_EXPY*				
OLS	0.1502(0.02)	0.1204	0.2240(0.02)	0.1948
RF	0.1519(0.02)	0.1220	0.2205(0.02)	0.2055
SVM(linear)	0.1503(0.02)	0.1178	0.2245(0.02)	0.1937
SVM(Gaussian)	0.1511(0.02)	0.1237	0.2247(0.02)	0.1983
GPR	0.1526(0.02)	0.1248	0.2227(0.02)	0.1954
(4)Initial GDP+HC+*ICA*				
OLS	0.2158(0.04)	0.1970	0.3279(0.05)	0.3135
RF	0.1472(0.02)	0.1133	0.2087(0.02)	0.1837
SVM(linear)	0.1534(0.02)	0.1152	0.2238(0.02)	0.1972
SVM(Gaussian)	0.1528(0.02)	0.1180	0.2239(0.02)	0.1932
GPR	0.1578(0.02)	0.1194	0.2237(0.02)	0.1998

Note: All regressions include period dummies. Hold-out randomly cases: reported values are the mean and standard errors (in parentheses) across 100 simulations.

Table 2.10: RMSE Report, Controlling for Initial GDP and Region Dummies

	5-year panels		10-year panels	
	Hold-out Randomly (i)	Hold-out Last Period (ii)	Hold-out Randomly (iii)	Hold-out Last Period (iv)
(1)Initial GDP+Regions				
OLS	0.1517(0.02)	0.1268	0.2292(0.03)	0.2170
RF	0.1503(0.02)	0.1290	0.2224(0.03)	0.2243
SVM(linear)	0.1519(0.02)	0.1271	0.2282(0.03)	0.2189
SVM(Gaussian)	0.1584(0.02)	0.1290	0.2354(0.03)	0.2204
GPR	0.1492(0.02)	0.1320	0.2170(0.03)	0.2112
(2)Initial GDP+Regions+*EXPY*				
OLS	0.1506(0.02)	0.1284	0.2262(0.03)	0.2120
RF	0.1489(0.02)	0.1273	0.2189(0.03)	0.2174
SVM(linear)	0.1507(0.02)	0.1281	0.2257(0.03)	0.2151
SVM(Gaussian)	0.1537(0.02)	0.1286	0.2196(0.03)	0.1982
GPR	0.1495(0.02)	0.1352	0.2149(0.03)	0.2046
(3)Initial GDP+Regions+*Adj_EXPY*				
OLS	0.1507(0.02)	0.1270	0.2272(0.03)	0.2120
RF	0.1499(0.02)	0.1297	0.2197(0.03)	0.2202
SVM(linear)	0.1510(0.02)	0.1265	0.2269(0.03)	0.2165
SVM(Gaussian)	0.1556(0.02)	0.1333	0.2224(0.03)	0.2213
GPR	0.1506(0.02)	0.1355	0.2175(0.03)	0.2115
(4)Initial GDP+Regions+*ICA*				
OLS	0.2149(0.04)	0.2102	0.3356(0.07)	0.3569
RF	0.1485(0.02)	0.1192	0.2134(0.03)	0.1954
SVM(linear)	0.1538(0.02)	0.1277	0.2278(0.03)	0.2165
SVM(Gaussian)	0.1527(0.02)	0.1241	0.2208(0.03)	0.1971
GPR	0.1548(0.02)	0.1294	0.2205(0.03)	0.2050

Note: All regressions include period dummies. Hold-out randomly cases: reported values are the mean and standard errors (in parentheses) across 100 simulations.