Spatial Econometric Analysis on Regional Economy and Human Capital Development in China

By

YANG Wenxuan

DISSERTATION

Submitted in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy in International Development

GRADUATE SCHOOL OF INTERNATIONAL DEVELOPMENT

NAGOYA UNIVERSITY

Approved by the Dissertation Committee:

Prof. OTSUBO Shigeru

Prof. UMEMURA Tetsuo

A/Prof. SOMEYA Masakazu

A/Prof. Christian OTCHIA

A/Prof. Carlos MENDEZ (Chairperson)

Approved by the Faculty Council: September 8, 2023.

Contents

List of Tables and Figures

Tables

Figures

Acknowledgments

I would like to begin by expressing my deepest gratitude to my parents and grandparents, whose unwavering support and encouragement have been the pillars of strength throughout my educational journey. Their sacrifices, love, and belief in my abilities have propelled me forward, and I am forever grateful for their presence in my life.

Next, I extend my sincere appreciation to Professor Someya Masakazu, my main-supervisor. Under his expert guidance, I embarked on the path of academic research with confidence and determination. Professor Someya's invaluable mentorship, insightful feedback, and unwavering support have shaped not only my research but also my growth as a scholar. His dedication to my success has been truly inspiring, and I am privileged to have had the opportunity to learn from him.

I would also like to express the gratitude to my sub-supervisors, Professor Christian Otchia and Professor Carlos Mendez. Their expertise in their respective fields, their commitment to academic excellence, and their willingness to share their knowledge and insights have been invaluable throughout my doctoral journey. Their guidance and encouragement have played a significant role in shaping my research direction and enhancing the quality of my work.

Furthermore, I would like to express my heartfelt gratitude to my research collaborators, Dr. Zhang Yifeng, Ms. Pi Xuedi, Mr. Kwadwo Tabi Amponsah, Mr. Samuel Otokunor Armah, Mr. Luo Qingfeng, as well as my doctoral colleagues from China, Ms. Zheng Hua and Ms. Chen Yilin. Their contributions, whether through data analysis, experimental assistance, or thought-provoking discussions, have enriched my research and expanded its scope. Additionally, I am grateful to all the seminar members who have provided valuable insights, constructive feedback, and a stimulating academic environment that has nurtured my intellectual growth.

Lastly, I express my appreciation to the School of International Development at Nagoya University for their generous financial support, which has allowed me to pursue my research goals and contribute to the academic community. The research grants and resources provided

by the institution have played a crucial role in facilitating the successful completion of my doctoral studies.

Once again, I extend my deepest gratitude to everyone mentioned above and to all those who have contributed to my academic journey. Your support, guidance, and friendship have been invaluable, and I am truly humbled by your presence in my life.

This work was financially supported by JST SPRING, Grant Number JPMJSP2125. The author (Initial) would like to take this opportunity to thank the “Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System.”

1 Introduction

According to the World Economic Situation and Prospects report compiled by the UN in 2020, the world's countries can be divided into developed economies, economies in transition, and developing economies. Only 36 countries or regions can be classified into developed economies, and they are mainly distributed in Europe, North America, and Australia. The countries that make up most of the world are developing countries or countries in transition that are approaching the developed world.

Developed countries usually have a high level of economic and social development and a high standard of living for their inhabitants. The general characteristics are a per capita Gross National Product (hereafter GNP), per capita income, the highest levels of industrialization in the world, a well-developed infrastructure, and a very high human development index.

All developing countries strive to develop their economies to become developed countries with a high per capita income. However, it is not easy for developing countries to become developed due to historical, political, and demographic factors. We can still find fast-growing economies such as the Four Asian Tigers (Korea, Taiwan, Hong Kong, and Singapore) from 1960 to 1990 and emerging economies (China, Brazil, India, etc.) after 2000. The development experiences of these countries have important implications for other developing countries.

This dissertation focuses on the issues of regional income inequality and imbalanced human capital development in China. Regional income inequality in China refers to the significant disparities in economic and income levels between different regions. Although China has achieved remarkable economic growth over the past few decades, there are still noticeable differences in development levels among regions, resulting in the problem of regional income inequality. On the one hand, the economic development in the eastern coastal

areas of China has been relatively rapid, with higher Gross Domestic Product (hereafter GDP) and per capita income levels. In contrast, China's central and western regions have lagged in economic development and have lower per capita income levels. This development imbalance has led to the widening wealth gap and migration issues, as many people choose to work in the eastern coastal areas or large cities, exacerbating regional disparities.

On the other hand, different regions' development levels also affect residents' living standards and the quality of public services. For example, public service facilities, such as healthcare and education, are generally more developed in the eastern coastal areas. At the same time, they are relatively underdeveloped in the western and central regions, resulting in regional disparities.

The problem of regional income inequality in China results from multiple factors. Firstly, differences in geography and natural conditions are among the leading causes of income inequality. Some regions have advantages in terms of climate, terrain, or resources, such as the eastern coastal areas and the mineral-rich areas of western China, which make it easier for these regions to develop their economies and attract investments, resulting in higher income levels. Secondly, regional economic policies are also one of the factors leading to income inequality. During the initial reform and opening-up policy stage, the central government invested more resources and provided policy support in the coastal economic zones to promote local development, which may have resulted in relative underdevelopment in other regions. Thirdly, differences in human capital are also a cause of income inequality, and regional income inequality may also result in regional human capital imbalance. Population in some areas receive better education and training, possess higher skills and knowledge, and therefore can engage in higher-value-added work, resulting in higher income levels. However, populations in

other regions may need more skills and knowledge and thus can only engage in low-skilled and low-income work.

In conclusion, regional income inequality in China results from multiple factors. Recently, the Chinese government has been committed to narrowing the regional income gap. It has implemented a series of policies to support the development of the central and western regions and enhance infrastructure construction. Whether these policies have been effective and China's regional economic disparities have been reduced in recent years is a question that deserves further research.

2 Related Literature

Schultz, who first introduced the concept of human capital, argued that investment in human capital is as much an explanation of economic growth as investment in physical capital (Schultz, 1961). Moreover, in the case of workers, the human capital investment could explain wage increases. This revolutionary theory emphasized the importance of people in productive activities and found an alternative source of growth for economic growth.

With the development of economics, the role of human capital for economic growth and social development has become a consensus in the academic community (Lucas 1988). The importance of human capital in reducing regional disparities and promoting balanced economic development has also received increasing attention (Acemoglu 2012). Conversely, the uneven distribution of human capital may also be a reason for the uneven regional development of the economy.

For China, the "strong east and weak west" development pattern has been a problem since the reform and opening up. To alleviate this problem, China has introduced a series of policies to balance regional development since 2000. The most representative of which is the Development

of China's Western Regions, which encourages university students to support the construction in the west and gives generous incentives. This policy has promoted the flow of talent and economic development in the western region, but it still has not changed its talent and economic development pattern.

2.1 Characteristics and Dynamics of Human Capital Distribution in China

Li et al. (2013) measured total and per capita human capital in six Chinese provinces based on the J-F lifetime income method and found that human capital and per capita income have a similar distribution pattern. The changes in human capital have structural characteristics, i.e., slow growth from 1985-1995 and rapid growth after 1995. In addition, the gap in human capital per capita between developed and underdeveloped provinces is widening.

Zhang and Huang (2020) and Peng (2019) use the Gini and Theil coefficients to measure per capita capital inequality in China and show that the human capital structure in China is characterized by distinct phases, with different trends in each decade. Spatially, spatial agglomeration is evident and has a polarization characteristic of "higher the higher and lower the lower." The study of Li and Chen (2019) also confirms this spatial unevenness and polarization.

Previous studies have shown that the distribution of capital per capita in China is characterized by a "high in the east and low in the west," which is similar to its per capita income, and that this imbalance is increasing, as evidenced by the fact that regions with high human capital stocks are growing faster and those with low human capital stocks are growing slower.

Most papers are using spatial modeling in measuring spillover effects from human capital. Using exploratory spatial data analysis (ESDA) in a spatial econometric approach, Lu and Zhou

(2014) found similar human capital and economic development characteristics in China and developed a spatial Lucas model. This study found significant positive spatial spillover effects of human capital in most Chinese provinces. The study by Fang and Luo (2016) reached similar conclusions and verified that the spillover effect of human capital has a growth effect on the economy using the GMM approach. Previous literature has verified that human capital has spillover effects, but scant research is discussing the role of spillover effects in human capital dynamics.

2.2 Spatial Spillover Effects of Human Capital

Using the exploratory spatial data analysis (ESDA) in a spatial econometric approach, Lu and Zhou (2014) found similar human capital and economic development characteristics in China and developed a spatial Lucas model. This study found significant positive spatial spillover effects of human capital in most Chinese provinces. The study by Fang and Luo (2016) reached similar conclusions and verified that the spillover effect of human capital has a positive growth effect on the economy using the GMM approach.

2.3 Markov Chains and Spatial Applications

This study mainly draws on Rey's (2001) Markov chain and spatially extended form in terms of methodology. Rey (2001) extends the Markov chain framework, integrates Markov chains and spatial correlation analysis, allowing Markov chains to incorporate with regional context. The empirical study of regional income from 1929-1994 concluded that geography impacts the evolution of regional income distribution. After this, many scholars have used Markov chains in the study of regional income changes (Le Gallo 2004) (Hammond 2004).

Human capital, as with many economic factors, is mobile and has spillover effects. So theoretically, the dynamics of human capital in a region can also be affected by its neighbors. Peng (2019) found that human capital inequality in China fluctuates every ten years, and there is a clear trend towards polarization of the spatial distribution. This implies that the role of space has an impact on human capital changes. In addition, Zhang and Huang (2020) as well as Lu and Zhou (2014) also confirmed human capital has spillover effects to neighboring regions.

However, these studies did not find whether spatial effects could impact human capital dynamics or not, due to the limitation of methodology. The innovation of this study includes the human capital index as a research variable within the research framework of Markov chains. It would be the first study to use Markov chains to study regional human capital changes. It is an attempt to move from static analysis to dynamic analysis in researching human capital issues.

2.4 Spatial Heterogeneity in China's Economy

Spatial autocorrelation and spatial heterogeneity are fundamental properties of geographic data and two critical concepts in spatial econometrics. Spatial autocorrelation (spatial autocorrelation) refers to the correlation between observations of some variables within the same distribution. Tobler (1970) states that all attribute values on a geographic surface are related to each other, but closer values are more strongly related than more distant ones. In general, we can measure spatial dependence through Moran's I or spatial modeling.

On the other hand, the spatial heterogeneity will be used to explain that the relationship between a set of variables will change in different regions. And this variation can generally be measured by a geographically weighted regression. A geographically weighted regression (GWR), as a spatial analysis technique in geography, is mainly applied in environmental and

resource-related topics. The literature using GWR models to study economic issues is relatively small, and exploring the factors influencing house prices is a hot topic in economics.

In the literature on China, some studies use the level of economic growth as the explanatory variable to investigate the effects of one or several explanatory variables on local economic growth in other regions. Su Fanglin examines the heterogeneous impacts of R&D using provincial and urban data, respectively. Su (2007) uses R&D intensity as the explanatory variable. Using data from 30 Chinese provinces from 1993-2002, Su (2007) conducted a regression and found that the estimated output elasticity of R&D has an apparent spatial difference, which decreases from the east to the west.

In another paper, Su (2010) regressed the number of researchers and the knowledge stock as the main explanatory variables for 202 Chinese cities from 1993 to 2002 and showed that research environment, research input, and knowledge stock all have a heterogeneous effect on the output. The research environment and the research input have the opposite effect. The elasticity of the output of knowledge stock is low in most cities of the three provinces in Northeast China, while it is high in the areas around Beijing, Gansu, and Sichuan. Similarly, Li et al. (2018) studied the effect of spatial knowledge spillover on regional economic growth, the authors combined a spatial Durbin model and a GWR model for the regression of provincial data and concluded that the contribution of spatial knowledge spillover to regional economic growth decreases from east to west, and the contribution gap narrows from 2010 to 2015.

Some other articles discuss the heterogeneous effects of multiple factors on regional economies. Mou (2010) combines per capita fixed asset investment, labor force, education expenditure, the proportion of output value of the primary industry, and the total amount of foreign capital utilized as explanatory variables and concludes that the influence of fixed asset investment per capita on different regions has the largest difference. The estimated parameter

values of Beijing and its surrounding areas are the highest. The parameter estimated by Bai and Zhang (2014) used urbanization level, the proportion of the secondary industry, the total power of agricultural machinery, per capita income of farmers, resource endowment index, as well as the economic location index, and concluded that the influence intensity of farmers' per capita net income, resource endowment, and total power of agricultural machinery on per capita GDP decreased successively from southeast to the influence of economic location and the proportion of secondary industry on per capita GDP falls from northwest to southeast. It shows that the economic growth of the Central Plains Economic Zone may also be affected by topography, climate, regional policies, etc.

3 Human Capital Dynamics across Provinces in China: A Spatial Markov Chain Approach

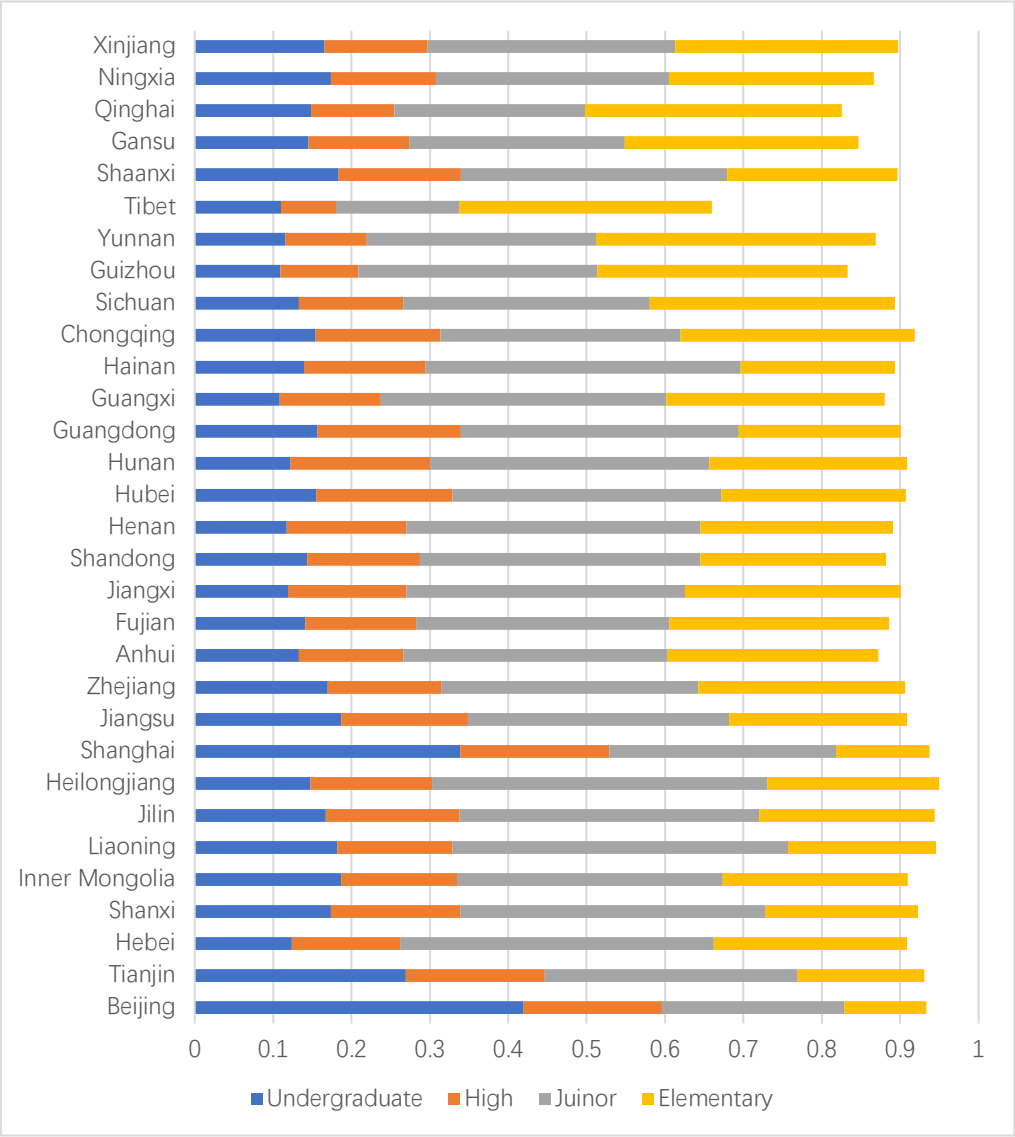
3.1 Introduction

In recent years, China has placed greater emphasis on talent. In March 2016, the Chinese Communist Party Central Committee issued the “Opinions on Deepening the Reform of the Institutional Mechanism for Talent Development”, proposing a more active, open, and effective talent introduction policy. Subsequently, local governments in China's second-tier cities introduced "New Talent Policies," which aimed at attracting college students and talents with certain professional knowledge or special skills to settle, including cash subsidies, employment settlements, and low-interest housing loans, in what the media called a "war for talent." By the end of April 2021, China had 3,191 items of local talent policies (Xia 2021).

On the contrary, the requirements for settling in big cities such as Beijing and Shanghai are getting higher and higher, and generally, only talents with master's degrees or above can meet the requirements for settling in these cities. Although these first-tier cities have introduced some

talent incentivizing policies, they are unattractive compared to the living and housing subsidies and fast-track settlement policies in second-tier cities. China's first-tier cities are rich in university resources and do not lack talent, so they do not need many welfare policies to attract talent to stay and work locally.

Figure 3.1 Regional Educational Attainment Percentage in 2020



Note: This percentage is for all personnel in different regions.

Source: National Bureau of Statistics of China.

The distribution of human capital and labor force in China is uneven. This distribution is influenced to some extent by the talent policy in recent years (Wang et al. 2021), and more importantly, it is related to the degree of economic development of a region. Figure 3.1 shows the proportion of the population with various levels of education in different regions. For human capital, we are more concerned with the proportion of the population with higher education. For most regions, the percentage of undergraduates is between 10% and 20%, but in the more developed provinces or municipalities, the number is much higher, such as Beijing (41.98%), Shanghai (33.87%), Tianjin (26.94%).

Large cities and developed regions can attract talents, while less developed regions will face the problem of brain drain. Nie and Liu (2018) found that talent flows come from provinces with many students and universities and tend to flow to regional economic centers or developed provinces, such as Beijing, Shanghai, and Guangdong Province. It may affect the dynamics of human capital and exacerbates the imbalanced distribution of human capital in China. Human capital is an essential factor for economic development, which can improve the efficiency of using physical capital. Chu and Cao (2019) as well as Verginer and Riccaboni (2020) have found that talent mobility will make talent gather in developed areas and lead to a shortage of talent in backward areas. It will intensify the imbalance of regional development.

Massive talent flow will cause changes in regional human capital, and these changes may be influenced by neighboring regions due to spillover effects. There is an emerging academic consensus that human capital has spillover effects like knowledge and technology (Fang & Luo 2019; Lu & Zhou 2014). With the help of spatial econometrics, we can capture the spillover effects of human capital. In general, the spillover effects of human capital are most pronounced for neighboring regions. It implies that the role of neighbors has an essential effect on human capital accumulation, which is also called the spatial effect.

Spatial effects should be considered when discussing the relationship between regional development and human capital accumulation. However, before Rey (2001), research using the Markov approach did not include the role of geography. Based on the definition of the Markov Chain, the probability of transferring from one state to another depends only on the state of the previous period. The Spatial Markov Chain developed by Rey (2001) integrated local spatial statistics into the Markov Chain framework. This is a new way to consider spatial effects in dynamics. In China's case, most papers are using spatial modelling in measuring spillover effects from human capital. However, spatial modelling cannot detect the role of spatial effects in human capital evolution. This paper aims to fill the research gap by combining spatial econometrics and a traditional Markov Chain methodology together and check whether spatial effects would affect human capital dynamics. The purpose of this paper is to investigate the characteristics of dynamic changes in regional human capital, to calculate the transition probabilities of each level of human capital by applying Markov chains, and to apply Markov chains in the spatially extended form to compare the differences in dynamic changes in spatial and non-spatial forms.

This paper shows that regional human capital imbalances are increasing in China and that spatial effects matter in this process. By comparing Markov chains in non-spatial and spatial forms we find that a region is more likely to improve if it is surrounded by neighbors with a high human capital level, and conversely, a region has a higher probability to relegate if it is surrounded by neighbors with less human capital.

This paper contributes to the literature of China's regional human capital development in three ways. First, an increasing number of articles use Markov chains to study changes in regional income (Bode & Nunnenkamp 2011; Rey et al. 2016; Kang & Rey 2018). In this paper, Markov chains are used in the human capital index to calculate the dynamic distribution and

transition probability of regional human capital in China. Second, a growing of literature finds that human capital has spatial effects (Lu & Zhou 2014; Xu & Li 2020). Based on the methodology of Rey (2001), this paper incorporates spatial factors into the Markov chain framework and verifies that spatial effects affect human capital dynamics. Lastly, most articles on spatial econometrics use spatial autocorrelation analysis to do static explorative analysis. This paper combines local spatial autocorrelation[1] with the Markov chain framework in a dynamic way.

3.2 Methodology and Data

3.2.1 Markov Chain

Traditional ESDA methods can only characterize the spatial features of a given year or compare the changes in spatial features between two years. Markov chains, on the other hand, can provide probabilistic information about dynamic changes. Specifically, Markov chains can discretize a random sequence of continuous states into several types in a specific application and calculate the probability distribution of each type and the general trend of its evolution to approximate the spatio-temporal characteristics of the variables (Tao & Qi, 2013).

According to (Rey 2001), in mathematical terms, assuming a total of k species and T times, the distribution of states at time t can be represented by a 1*k vector of

$$P_t = [P_{1,t}, P_{2,t} \dots P_{k,t}] \tag{3.1}$$

This vector represents the distribution probability of the different species at time t.

The probability of transition in each region can be represented as $m_{t,i,j}$, where t is the time, i is the initial state and j is the end state. $m_{t,i,j}$ represents the probability that at time t, this region is converting from state i to state j at the following period. In the basic Markov model,

we consider transition probability as time-invariant, which means $m_{t,i,j} = m_{t+b,i,j}\ \forall b$. Therefore, we can write weight matrix as

$$M = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix} \tag{3.2}$$

which must meet the condition $\sum_{j=1}^{n} m_{ij} = 1$.

Assume one simple scenario where there are only two states, Low and High, for a variable. These two states can switch to the other state or stay the same in the following period. Moreover, there are only two times: t_0 and t_1. Then, the transition probability matrix M for these two states can be represented in Table 1.

Table 3.1 Simplified Markov Transition Matrix

t_0 \ t_1	Low	High
Low	m_{LL}	m_{LH}
High	m_{HL}	m_{HH}

Notes: This table is drawn by author himself based on (Rey 2001).

With the help of the transition probability matrix M, we can compute the state distribution for period t + 1 as follows.

$$P_{t+1} = P_t M = P_{t-1} M^2 = \cdots = P_0 M^t \tag{3.3}$$

where P_0 is the initial state distribution. From this equation, we know the evolution of the Markov Chain is totally decided by a transition probability matrix.

Markov chains can also give us some other information, such as the time it takes to reach a steady state and it's time to transit from one state to another. The details will be shown in the empirical analysis.

3.2.2 Spatial Expansion of Markov Chain

There is an assumption in the traditional Markov chain, which is that the effect of spatial factors or spatial homogeneity is not considered. We expand the k*k matrix into the form k*k*k to include effects from neighbors. For simplicity, as in Table 3.2, we consider here a simple scenario. The variables have only Low and High states, and the neighbors also have only Low and High states. t_0 and t_1 denote the starting and ending time points, respectively.

Table 3.2 Simplified Spatial Markov Transition Matrix

Spatial Lag	t_0 \ t_1	Low	High
Low	Low	$m_{LL\|L}$	$m_{LH\|L}$
	High	$m_{HL\|L}$	$m_{HH\|L}$
High	Low	$m_{LL\|H}$	$m_{LH\|H}$
	High	$m_{HL\|H}$	$m_{HH\|H}$

Note: This table is drawn by author himself based on Rey (2001).

In Table 3.2, spatial lag indicates the state of neighbors, which is analogous to the presence of a precondition. For example, $m_{LH|L}$ denotes the transition probability of the region moving

from Low to High when the neighbor is Low. By comparing the spatial transition matrix with the traditional transition matrix, we can see if the spatial context has a significant effect on the transition probability. For example, $m_{LH|L} < m_{LH}$ suggests that having poorer neighbors is detrimental to a region's conversion from Low to High. Conversely, if the spatial context does not have a noticeable effect on the transition probability, then

$$m_{ij|1} = m_{ij|2} = \cdots = m_{ij|k} = m_{ij} \; \forall \; i, j. \tag{3.4}$$

Spatial Markov chains can provide information on whether transition across classes is related to neighbors and the magnitude of the influence from neighbors. But spatial Markov chains cannot measure the specific magnitude of the impact from neighbors.

3.2.3 Data

The data for this study come from the Human Capital Index Project of China Center Human capital and labor Market Research (CHLR). This project uses and improves upon the internationally widely used Jorgenson-Fraumeni income calculation method (hereafter referred to as the J-F method) to estimate regional human capital in China. Essentially, the J-F lifetime income approach measures the level of human capital as the present value of an individual's lifetime income over his or her expected lifetime. Assuming that an individual's human capital can be traded in the market like physical capital, the price is the present value of the individual's future lifetime earnings over his or her expected lifetime.

The J-F income-based approach is the most widely used measure of human capital today. It estimates expected future lifetime earnings based on the currently observed earnings of a cross-section of individuals. In this method, personal income is expected to grow at a certain growth rate, discounted to the present at a fixed discount rate. An individual's expected lifetime earnings are calculated from variables such as years of schooling, gender, age, observed average

earnings of the cohort, and induction rate. The total human capital stock of the region is obtained by multiplying the individual human capital of the region by the total number of individuals (Li et al. 2013). The lifetime income approach can accurately reflect the role of long-term investments such as education and health in human capital accumulation.

This paper is using the calculation results from Project 2021, which covers the period from 1985 to 2019. The unit of the human capital index is thousand yuan. This paper used real average human capital value for each province, which is adjusted by Consumer Price Index (CPI) and the base year is 1985. This database covers 31 provinces and municipalities in mainland China, and the Human Capital Index data from 1985-2019 are used in this paper.

Table 3.3 Descriptive Statistics of Regional Human Capital Index

	In 1985				In 2019			
	West	Middle	East	Northeast	West	Middle	East	Northeast
Mean	45.08	45.83	69.9	52.67	406.83	494.5	649.8	422
S.D.	5.68	4.17	18.94	2.31	105.68	92.80	211.00	71.97
Min	36	38	52	50	225	374	357	339
Median	44.5	47	62.5	54	434	501.5	636.5	460
Max	53	49	108	54	563	617	1082	467
Obs.	12	6	10	3	12	6	10	3

Note: China are divided into west, middle, east and northeast, according to the 2011 classification by the National Bureau of Statistics of China. S.D. = standard deviation, Min = minimum value, Max = maximum value, Obs. = observation.

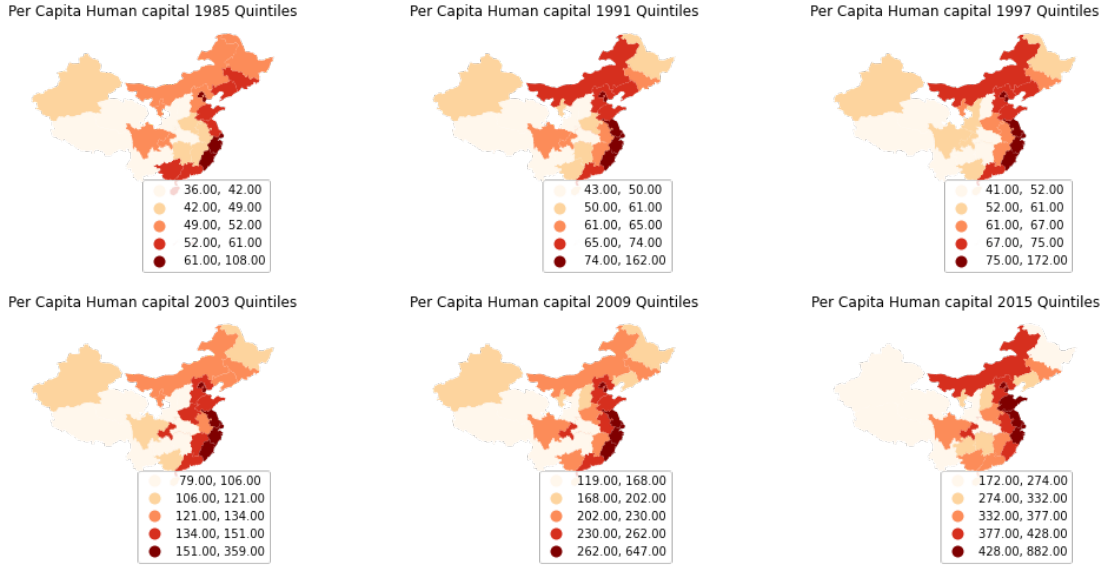
Source: Author's calculation using data from China Center for Human Capital and Labor Market Research (CHLR).

Table 3.3 compares the descriptive statistics of the human capital index for the provinces between the four geographical regions of China for the starting year (1985) and the final year (2019). Both the mean and median indicate that in both years, China's regional human capital is characterized by a high east and low west. However, the relative position of the Northeast has changed. In 1985, the human capital index of Northeast was between the East and Middle. But in 2019, the index for the Northeast is lower than that of Middle. This result implies China's regional human capital pattern is changing.

3.3 Empirical Analysis

3.3.1 The Dynamics of Human Capital

Figure 3.2 Change in Human Capital Index Distribution

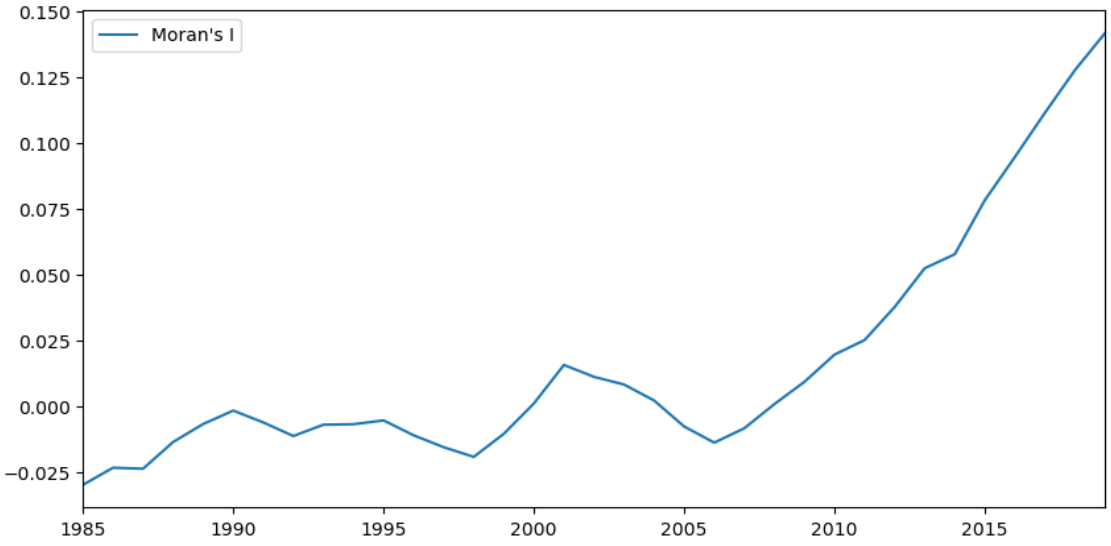


Source: Author's calculation using data from China Center for Human Capital and Labor. Market Research (CHLR).

Figure 3.2 shows the regional differences and evolution of human capital in China. Based on the quintiles of the human capital index, Chinese provinces are divided into five levels, with darker colors indicating higher human capital. Some features are concluded from these

evolution maps. First, the regions with higher human capital index are concentrated in the developed eastern coastal provinces, and the neighboring regions are also at a relatively high level. Second, regions in western China have lower human capital indexes, with Xinjiang province dropping to the lowest level in the latest graph. Third, China's northeastern region drops from the higher level to the lowest level. Although the human capital levels of Chinese provinces are in dynamic change, the pattern of higher in the east and lower in the west is unchanged. It implies that the gap between the eastern and western provinces does not decrease over time, and there is an increasing spatial clustering of human capital.

Figure 3.3 Yearly Moran's I of Human Capital Index[2]



Source: Author's calculation using data from China Center for Human Capital and Labor Market Research (CHLR).

From the maps in Figure 3.2, we find the sign of spatial clusters. Then we can use Moran's I as a measure to describes the global spatial correlation of variables. Figure 3.3 shows Moran's I of human capital from 1985 to 2019. We can see the spatial correlation in human

capital has experienced an increase from negative to positive during this period, although the change in values is small. This implies positive neighboring effects in human capital are increasing.[3] It also implies that we should consider spatial effects in researching on the changes of regional human capital.

3.3.2 Markov Chain

Based on the way of Rey (2001), we divided Chinese provinces into five quintile levels according to the human capital per capita index, namely, poor, lower, middle, upper and rich. The level of each region is changeable over time. Markov chains can record the frequency of these changes in each year and calculate the probability of moving from one state to another. This is an important message in a dynamic change. I only used the full sample in this research since it requires a large sample in Markov chain analysis. For example, Rey (2001) applied 3,120 samples in his research. My paper only has 1,085 samples, if I divide the full sample into several sub-groups, movement between different groups will be less captured.

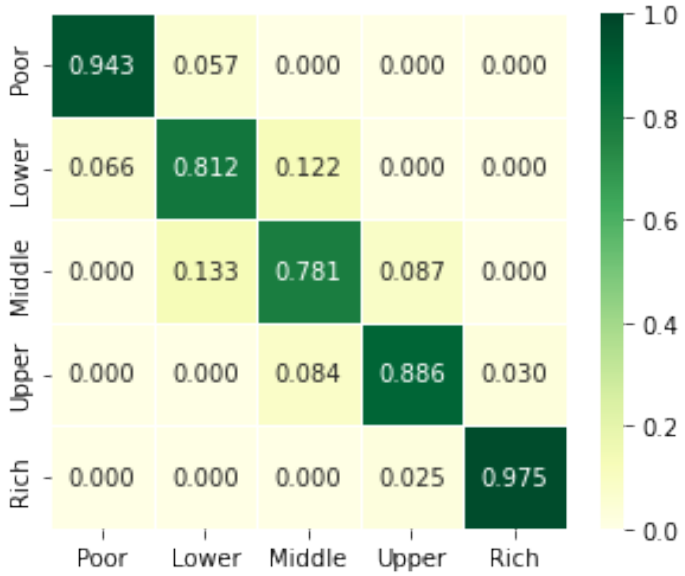
Table 3.4 Classic Markov Transition Frequency

	Poor	Lower	Middle	Upper	Rich
Poor	230	14	0	0	0
Lower	14	173	26	0	0
Middle	0	26	153	17	0
Upper	0	17	179	6	0
Rich	0	0	0	5	194

Note: This frequency records the transition of 31 regions in China from 1985-2019. The. transition probability of Figure 3.4 is calculated based on this frequency.

Source: Author's calculation using data from China Center for Human Capital and Labor. Market Research (CHLR).

Figure 3.4 Classic Markov Transition Matrix



Source: Author's calculation using data from China Center for Human Capital and Labor Market Research (CHLR).

Table 3.4 and Figure 3.4 record the transfer frequency and transfer probability of sample points among the five levels, respectively. The transfer probability is derived from the transfer frequency, so they convey the same information. For example, in Table 3.4, we focus regions on the Poor class in the first period. 230 sample points are still on the Poor class in the next period, and 14 sample points enter the Lower class. No points enter the higher levels of Middle, Upper and Rich. Combining the results with the other rows we can see that all transfers are made in the level adjacent to the starting point and no jumps across levels occur. The values on the diagonal are much larger than the other values, which indicates that most of the points stay in their original level, and even Middle, which is the least stable, has a 78.1% probability of staying in its original position in the next period. In summary, the transition matrix in Figure

3.4 implies that the richest and poorest areas (Rich class and Poor class in Figure 3.4) are very difficult to change, while Middle class has a relatively higher probability of making changes.

3.3.3 Spatial Markov Chain

Table 3.5 Spatial Dependance Test

Number of classes: 5		
Number of transitions: 1054		
Number of regimes: 5		
Test	Likelihood Ratio	χ^2
Stat.	47.599	46.883
DOF	29	29
p-value	0.037	0.043

Source: Author's calculation using data from China Center for Human Capital and. Labor Market Research (CHLR).

Likelihood Ratio test (hereafter referred to as the LR test) and the χ^2 test are used to test whether the regional context has a significant effect on the estimation results of the Markov Chain. The null hypothesis of this test is that the variables are spatially independent. In Table 3.5, the P-value shows that both the LR test and χ^2 test reject the null hypothesis, which indicates that the variables are not spatially independent. In other words, we cannot ignore the interaction of variables in space, which is why we use the spatial Markov Chain.

Figure 3.5 Comparison of Classic and Spatial Markov Transition Matrix

Source: Author's calculation using data from China Center for Human Capital and Labor Market Research (CHLR).

Figure 3.5 shows both the classical Markov probability matrix and the spatial Markov probability matrix. A classical Markov chain does not consider the influence of neighbors, while a spatial Markov chain considers this spatial influence from neighbors. In Figure 3.5, I divide the neighbors into 5 classes from small to large by quintiles. Levels 1-5 correspond to 'Poor', 'Lower', 'Middle', 'Upper', and 'Rich', indicating the stratum in which the neighbors' human capital levels are located.

Different neighbors have a significant effect on the movement of human capital levels across classes. If a region is surrounded by neighbors with rich human capital, it will have a higher probability of staying in the same class or moving to a higher class, while if it is surrounded by neighbors with less human capital, the region will have a higher probability of falling into a lower level. For example, the middlemost grid in each graph indicates the probability that a

'Middle' region will remain in 'Middle' in the next stratum. When we do not consider the effect of spatial factors, this probability is 0.781. when surrounded by poor areas, this probability decreases to 0.615. Conversely, if one region is surrounded by rich areas, the probability of remaining in this stratum is 0.875. In addition, 'Middle' regions with poor neighbors have a 0.308 probability of falling into the 'Lower' class, which is much higher than if space is not a factor (0.133) and if they have rich neighbors (0.021)

Cases in other levels also support the neighbor's effects. The neighbor's level will affect the transition probability of moving up and dropping down. It implies the role of spatial effects. According to Tobler's First Law of Geography, geographical things are correlated with each other in terms of spatial distribution. Therefore, the results of this study are consistent with Tobler's First Law.

3.4 Concluding Remarks

This article aims to explore the characteristics of regional human capital dynamics in China and compare the difference between considering and not considering spatial effects. To calculate the transition probabilities, this paper applies Markov chains and spatially extended Markov chains methods in regional human capital index. Results indicate that China's human capital development is uneven, with some wealthy coastal regions having much higher per capita human capital than others. Results imply that the dynamic change may be correlated to the increasing spatial effects. From 1985 to 2019, the degree of spatial autocorrelation increased. It turned from an insignificant negative correlation to a significant positive correlation.

Another important finding is that regional context matters in the transition process, which is similar with Wang (2013) but using the spatial Markov Chain method. The likelihood ratio test and the Chi-square test indicate that we cannot ignore spatial effects in regional human capital.

Comparing the Markov chain result and the spatial Markov chain result, we find how spatial effects influence the transition process. If one region is surrounded by rich regions, it is more likely to move up to the upper level. Conversely, regions surrounded by poor neighbors have a higher possibility of dropping down to a lower level.

Finally, further research about dynamics in China's human capital can be extended in at least three fronts. First, to confirm the role of neighbors in human capital, more robustness checks can be done by using other datasets. Second, alternative dynamic explorative analysis could be considered, like directional LISAs and a local indicator of mobility association (LIMA). These methods can give more information about the transition process. Finally, to measure the magnitude of the neighbors' effect, spatial modelling for human capital is needed. For instance, Lu and Zhou (2014) used a spatial-extended Lucas model to estimate the human capital spillover effect between provinces.

4 Re-estimate economic convergence in China: Using satellite nightlight data

4.1 Introduction

The core issue of economics is how to maintain economic growth and reasonable income distribution. However, in most developing countries, income inequality is an important issue that affects sustained economic development and social stability. In China, this problem is mainly reflected in regional differences.

The convergence hypothesis is a critical way to measure the variation of regional economic differences. β convergence[1] implies that regions with a disadvantaged background will grow faster, which means that the development gap tends to be close within one economy. After China established the policy of reform and opening up in 1978, some coastal cities were set up as special economic zones. Special economic zones are granted more free market-oriented economic policies and flexible governmental measures. Then they attracted more foreign investment and became the windows to doing foreign business. As a result, the coastal areas in eastern China developed rapidly, but the gap between east and west widened gradually. The goal of the Chinese Communist Party is common prosperity, but at the beginning of reform and opening up, "A part of people who enrich first can be the model, then arouse and help the rest." (citation?) Has China achieved common prosperity now? From the perspective of β convergence, China has gone through a process of "convergence-divergence-convergence" (Dong & Chi, 2020), which means regional economic disparities have gone through three stages of "narrowing-widening-narrowing" (Chen et al., 2018).

Most studies support the existence of absolute convergence in China (Hong et al., 2010; Chen et al., 2018; Si & Wang, 2021). In addition, recent studies have realized the role of spatial

[1] β convergence refers to a process in which poor regions grow faster than rich ones and therefore catch on them (Young et al., 2018).

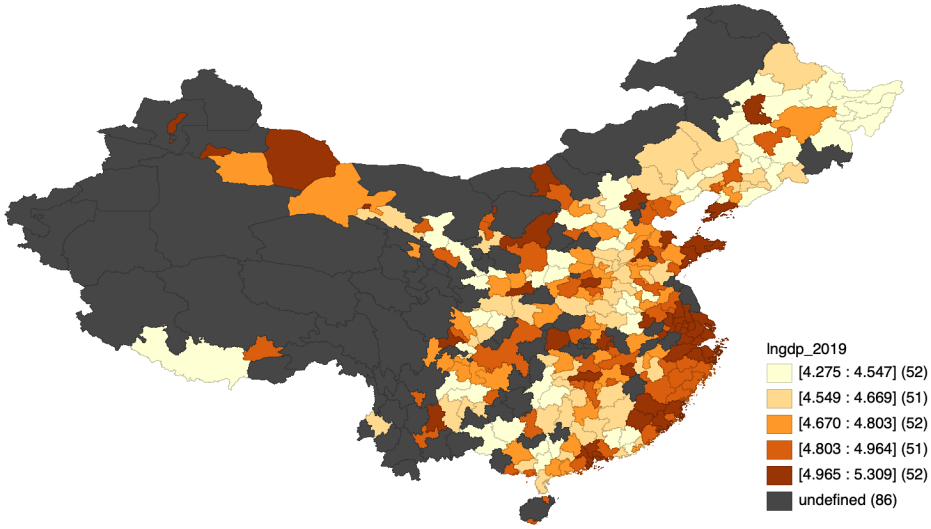
factors and expanded the spatial convergence model by incorporating spatial factors into the traditional convergence framework (Hong et al., 2010; Dong & Chi, 2020). However, these papers with spatial models emphasize the role of spatial correlation, but they ignore another vital feature in space-spatial heterogeneity.

This paper aims to investigate the regional differences in convergence rates in China. Firstly, this paper uses satellite lighting data to fill in the missing regional economic data. Then the regional convergence rates are calculated based on the city level. Last, this paper applies multi-scale geographically weighted regression in calculating the influence of impacting factors on the convergence process.

4.2 Data

4.2.1 Missing Value Issue in China

Figure 4.1 Regional GDP per capita in China in 2019



Note: The underdefined grey areas on the map are missing values; only mainland. city data is available.

Source: Author's calculation using data from *China City Statistical Yearbook 2020.*

Missing values at the city level is a severe problem in China. Basic economic data cannot be collected in some remote areas of China. For example, Figure 4.1 shows the regional GDP per capita of 2019 in China, with 86 of all 344 prefecture-level data missing. The missing value percentage is 25%. The proportion of missing values in earlier years before 2019 is even higher.

4.2.2 Application of Satellite Lighting Data

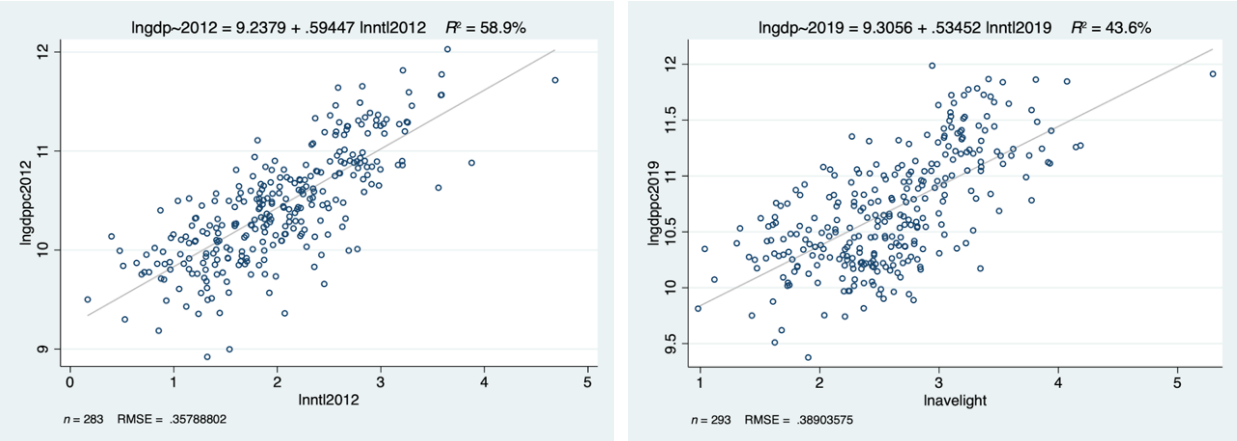
Recently, nightlight data (hereinafter referred to as NTL data) has been widely used in economics. NTL data can be seen as a sign of economic activity, so it is widely used as a proxy

for economic indicators. Henderson et al. (2012) show that night-lighting data can play a role in measuring economic performance, especially for low-income and middle-income countries with poor data quality. In addition, Chen and Nordhaus (2011) conclude that luminance data can be used as additional information for countries and regions with missing or poor data quality. Further, Lessmann and Seidel (2017) validated the relationship between nighttime lighting data and regional income on a global scale. They measured the degree of convergence and the determinants of convergence using out-of-sample predictions. All of the above papers validate the correlation between light intensity and economic activity and point to the promising future of nightlight data in economics research.

Compared to DMSP, VIIRS offers significant improvements in spatial resolution, dynamic range, quantization, calibration, and availability of spectral bands suitable for discriminating thermal radiation sources (Elvidge,2013).

4.2.3 The relation of lights and GDP in China

Figure 4.2 Scatter plot of light p.c. and GDP p.c. in 2012 and 2019



Source: Author's calculation using data from *China City Statistical Yearbook 2013 and 2020.*

We selected 2012 as the initial year and 2019 as the final year in this research. Figure 4.2 shows the scatter plots of regional total luminance per capita and regional GDP per capita in the year of 2012 and 2019. The scattered points are evenly distributed on both sides of the regression line, which implies the positive relationship between nightlight brightness and GDP. Then we use a simple linear regression model to further verify the relationship. Equation 1 directly verifies the relationship between GDP per capita and lights per capita at the city level. Based on Eq. (1), Eq. (2) includes GDP per capita at the provincial level.

Referring to Lessmann and Seidel (2017), we added regional income at a higher dimension to increase the accuracy of the estimation.

$$\log(\mathrm{GDPpc_i}) = \beta_0 + \beta_1 \log(\mathrm{light_i}) + \varepsilon \tag{4.1}$$

$$\log(\mathrm{GDPpc_i}) = \beta_0 + \beta_1 \log(\mathrm{light_i}) + \beta_2 \log\left(\mathrm{GDPpc_j}\right) + \varepsilon \tag{4.2}$$

where $\mathrm{GDPpc_i}$ means GDP per capita at the city level, $\mathrm{GDPpc_j}$ means GDP per capita at the provincial level, and $\mathrm{light_i}$ means light intensity per capita at the city level.

Table 4.1 Regression of Light p.c. and GDP p.c. in 2012 and 2019

Dependent variable: $\log(GDPpc_i)$	2012		2019		Difference (2012-2019)	
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Controlled provincial GDP	OLS	Controlled provincial GDP	OLS	Controlled provincial GDP
$\log(light_i)$	0.594^{***}	0.529^{***}	0.535^{***}	0.448^{***}	0.193^{***}	0.169^{***}
$\log(GDPpc_j)$		0.432^{***}		0.650^{***}		1.682^{***}
Constant	9.238^{***}	4.869^{***}	9.306^{***}	2.514^{***}	0.154^{***}	-0.488^{***}
Obs.	283	283	293	293	282	282
R-Squared	0.590	0.645	0.436	0.607	0.027	0.537
AIC	223.527	184.066	280.258	176.193	116.120	-91.261

Note: *, **, *** denote statistical significance at 10%, 5% and 1%, respectively. Robust standard errors in parentheses.

Source: Author's calculation using data from *China City Statistical Yearbook 2013 and 2020.*

The results of Eq. (1) and Eq. (2) are reported in Table 4.1. The regression results in Column (1) and Column (3) verify that lights and GDP are significantly and positively correlated at the city level in 2012 and 2019. After including provincial GDP, Column (2) and Column (4) still hold the relationship between lights and GDP. By comparing the results in Column (1) and Column (2), we find that the R-squared increases from 0.59 to 0.645 with the inclusion of provincial GDP data, while the R-squared increases from 0.436 to 0.607 in 2019. This means that including provincial GDP gives a better estimation, so we should use the model with the inclusion of provincial GDP for out-of-sample estimation. The results of the AIC also support this conclusion. In addition, we separately differenced the GDP data and the lighting data over the period 2012-2019 to derive the amount of growth in the GDP data and the lighting data. Column (5) and (6) are regressions for the relationship between the growth amounts of these two variables, and the results show that it is also significant in growth amounts. The growth of light intensity can also represent the growth of GDP. It implies that lighting data can

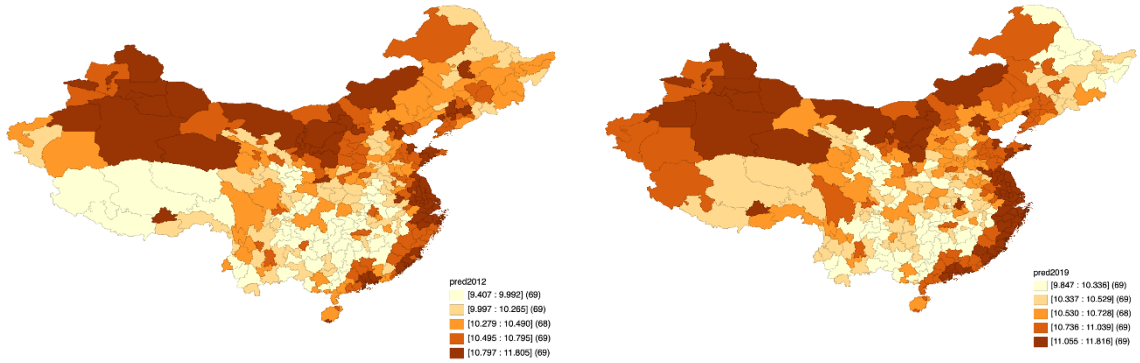
be used as a proxy for economic data, not only for cross-sectional data, but also for intertemporal comparisons.

Based on the regression results in Table 4.1, we derive the estimated equations (Eq. (4.3) and Eq. (4.4)) for 2012 and 2019 respectively. On the right-hand side of the equation, we have complete municipal lighting data and provincial GDP data. We computed municipal GDP data for 2012 and 2019 based on Eq. (4.3) and Eq. (4.4).

$$\log(\mathrm{GDPpc_i}) = 4.869 + 0.529\log(\mathrm{light_i}) + 0.432\log(\mathrm{GDPpc_j}) \tag{4.3}$$

$$\log(\mathrm{GDPpc_i}) = 2.514 + 0.448\log(\mathrm{light_i}) + 0.650\log(\mathrm{GDPpc_j}) \tag{4.4}$$

Figure 4.3 Compare of estimated GDP p.c. in 2012 and 2019



Note: The left graph shows estimated results of 2012; the right graph shows results of 2019.

Source: Author's calculation using data from *China City Statistical Yearbook 2013 and 2020.*

Figure 4.3 compares the regional development levels from 2012 to 2019 based on the estimated data. In the absence of missing values, we can have a comprehensive perception of the regional development of China. In general, the North is more developed than the South, and

the coastal is more developed than the inland region. The top 20 percent of cities in terms of GDP per capita is located in Xinjiang, Inner Mongolia, the Yangtze River Delta, and the Pearl River Delta. By 2019, the proportion of developed cities in coastal provinces had gradually increased, especially in Zhejiang Province. In addition, the biggest changes are the decline of the Northeast and the development of Tibet and its surrounding areas. Most cities in the Northeast decreased by one quintile, while most cities in Tibet Province improved by one quintile.

4.3 Methodology

4.3.1 Absolute Convergence Model

$$g_{i,0-T} = a + \beta \ln\, y_{i0} + \mu_i \tag{4.5}$$

Where $g_{i,0-T}$ is the growth of region i during from initial period to time T, y_{i0} is the regional income of region i in the initial year. β reflect the relationship between growth rate and regional income. If the value of β is negative, it means beta convergence in this economy. (Barro and Sala-i-Martin, 1992)

4.3.2 Conditional Convergence Model

$$g_{i,0-T} = a + \beta \ln\, y_{i0} + \sum_j \phi_j X_{it} + \mu_{it} \tag{4.6}$$

Where X_{it} is the conditioning factor to the convergence process, $\sum_j \phi_j X_{it}$ is the linear combination of the conditioning factors, β reflect the relationship between growth rate and regional income. According to Pan (2015), we consider physical capital, human capital,

marketization, openness and industrialization as conditioning factors. Variables and meanings are explained in Table 4.2.

Table 4.2 Meanings of Conditioning Factors

Variables	Term	Meaning
SK	Physical capital investment	The ratio of physical capital investment in GDP
LAB	Human capital	The ratio of employed people in total population.
GOV	Government expenditure	The ratio of government expenditure in GDP
FDI	Openness	The ratio of FDI in GDP
IND	Industry structure	The ratio of 2nd industry in GDP

4.3.3 Geographically Weighted Regression (GWR) and Its Extensions

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{n} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i \tag{4.7}$$

Where (u_i, v_i) is the geographical coordinates of point i, $\beta_j(u_i, v_i)$ is the coefficient of variable j at point i. In OLS regression, for each independent variable we will get only one coefficient. But in GWR, coefficient varies with the geographical location, it allows to obtain different coefficients at different observation points.

MGWR is an extension of GWR, it allows different bandwidths for different independent variables. The bandwidth indicates how many points are included in the local regression. The GWR model uses the same bandwidth for all independent variables, while MGWR allows the relationship to vary over different spatial scales.

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{n} \beta_{bwj}(u_i, v_i)x_{ij} + \varepsilon_i \quad (4.8)$$

where *bwj* in β_{bwj} indicates the bandwidth used for calibration of the jth variable.

4.4 Results

4.4.1 β Convergence Results

Table 4.3 β Convergence Results

	Dependent variable: Growth rate (2012-2019)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Absolute conv.	SK	LAB	GOV	FDI	IND	ALL
β	-0.192^{***}	-0.217^{***}	-0.204^{***}	-0.200^{***}	-0.203^{***}	-0.179^{***}	-0.220^{***}
$\beta\ of\ SK$		-0.205^{***}					-0.185^{***}
$\beta\ of\ LnSH$			0.092				0.009
$\beta\ of\ GOV$				-0.139^{**}			-0.112
$\beta\ of\ FDI$					0.749		0.831
$\beta\ of\ INDUSTRY$						-0.282^{**}	-0.306^{***}
Speed of conv.	3.05%	3.49%	3.26%	3.19%	3.24%	2.82%	3.55%
Halflife time	22.76	19.83	21.27	21.74	21.38	24.60	19.53
R-Squared	0.221	0.278	0.223	0.233	0.225	0.244	0.312
Obs.	270	270	270	270	270	270	270

Note: *, **, *** denote statistical significance at 10%, 5% and 1%, respectively. Robust standard errors in parentheses.

Source: Author's calculation using data from *China City Statistical Yearbook 2013 and 2020.*

Table 4.3 compares the results of absolute convergence and conditional convergence with the inclusion of conditioning factors. Column 1 shows the results for absolute convergence, and the newly estimated city-level data supports the conclusion of absolute convergence when no other control variables are considered. The convergence rate of absolute convergence is 3.05%, which is higher than the 2% convergence rate of Barro and Sala-i-Martin (1992). Since China

is a developing country with a fast economic growth rate, it is reasonable to have a faster convergence rate in the short run.

Columns 2 to 6 show that controlling for physical capital, labor, government spending, and FDI can accelerate the convergence rate, and physical capital has the greatest impact on the convergence rate. Controlling for physical capital increases the convergence rate from 3.05% to 3.49%, while the other factors have less impact on the convergence rate. This suggests that the physical capital factor has the greatest impact on economic growth in China. A surprising result is in Column 6, where the overall rate of convergence decreases after controlling for the industrial structure. One possible reason is that China's economic growth engine is gradually shifting from manufacturing (secondary industry) to services and IT industries (tertiary industry). Therefore, over-reliance on manufacturing is not conducive to economic convergence. In column 7, if we consider all the conditional factors, the convergence rate increases from 3.05% to 3.55%. Accordingly, Half-life time decreases from 22.76 years to 19.53 years.

4.4.2 GWR Results

Table 4.4 GWR Results

	OLS	GWR
β	-0.450^{***}	
Mean of β		-0.589
S.D.	0.048	0.280
Min		-1.259
Median		-0.598
Max		0.137
R-Squared	0.202	0.700
AIC	902.436	639.696
Obs.	344	344

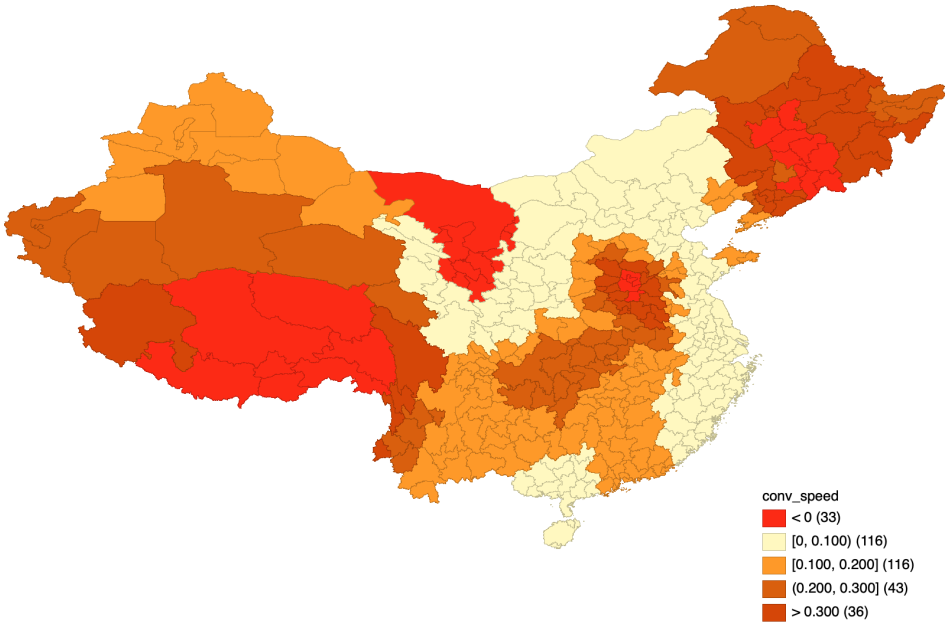
Note: *, **, *** denote statistical significance at 10%,5% and 1%, respectively. Robust standard errors in parentheses.

Source: Author's calculation.

Table 4.4 compares the results of using the OLS and GWR models to measure the rate of convergence. OLS results show that there is significant convergence from 2012 to 2019. The coefficient of β (-0.45) represents the average effect of all the observations, i.e., the economy of a lagging region is growing faster than that of a developed one. On the other hand, GWR is essentially running local regressions centered on each sample point, so it can give more information about the details of the regressions. It assigns large weights to samples close to the center and small weights to samples far from the center. For example, the median, minimum and maximum values of β are given in Table 4.4. The maximum value is positive, which

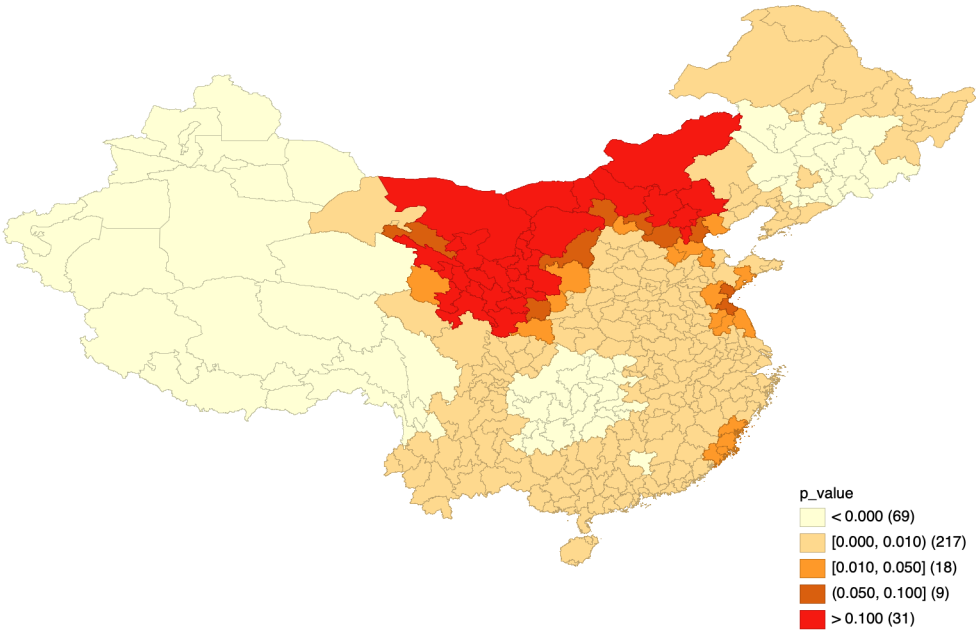
indicates that the results for some regions do not support convergence. It is worth noting that when we use the GWR model, the R-squared increases from 0.202 to 0.7 and the AIC decreases from 902.436 to 639.696, which indicates that the GWR model fits much better than the normal OLS regression.

Figure 4.4 Local Convergence Speed in 2012-2019



Source: Author's calculation.

Figure 4.5 Local Convergence Significance Level in 2012-2019



Source: Author's calculation.

GWR gives the regression coefficients for each city derived from the GWR model, then we can calculate the rate of convergence for each city. Figure 4.4 shows the rate of convergence. The regions marked in red are the regions with negative convergence speed, in other words, these regions do not have significant convergence and are therefore not discussed in this study. Those with a convergence speed of 0-0.1 account for 1/3 of the total, mostly in the north-central region and the southeast coast. Regions with a convergence rate of 0.1-0.2 also account for 1/3 of the total, and are distributed in northern Xinjiang, southwestern and southern regions. Other regions with convergence rates greater than 0.2 are mainly in southern Xinjiang, south-central, and most of the northeastern regions. In general, the speed of convergence is inversely related to the level of regional development. Regions with a high GDP per capita in the initial year will have a slow convergence rate. This is consistent with the results of this study. The north-central

and southeast coast of China are more developed; accordingly, the convergence rate is slower.

Figure 4.5 shows the significance of each local regression. Most of the regions in China are significant, with only 31 cities marked in red as insignificant, a proportion of less than 10%. The insignificant regions are concentrated in the central and northern parts of China, mainly in the Inner Mongolia Autonomous Region and Gansu Province.

4.4.3 MGWR Results

Table 4.5 MGWR Variable Bandwidth

	Bandwidth	
Variables	GWR	MGWR
GDP p.c. in 2012	44	50
Physical Capital	44	142
Human Capital	44	127
Marketization	44	233
FDI	44	68
Industry Structure	44	83

Note: *, **, *** denote statistical significance at 10%,5% and 1%, respectively. Robust standard errors in parentheses.

Source: Author's calculation.

Table 4.6 MGWR Results

	OLS	GWR	MGWR
β	-0.470^{***}		
Mean of β		-0.620	-0.688
S.D.		0.279	0.22
Min		-1.427	-1.227
Median		-0.586	-0.632
Max		-0.134	-0.322
R-Squared	0.221	0.686	0.786
AIC	702.866	513.481	455.976
Obs.	270	270	270

Note: *, **, *** denote statistical significance at 10%,5% and 1%, respectively. Robust standard errors in parentheses.

Source: Author's calculation.

Table 4.5 shows the comparison of the optimal bandwidth for each variable in the GWR and MGWR models. In GWR, the bandwidths of the different variables are fixed to 44. In contrast, different independent variables are allowed to be given different bandwidths in MGWR, meaning that different independent variables are included in different numbers of sample points when doing local regressions. In other words, the bandwidth also indicates the magnitude of spatial heterogeneity. For example, the bandwidth of the degree of marketability is 233, this means that the local regression of this variable needs to incorporate into nearly 2/3 of all sample points, which is the least heterogeneous of all the control variables.

Table 4.6 compares the results of OLS, GWR and MGWR. There is minor difference

between GWR and MGWR in the coefficients of β, but the results of R-squared and AIC show that MGWR has a better fit than GWR and is much better than the OLS regression. MGWR uses different bandwidths, i.e., different geographic scales, for each independent variable (Fotheringham, 2017). A separate neighborhood is used for each explanatory variable to account for the relationship between each explanatory and dependent variable. Therefore, the local coefficients estimated by MGWR are more accurate.

4.5 Concluding Remarks

This paper aims to examine whether regional economies in China converge, and compare the difference between considering and not considering spatial effects. This study first addresses the issue of missing GDP data. Using nighttime light data and provincial GDP data, linear regression is employed to construct an estimation strategy for city-level GDP, which results in a comprehensive dataset. Results show a significant positive correlation between nighttime light data and per capita GDP, indicating that nighttime light data can serve as a substitute indicator for per capita GDP.

Another important finding is that China's economy exhibits absolute convergence from 2012 to 2019. After considering spatial effects, the convergence rate increases from 3.03% to 4.42%, suggesting that spatial correlation accelerates the convergence process. In conditional convergence analysis, this study uses five control variables: physical capital, human capital, government expenditure, foreign investment, and industry structure. Physical capital, government expenditure, and foreign investment contribute to economic convergence, while human capital and industry structure slow down the convergence rate. Overall, the five variables' combined effect accelerates economic convergence's speed.

Finally, this study uses geographically weighted regression to measure the spatial heterogeneity of China's economic convergence. Results show that the speed of regional economic convergence in China varies greatly, with faster convergence in the underdeveloped western, northeastern, and southwestern regions and slower convergence in the eastern coastal areas. This finding is consistent with the theory of economic convergence, which posits that poor economies grow faster than rich ones.

5 Regional Human Capital Imbalance and Influencing Factors in China: A New Perspective at City Level

5.1 Introduction

Human capital, as one of the most important capital investments for social and economic development, is deeply affected by the accumulation and transaction of the knowledge, skills, and health resources. These resources are directly related to the education, career training, agriculture, food, medical systems, and so on. More crucially, the efficiency of transforming these resources into human capital should also be noticed as their qualities may be weighted as heavy as quantities.

Since the 1960s, economists have introduced human capital into traditional models of economic growth. On the one hand, according to Schultz (1961), combined with the theory from Smith (1776), Thünen (1875), and Fisher (1906), the modern human capital accumulation can be recognized as important as the physical capital investment. On the other hand, Becker (1962) extended the sources of human capital investment with the contents covering knowledge, skills, and health which people agreed with nowadays. There are already countless studies that include human capital as a crucial variable in the economic research. It has become a consensus that human capital plays a vital role in economic growth. However, throughout human history, none of the regions can experience a balanced and entirely sustainable economic development. The imbalances among the regions, communities, industries, and population groups are commonly detectable.

In China, regional imbalances remain a significant issue affecting sustainable economic development. The extent to which regional imbalances are caused by the uneven distribution of human capital and what factors influence the formation of regional human capital is still under

discussion. In the recent years, there is one leading human capital project which has been continuously establishing and upgrading annually. The Human Capital Index Project, measured by the China Center for Human Capital and Labor Economics Research (CHLR), is a project that calculates human capital levels in 31 provinces based on the Jorgenson-Fraumeni method (hereafter J-F lifetime income approach) (Jorgenson & Fraumeni, 1989, 1992b). This project is now widely considered the most reliable and systematic measurement of human capital indices in China. However, when studying regional economies, especially using spatial econometric analysis, we prefer a smaller data unit rather than using provincial data. Unfortunately, the Human capital index calculated by CHLR is a provincial level result. Therefore, this study aims to measure China's human capital index in terms of the city level and explore the relationship between human capital and regional economic imbalances.

Based on CHLR project, recent research uses the J-F lifetime income approach in calculating human capital (Li et.al, 2010; Li et.al, 2014; Li and Tang, 2015). It assumes that human capital can be traded in the market and that the transaction price is measured by the present value of expected future lifetime income. It uses a backward-looking approach to estimate future income through survival, progression, and employment rates, while the growth rate of labor income and its discount rate are also considered to calculate the present value of income. This method is currently the most common method used to study human capital. However, while the J-F method is based on years of schooling and income, it ignores the role of the quality of education and health. In addition, the J-F method requires survey data to calculate the income of different age groups. In China’s case, the J-F method can be used in calculating provincial human capital but cannot be applied to a smaller geographical unit.

Unlike the CHLR project, this study refers to a measurement of the Human Capital Project (HCP) from the World Bank, which considers three aspects: survival, education, and health

(World Bank, 2018, 2020). Based on the HCP method, this study aims to improve the calculation method in three ways. Firstly, analogically to the quality of education, this study considers the impact of health quality on human capital formation. Without nationwide health survey data, this study used environmental and medical resource data as approximate proxies for the health quality index. Second, regarding the Mincer Equation, this study considers the effect of returns on education and analogously proposes the concept of health returns. Both returns are incorporated into the calculations of this study. Last, The World Bank standardizes the return to education worldwide at 0.08. This study considers regional heterogeneity in both education and health sides. Therefore, this study calculated the return to education and the return to health for each province using micro-survey data.

This study derived the human capital index for 2010 and 2019 at the city level based on an improved calculation method. The analysis between 2010 and 2019 explores the spatial distribution and temporal trends of regional human capital in China. Finally, the random forest algorithm in machine learning is also utilized to estimate the contribution of resource inputs to human capital from four perspectives: education, healthcare, employment and government expenditure.

This study has three leading findings. First, this study calculated the prefecture-level human capital index as well as education and health sub-indexes. Second, this study finds that China's regional human capital distribution is unbalanced, and this imbalance persisted from 2010 to 2019. Finally, using machine learning algorithms, this study ranks the importance of factors in forming human capital. Investment in science and technology has the greatest impact on human capital formation. In addition, the neighboring effect of human capital cannot be ignored.

5.2 Methodology

5.2.1 Measuring Human Capital Index for City level in China

Extended from the methodology presented by Human Capital Project (World Bank, 2018, 2019, and 2020), the whole process of human capital index measurement starts from the estimation of return to years of schooling and return to experience. By ensuring the household survey covers all across China, the Mincer type equation (Mincer, 1974: 124) become utilizable.

$$\ln Y_i = \alpha_1 + \beta_1 S_i + \beta_2 Exp_i + \beta_3 Exp_i^2 + \varepsilon_i \quad (5.1)$$

In equation (5.1), $\ln Y_i$ is the natural logarithm of an individual's annual salary, S_i denotes the schooling years. Exp_i represents years of experiences, and Exp_i^2 is the squared terms. ε_i is the error term. The purpose is to estimate the return to years of schooling: β_1, and return to experiences: β_2. These two coefficients are critical for transforming education and health data into human capital index. However, due to the data shortages, it is nearly impossible to estimate these two coefficients in China's city level. The existing datasets of household survey can only derive a return to years of schooling and a return to experiences across China. Therefore, the China Family Panel Survey (CFPS) is chosen for the evaluation as the similar methodology and results were used in the author's previous research (Zhang, 2022). However, different provinces have different sample sizes, and some regression results are not significant. It cannot provide a reasonable return to education and experience. By compromising the detailed information of each city and province, the country level regression result is selected for all the cities in the same year. It is still better to choose a constant number for all the years.

Referring to the World Bank's Human capital project, education, health, and survival are

also calculated separately. Before diving into the direct calculation, we should make quality-based adjustments for education and health indexes.

According to the original methodology introduced by the World Bank (2018), education quality was evaluated through harmonized test scores all around the world. Meanwhile, the same idea was strongly emphasized by Filmer et al. (2020) as they calculated the learning adjusted years of schooling (LAYS) which demonstrates a more realistic education status quo of the examined areas. By fowling the same intention on involving quality based human capital index evaluation, equation (5.2) and (5.3) are modified as education and health quality adjustment indicators, respectively.

$$R_i^b = \frac{L_i}{L_b} \quad (R_i^b = Education\ Quality\ Adjustment\ Indicator) \tag{5.2}$$

$$\theta_i^b = \frac{H_i}{H_b}\ (\theta_i^b = Health\ Quality\ Adjustment\ Indicator) \tag{5.3}$$

R_i^b is the education quality adjustment indicator for city i over benchmark city b. L_i represents the education scores for each city while L_b denotes the benchmark score. Here the benchmark score is the highest education score among all the cities. Under the same design, θ_i^b offers health quality adjustment for health sub-index. It is calculated with the health score of each city H_i over the benchmark city's health score H_b. The education score is the college entrance examination relative score provided by Best China University Rankings from Shanghai Ranking Consultancy's annual publication. In contrast, the health score is a total score of cities' air quality and hospital condition data aggregated by a de-dimensional principal component analysis algorithm. With the quality adjustment indicators, the adjusted average years of schooling and adjusted life expectancy can be transferred into quality-based variables.

Equation (4) and (5) are implemented for the estimation.

$$Adjusted\ AYS = Average\ Years\ of\ Schooling\ \times\ R_i^b \tag{5.4}$$

$$Adjusted\ LE =\ Life\ Expectancy \times \theta_i^b \tag{5.5}$$

With all the preparations settled, the following equation (5.6), (5.7), and (5.8) are designed for education, health, and survival indexes, respectively. Education index formed a similar structure to the World Bank's methodology, but here the subtracted 14 years of standard education should be explained as maximum years of schooling before 20 years old according to China's reality. Health index seemed to be quite different from the World Bank, but it also followed a similar concept. The adjusted life expectancy and the return to experience are utilized for the calculation. The benchmark of maximum age that should be subtracted in equation (5.7) are different due to the change in life expectancy. Finally, the survival index calculated with equation (5.8) is identical to the World Bank which involved under five years old mortality rate in each province for that year.

$$I_{Education} = \mathrm{e}^{\beta_1(Adjusted\ AYS-14)} \tag{5.6}$$

$$I_{Health} =\ \mathrm{e}^{(\beta_2(Adjusted\ LE-Benchmark)+\beta_3(Adjusted\ LE-Benchmark)^2)} \tag{5.7}$$

$$I_{Survival} =\ \frac{1-Under\ 5\ Mortality\ Rate}{1} \tag{5.8}$$

$$Human\ Capital\ Index_i = I_{Health} \times\ I_{Education} \times\ I_{Survival} \tag{5.9}$$

Lastly, the overall aggregation of the human capital index requires the multiplication of the three sub-indexes. Equation (5.9) demonstrates the final step of the human capital calculation.

5.2.2 Random Forest Algorithm

The random forest algorithm is essentially a bootstrap aggregating method (or bagging). Bootstrap (or bagging) means resampling with replacement (Chen, 2020). The specific steps of bagging are as follows. First, we do resampling with replacement to get B bootstrap samples. For example, B equals 500 in this paper. We can estimate a decision tree with each bootstrap sample. According to Chen (2020), we record the estimation results as:

$$\{\hat{f}^{*b}(\boldsymbol{x})\}, b = 1, \cdots, B \tag{5.11}$$

Then we take the average of the estimation results from B decision trees:

$$\hat{f}_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(\boldsymbol{x}) \tag{5.12}$$

Since the bagging method averages many decision trees, it reduces the variance of the estimates and thus increases the prediction accuracy of the model.

Bagging also provides an easy way to estimate test errors. Since we do resampling with replacement, each tree has some observations that are not used. These are called out-of-bag observations (hereafter OOB observations) and can be used as a test set.

Specifically, for any observation x_i, it does not appear in about B/3 of the total decision

trees, so it is an OOB observation for these B/3 decision trees. To obtain the out-of-bag prediction y for the ith observation, we average the predictions of the OOB observations.

$$\mathrm{MSE}_{\mathrm{OOB}} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{i,\mathrm{OOB}} - y_i\right)^2 \tag{5.13}$$

We can calculate pseudo R^2 using out-of-bag mean square error (OOB MSE).

$$\text{Pseudo } R^2 \equiv \frac{\mathrm{Var}\,(y) - \mathrm{MSE}_{\mathrm{OOB}}}{\mathrm{Var}\,(y)} = 1 - \frac{\mathrm{MSE}_{\mathrm{OOB}}}{\mathrm{Var}\,(y)} \tag{5.14}$$

5.3 Data

Table 5.1 Composition of Human Capital Index

Aspects	Variables	Source
Education quantity	Educational year	*China population census yearbook 2010 and 2020*
Education quality	College entrance exam score	Chinese Universities Ranking from Shanghai Ranking Website URL:https://www.shanghairanking.cn/rankings/bcur/2019
Health quantity	Life expectancy	Provincial Health Commission website (collected by authors)
Health quality	Surface PM 2.5, medical resources scores	Atmospheric composition analysis group URL: https://sites.wustl.edu/acag/datasets/surface-pm2-5/ (Wan et.al, 2021) URL:https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-07119-3#citeas
Survival	Under-5 mortality rate	Provincial Health Commission website (collected by authors)

Note: Data sources are listed on the table.

Data used for calculating the human capital index are listed in Table 5.1. Life expectancy and Under-5 mortality rate are based on provinces, collected by authors from the provincial Health Commission website. The remaining variables are city-level data. Since there is no countrywide health survey in China, we integrated the health quality index by merging surface PM 2.5, medical resources scores through PCA methods. Air quality and medical service scores are used as the basis of health quality since they are the determinants of health (WHO, 2017) (Li et.al, 2021) (Wang et.al, 2022).

Table 5.2 Feature Variables at City Level

No.	Variables	Meaning	Classification
1	gespc	government expenditure on science and technology per 10000 persons	Government Expenditure
2	geepc	government expenditure on education per 10000 persons	
3	espc	employees in science and technology service per 10000 persons	Employment
4	eepc	employees in education per 10000 persons	
5	ehpc	employees in healthcare service per 10000 persons	
6	ntpc	tertiary schools per 10000 persons	Education Inputs
7	nsspc	senior secondary schools per 10000 persons	
8	nppc	primary schools per 10000 persons	
9	ntr	tertiary education teacher-student ratio	
10	nsvr	secondary vocational school teacher-student ratio	
11	nstr	secondary school teacher-student ratio	
12	npr	primary school teacher-student ratio	
13	lcpc	library collection per 10000 persons	Cultural and Healthcare Input
14	nhpc	hospitals per 10000 persons	
15	nhbpc	hospital beds per 10000 persons	
16	nldpc	licensed doctors per 10000 persons	

Source: *China City Statistical Yearbook 2011-2020.*

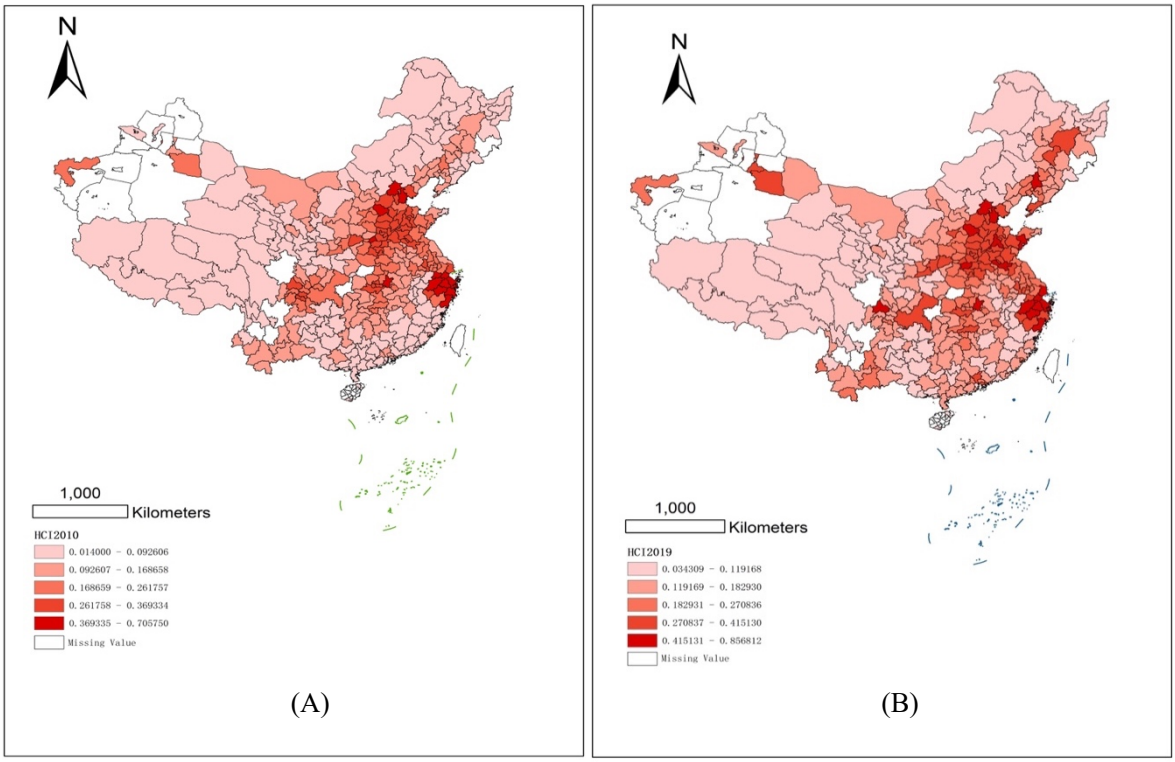
Feature variables used for machine learning are listed in Table 5.2. These variables can be considered as human capital inputs and can be classified into four categories. Data are collected by authors from *China City Statistical Yearbook (2011-2020).*

5.4 Results

5.4.1 Results of the Human Capital Formation

The human capital index in the following Figure 1 demonstrates with geographic locations on the map of China. Although the major human capital distribution in China stays similar between 2010 and 2019, there are several important differences that can be detected within the comparison. This section will focus on the results demonstrated by this newly derived human capital index.

Figure 5.1 Regional HCI Calculation Results in 2010 and 2019

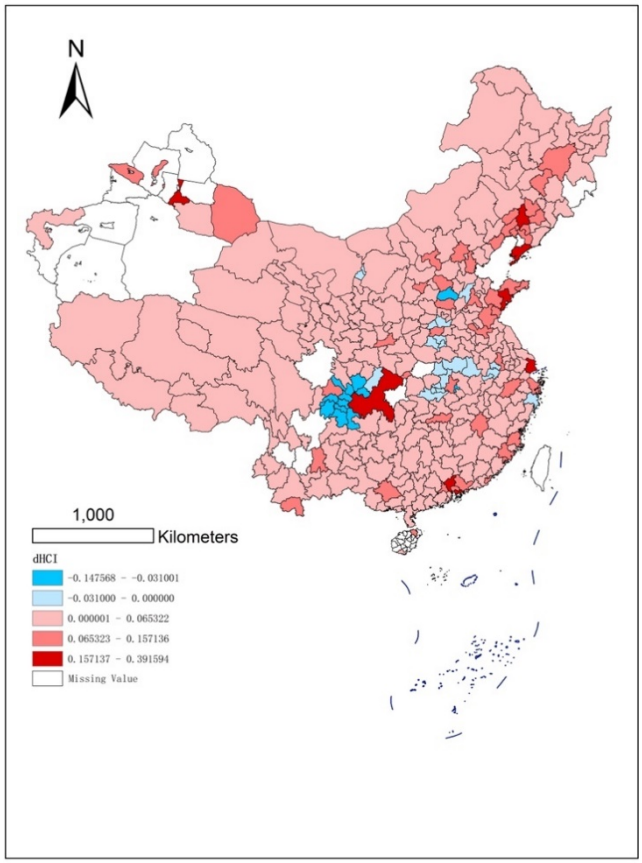


Source: Author's calculation.

In 2010 (Figure 5.1-A), China's human capital accumulation illustrated a clear imbalance between East and West, north and south regions. In the east coast regions, many cities around Shanghai, Jiangsu Province, and Zhejiang Province reflected its highly developed human capital. Similar patterns can also be detected around Beijing, Tianjin, the southern part of Hebei Province, and Shandong Province. Between the central and western area of China, some cities in Hubei Province, Chongqing, and Xinjiang Uygur Autonomous Region have demonstrated better human capital than their surrounding regions. Such a clustering effect also appeared in 2019 (Figure 5.1-B). Most already-developed cities and clustering regions in 2010, such as Beijing, Tianjin, Shanghai, Wuhan, and Chongqing, maintained high-level human capital. In the meantime, significant improvements of human capital development can be recognized around several new regions. In Northeast China, we can see Shenyang and Dalian have become the new core. In Southern China, Shenzhen has risen up, and also in the northwest of China, Tulufan and Urumqi developed significantly. However, some cities maintained low positions in their human capital development. They are mainly located in the western and northern regions of China which continuously stayed in the lowest group of human capital status. We also provide detailed information of China's human capital index map according to four different geographical groups in both 2010 and 2019 in the Appendix.

To furtherly identify the human capital changes, Figure 5.2 to 5.4 are designed to illustrate the value change of aggregated human capital index, education index and health index, respectively. The red areas indicate rising values and the blue areas indicate falling values. The darker the color, the greater the increase or decrease in value.

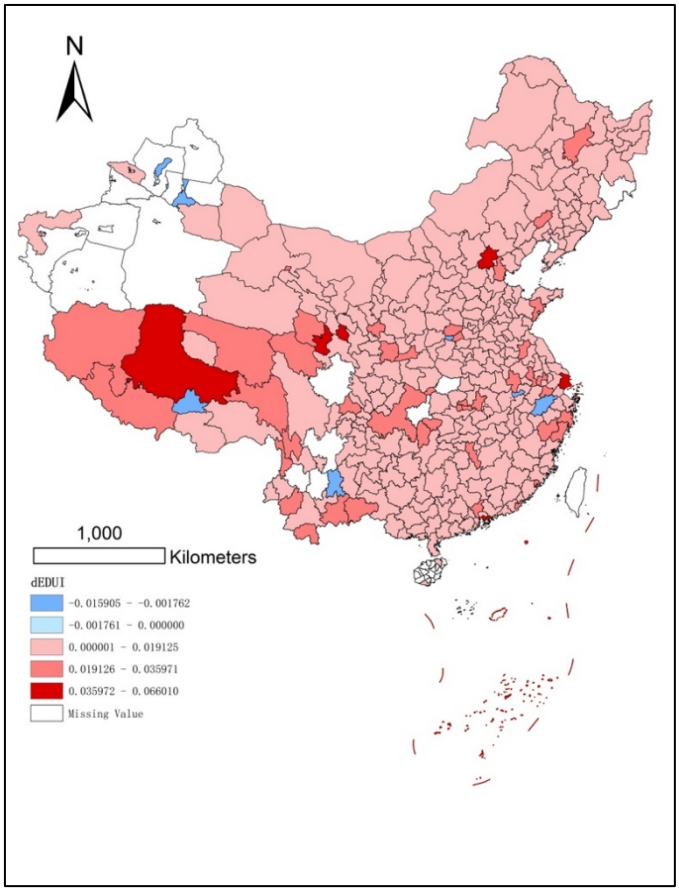
Figure 5.2 Regional HCI Difference Between 2010 and 2019



Source: Author's calculation.

From 2010 to 2019, the Human Capital Index in most regions in China experienced positive development. Among the fastest-growing areas, there are Chongqing, Shanghai, the eastern coastal part of Shandong Province, and the southern coastal region of Liaoning Province. In contrast, there are also some other regions where the human capital index has declined. One of the most significant declines was in the southeastern part of Sichuan Province, next to Chongqing. Some cities in the central region, Hubei and Hebei provinces, also experienced slight declines.

Figure 5.3 Regional Education Index Difference Between 2010 and 2019



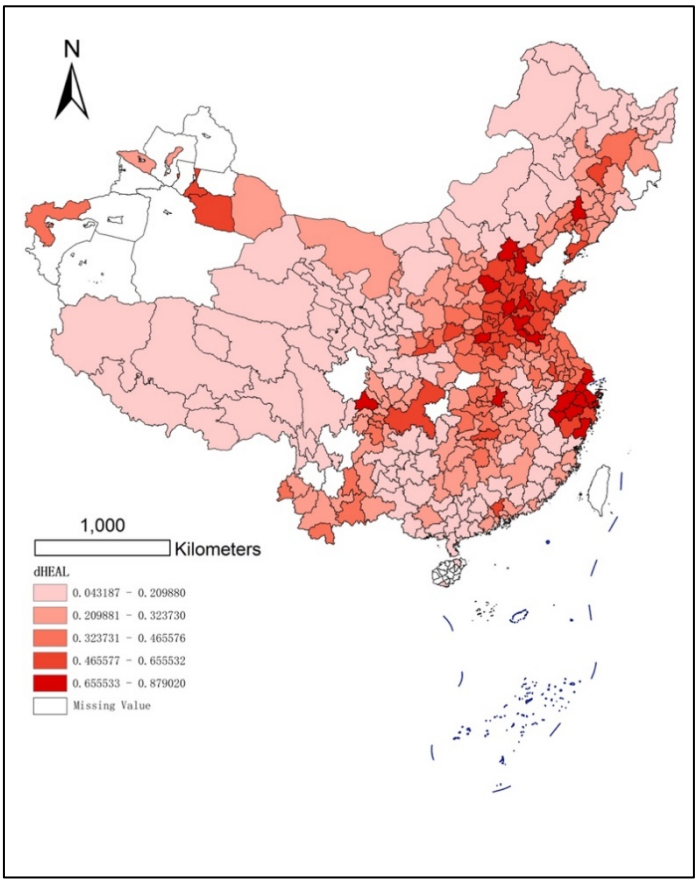
Source: Author's calculation.

Education and health are the two most essential aspects in constructing the human capital index. Here we also provide separated information on both education and health index changes with Figure 5.3 and Figure 5.4.

Figure 5.3 shows the results of the education indicator. The vast majority of regions are also offering an upward trend. The most significant increase is in the Tibetan and Qinghai area in western China, Beijing and Shanghai also demonstrated great development in the education field. Meanwhile, smaller scales of educational increases cam be recognized inmost cities from central, north, south, and eastern regions. The regions where the education index declined can

be found in Xinjiang Province, Tibet Province, Yunnan Province, and many cities in Zhejiang Province.

Figure 5.4 Regional Health Index Difference Between 2010 and 2019



Source: Author's calculation.

Figure 5.4 shows the regional health index changes between 2010 and 2019. All Chinese cities have improved in health indicators within this decade. The most significant increases are in the coastal regions of the eastern coastal line of China, especially in Shandong Province, Zhejiang Province, and the southern part of Hebei Province. Health indicators are composed of two quality indicators: air quality, health care provider scores, and one quantity indicator: life expectancy. As life expectancy and health care providers all around China have been

continuously increasing at a slow but steady pace in the past several decades, the most remarkable change among these three indicators is the air quality status. Therefore, the shift in health indicators mainly reflects air quality improvement.

With all the results from separated sub-indices to the general index and covering the detailed changes from 2010 to 2019 on China's city-level human capital index, the discussion and the overviews can be drawn at this stage. Since the human capital index across China demonstrates severe geographical characteristics, the imbalance issue is undoubtedly confirmed in the first place. However, several human capital development issues are still uncovered by these results that are urging further discussions.

One of the most important findings of the human capital index on China's city level is the significant geographical clustering effect. The economic or politically centered regions such as the Beijing-Tianjin-Hebei region, Yangtze River Delta Region (Shanghai, Jiangsu, and Zhejiang), Middle Yangtze River (Wuhan, etc.), Chongqing Chengdu City Cluster, and Guangzhou-Shenzhen-Hong Kong Greater Bay Area, were the most advanced developed regions in human capital from 2010 to 2019. Similar patterns can also be seen in the northeast and northwest of China, for these areas reflected visible growth of human capital development after ten years. The regional overlap between economic and human capital reminds researchers about the inner relationship of how dedicated policies may help generate sustainable growth.

In the meantime, the fundamental components of human capital, such as education and health, indicate more critical information on the potential policy issues of solving the human capital imbalances for those not included in any of those clustering regions. To reduce the education and health gaps and even reverse some of the declining situations, targeted studies on those policy determinants related to human capital and economic development are required.

5.4.2 Human Capital Policy Determinants with Machine Learning in 2010-2019

Machine learning, as a research method from computer science, is now gaining attention in economics research. Machine learning has advantages that traditional econometric models cannot match when using big data for prediction. In addition, Storm et al. (2020) points out that the functional forms used in econometric analysis often lack a theoretical basis and must be more flexible to capture the multiple interactions, nonlinearities, and heterogeneities common in social processes. In contrast, machine learning methods allow for highly flexible estimation, resolve model uncertainty, and handle large sample sizes effectively.

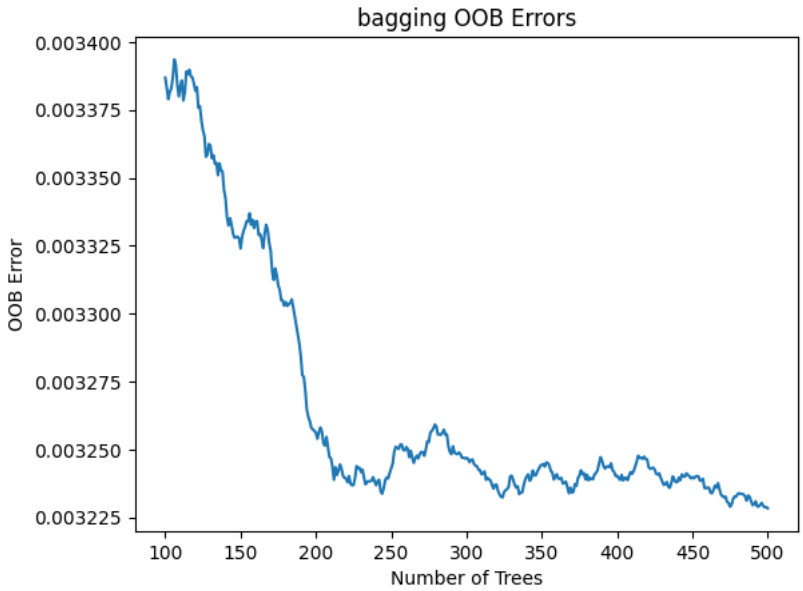
This section aims to find the impact factors of the human capital index. Based on the World Bank definition, we currently construct the human capital index regarding education, health, and so on. From the perspective of local governments, government inputs or urban environment indicators cannot directly act on indicators such as years of schooling, life expectancy, and others., and are not directly related to the urban human capital index. First, government inputs affect human capital development, a long-term, complex process that may have a time lag. Second, the relationship between government inputs and human capital is likely nonlinear and complex. Generally, regions with good regional economic development also have higher levels of human capital, so there may be a two-way causal endogeneity problem between urban variables and urban human capital. In summary, the relationship between these urban variables and urban human capital needs to be better explained using econometric models.

Otchia and Asongu (2021) used a machine-learning approach to predict African industrial development. They validated four sets of variables that influence industrial development through a random forest algorithm and found the most critical determinants from them. In this study, we adopt the methodology from Otchia and Asongu (2021) which originally referred to Breiman (1996;2001). In the following section, we use our calculated human capital index as

the response variable and the urban environment variable as the characteristic variable to find the determinants of the urban human capital index.

In machine learning, the objective is to find the importance of environmental variables for human capital at the city level. The response variable (Y) is the human capital index we calculated in Chapter 5.2. The feature variables (X) are city-level environmental variables, which are listed in Table 5.2. The feature variables are inputs regarding education, healthcare and government expenditure, which are related with regional human capital formation.

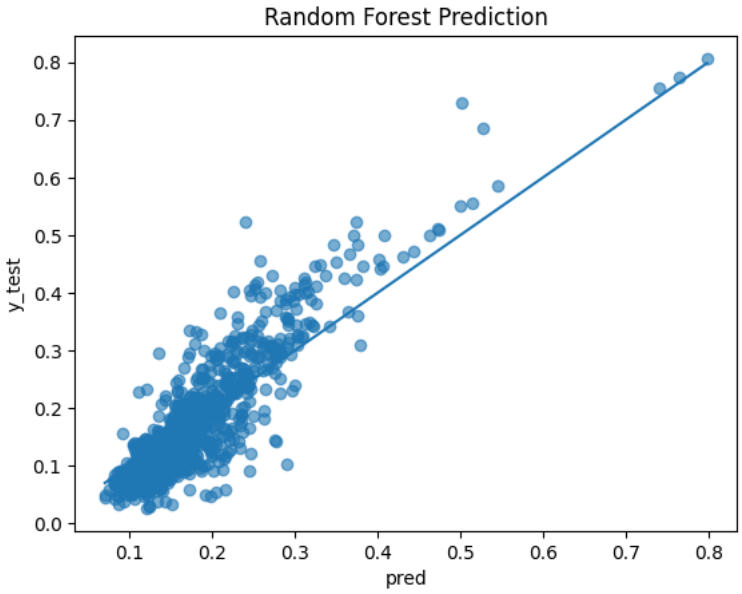
Figure 5.5 Bagging OOB Errors



Source: Author's calculation.

In Figure 5.5, the X-axis is the number of decision trees, and the Y-axis is the out-of-bag mean squared error. The mean squared error decreases as we increase the number of decision trees. When the decision tree is over 300, the mean squared error gradually stabilizes. In practice, we set the number of decision trees to 500.

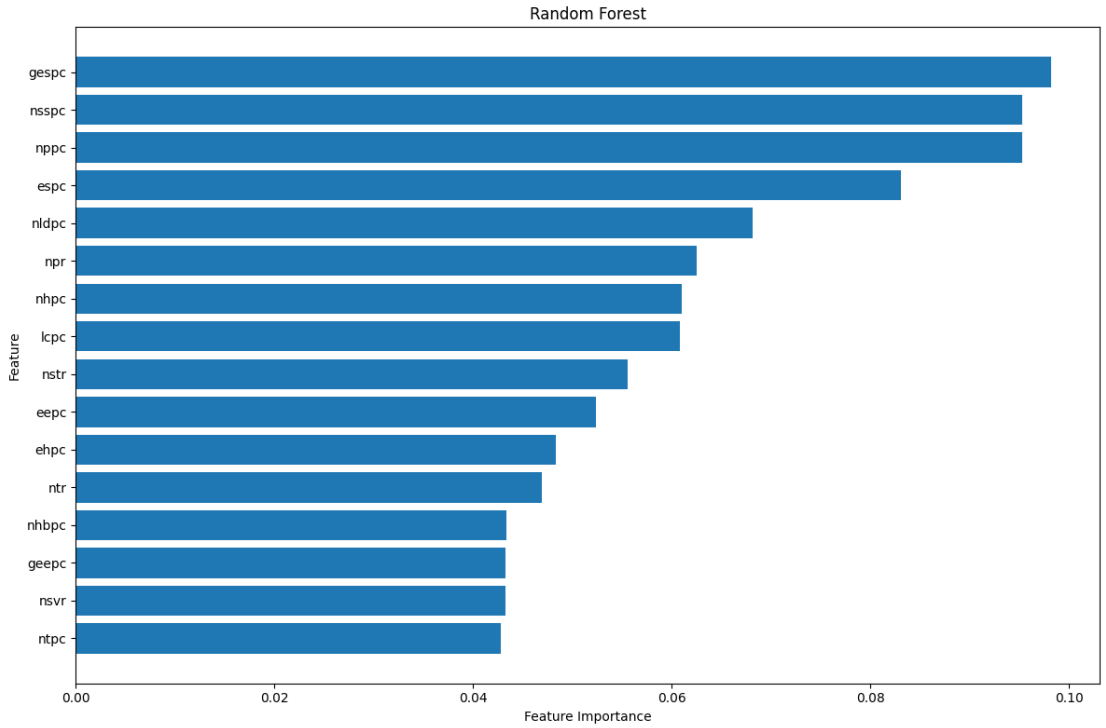
Figure 5.6 Scatterplot of Predicted Value and Actual Value in Random Forest Predication



Source: Author's calculation.

Figure 5.6 shows the scatterplot of predicted value and actual value. Most of the scatters are distributed on both sides of the 45-degree line, which indicates that the random forest achieves good prediction results. Pseudo R^2 is 0.760, which is much higher than 0.271 from linear regression algorithm. The prediction results have the potential to improve. Now we are using a dataset of 16 feature variables, 10 years and 344 cross-sectional units. If we can have a larger dataset, the prediction results should be better.

Figure 5.7 Variable Importance Ranking in 2010-2019



Note: Please refer to Table 5.2 for the full name of feature variables.

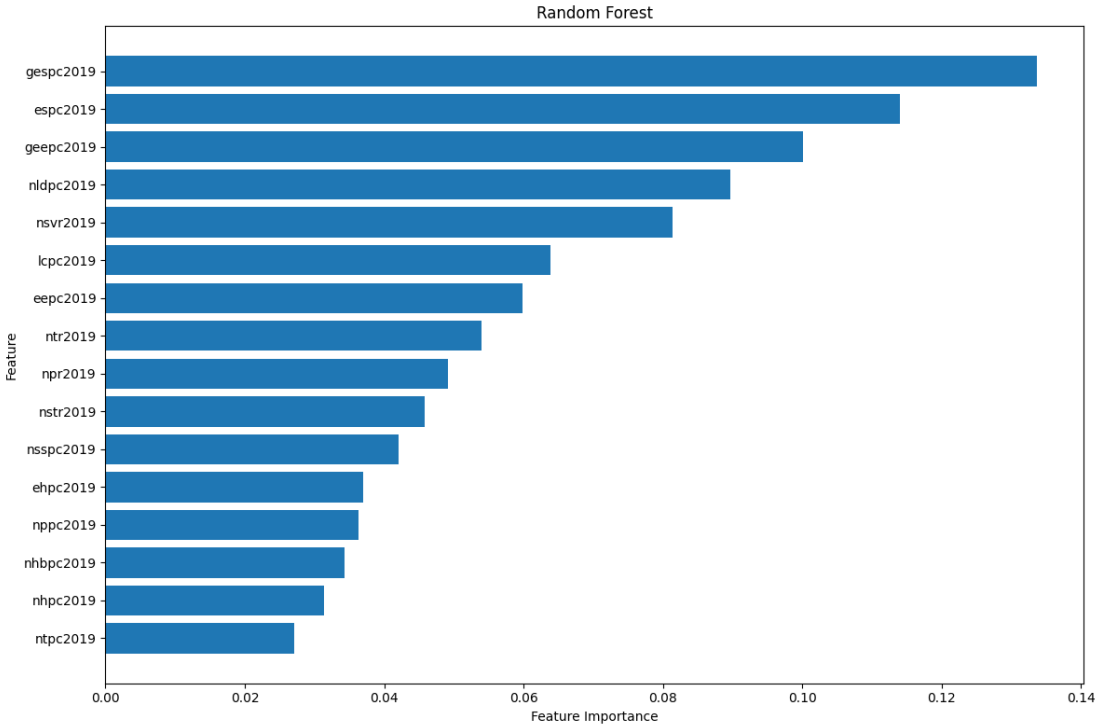
Source: Author's calculation.

In the decision tree algorithm, only one variable is used for each node split. We can use the magnitude of the decrease in the sum of squared residuals at the time of node splitting as the contribution of each variable. In the random forest, we can derive the importance of the variables by averaging the decreasing magnitude.

Figure 5.7 ranks the importance of the feature variables (X) for the human capital index (Y). The most important explainer is the government expenditure on science and technology. The number of senior secondary schools and primary schools, employees in science and technology service, and the number of licensed doctors is also playing vital roles in regional human capital.

5.4.3 Machine Learning with spatial lag variables in 2019

Figure 5.8 Variable Importance Ranking in 2019



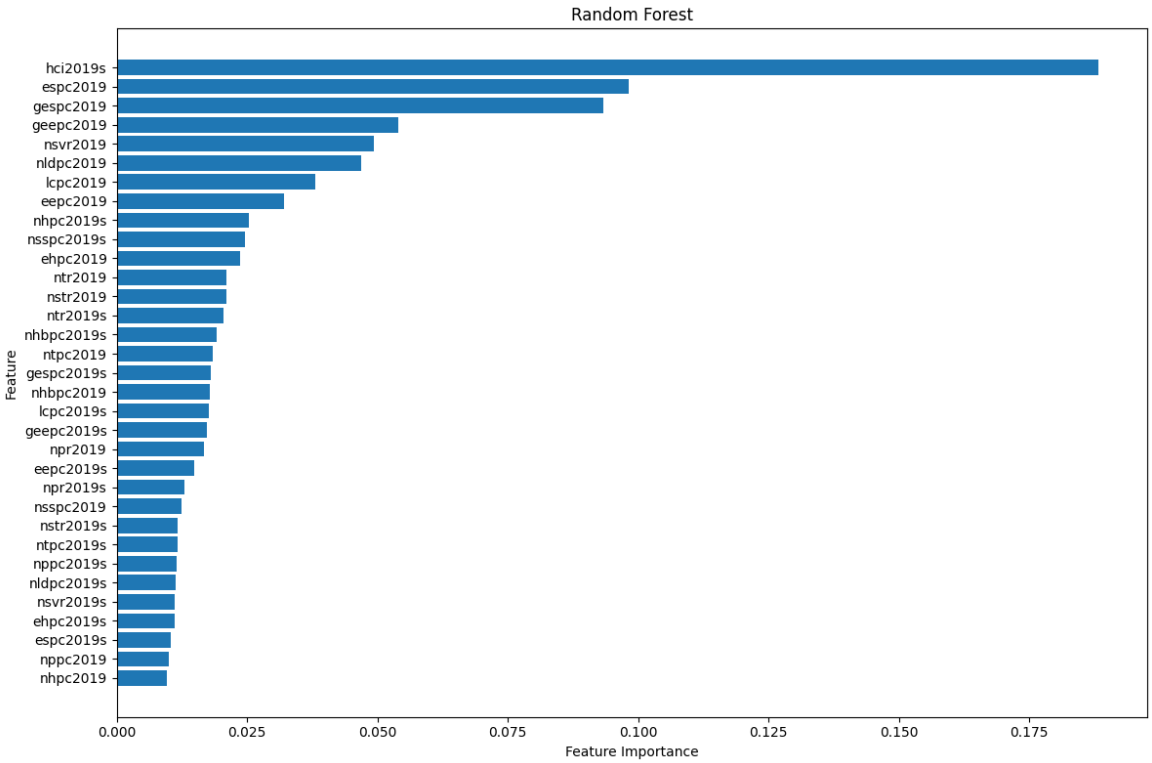
Note: Please refer to Table 5.2 for the full name of feature variables.

Source: Author's calculation.

In Figure 5.8, we have selected a sample from 2019 to repeat the machine learning process above. Compared to the overall sample for the decade 2010-2019, government expenditure on science and technology remains the most important feature variable, employees in science and technology services and licensed doctors are also in the top five, which indicates that the results derived from machine learning are robust. However, senior secondary schools and primary schools, which ranked second and third, respectively, in the overall sample, rank 11th and 13th

in the 2019 single-year sample. Machine learning requires a large sample size, and the single-year sample is 1/10th of the overall sample, so the results of the total 10-year sample are more credible.

Figure 5.9 Variable Importance Ranking in 2019 (Adding Spatial Lag Variables)



Note: Please refer to Table 5.2 for the full name of feature variables**;** the end 's' means. spatial lag of the original variable.

Source: Author's calculation.

To address the problem of insufficient sample size, we generated its spatial lag term for each variable in 2019 (Figure 5.9). The spatial lag term is the weighted mean value of that variable in its spatial neighborhood. Because human capital has spillover effects, we assume that investments in education, health, etc. in neighboring regions may also have an impact on human capital formation in the region. With the inclusion of the spatial lag term, our

characteristic variables increase from 16 to 32, and the sample size doubles. In addition to this, we also added the spatial lag term of the response variables.

The results show that the spatial lag of human capital has the greatest impact on human capital formation. This suggests that human capital development in one region is associated with human capital in neighboring regions, which is consistent with the findings of Yang (2023). Yang (2023) uses Markov chains to demonstrate that rich neighboring provinces increase the probability that a province becomes better off. The findings of Figure 5.9 similarly highlight the impact of spatial correlation on regional human capital development.

Similar to the findings in Figure 5.8, even with the introduction of the spatial lag term, employees in science and technology service, government expenditure on science and technology and the number of licensed doctors are still important factors influencing human capital. Notably, government expenditure on education and secondary vocational school teacher-student ratio occupy the fourth and fifth positions, respectively. This result will be discussed at the end. Finally, except for the spatial lag of human capital, the spatial lag term of the characteristic variables has little effect on human capital.

5.5 Concluding Remarks

The purpose of this paper was to calculate regional human capital index using city-level data. We referred to the World Bank's calculation methodology on human capital and adapted it to the reality of China. According to the World Bank's calculation methodology, the human capital index consists mainly of years of schooling, life expectancy, and under-five mortality. One of the novelties of this study is the introduction of the concepts of education quality and health quality. The lack of data at the municipal level, especially on health aspects, is a challenge in the calculation. We tried to find proxies and calculated city human capital indices for 2010

and 2019.

Based on the calculated results, we extracted the education and health indexes from the final calculation results and compared the change of the sub-index. Finally, we applied machine learning to human capital issues. We ranked the importance of 16 environmental variables through a random forest algorithm. As an attempt, we generated spatial lag variables to introduce spatial effects.

The main results indicate persistent regional imbalance in human capital from 2010 to 2019. High human capital values are clustered in eastern coastal provinces, especially in Shandong and Zhejiang Province. Other relatively high values appear in other provinces adjacent to Shandong Province (Hebei and southern part of Henan), as well as in northern Xinjiang. The relatively high value in Xinjiang seems abnormal because most western areas are undeveloped. We assume the high value in northern Xinjiang to be driven by high health quality. The results of the Sub-index show that most regions have improved in both health and education, with the education index improving more in western China and the health component improving more in the eastern coastal areas of China.

Last, we found that the random forest algorithm has more accurate estimation results than linear regression. The most crucial finding in this part is the ranking. Specifically, the top influencing factors are government expenditure on science and technology and senior secondary schools. These variables are highly related to education and R&D inputs. After introducing spatial lag variables, we found that the most critical finding is the spatial lag of the human capital index, which indicates that regional human capital has a strong neighbor effect and neighbors' high human capital levels contribute to the growth of local human capital. But the spatial lags of other environmental variables do not play an essential role in human capital growth. On the other hand, the results are robust in 2019 with or without the inclusion of

spatially lagged variables. Government investment in science and technology, government investment in education, the number of scientific and technical employees, and the number of practicing physicians are all important influences.

6 Conclusion

This doctoral dissertation explores the impact of human capital on China's economic development, and on regional imbalances, from three perspectives. The first analytical chapter is an exploratory analysis focusing on regional human capital. First, the regional distribution of human capital in China is uneven, with high human capital concentrated in the eastern coastal regions of China. Second, the spillover effects of human capital are detected with the help of Moran's I and geographic information visualization instruments. Finally, by comparing the results of traditional Markov chains and spatial Markov chains, this study identifies the role of spatial factors in regional human capital dynamics. That is, a high level of human capital in neighbors will contribute to the level of human capital in the region. Conversely, a low human capital levels of neighbors will also cause the deterioration of human capital in the region.

The second analytical chapter re-estimates the economic convergence in China using nighttime lighting data. First, to address the missing values at the city level, this study estimates urban GDP using nighttime lighting data and provincial GDP data. Second, after obtaining the complete dataset, this study verifies the existence of absolute beta convergence and conditional beta convergence for the Chinese economy from 2012 to 2019. Finally, this study finds the heterogeneity of the convergence process in the Chinese economy through geographically weighted regressions and measures the rate of convergence for each city. The regional convergence rates of China measured in this study are consistent with the convergence hypothesis, i.e., the convergence rates are faster in less developed regions such as central, western, and northeastern China.

The third analytical chapter measures the level of human capital in China from the city dimension, referring to the World Bank's algorithm. This study supplements the health quality and education quality aspects of the World Bank algorithm with substitutions based on the

actual situation in China. This study mainly uses the random forest algorithm in machine learning to rank the importance of factors for the years 2010-2019, and the results show that number of primary and secondary schools, government expenditure and employees on science and technology are important factors in human capital formation. In addition, we introduced spatial effects in machine learning for the single years 2010 and 2019, i.e., we included spatial lagged terms for 16 feature variables and one original response variable. The results show that neighbors' human capital is the most important factor to regional human capital formation.

In general, this study conducted an exploratory analysis of China's uneven development from economic and human capital perspectives. This study has commonalities and differences from the two perspectives. In China, regional economy and human capital exhibit similar regional distributions, with the eastern coastal regions significantly higher than the western inland regions.

Different research methods were employed for this study's economic and human capital sections. For the regional economy section, this study used β convergence to examine whether absolute or conditional convergence exists at the city level in China. It can indicate whether regional development gaps are narrowing. For the human capital section, the study utilized Markov chains and machine learning to explore the factors influencing human capital formation. Additionally, a comparison between non-spatial and spatial econometric models was an innovative aspect of this study. The emphasis on exploring the impact of spatial effects on regional economic and human capital formation varied. In the section on regional economy, the study emphasized spatial heterogeneity, highlighting different economic relationships across various spatial locations. On the other hand, in the human capital section, the study emphasized spatial correlation, indicating that neighboring areas have spatial interdependence and influence.

This study partially discusses the relationship between regional economic development

and human capital. Chapter Four discusses the role of human capital as an explanatory variable in regional economic convergence. The results indicate that human capital accelerates economic convergence at the city level. Furthermore, Chapter Five explores the impact of economic investments in education, health, and other areas on human capital. The results rank the importance of each economic input, with government funding for science and technology being the most significant.

Regarding the above conclusions, this study proposes the following policy recommendations:

1. To address the regional imbalance in China's human capital development, it is vital to encourage talent mobility from developed to underdeveloped areas to balance the human resources situation in different regions. Providing incentives such as tax benefits, housing subsidies, and career development opportunities can attract high-quality talent to underdeveloped areas, promoting balanced human capital development nationwide.

2. This study confirms the positive neighborhood effect of human capital in China, meaning that neighboring regions influence a region's level of human capital. The government should establish robust regional cooperation mechanisms to facilitate collaboration and communication among areas. Mainly, it is essential to strengthening the establishment of talent exchange and training mechanisms, encouraging talent mobility and cultivation across regions, thereby harnessing the positive neighborhood effect of human capital.

3. Emphasis should be placed on technology investment and primary and secondary education. Machine learning has identified R&D expenditure as the most significant factor affecting human capital formation. The government should increase the proportion of GDP allocated to scientific and technological research and development. It can be achieved by enhancing funding for research projects and establishing innovation funds to support

researchers in conducting basic and applied research, thus promoting the growth of technological innovation. Additionally, the government should increase investment in primary and secondary education to ensure the equitable distribution of educational resources. It can be done by increasing the education budget and the proportion of GDP allocated to education, ultimately improving the quality of primary and secondary education.

References

Aaron van Donkelaar, Melanie S. Hammer, Liam Bindle, Michael Brauer, Jeffery R. Brook, Michael J. Garay, N. Christina Hsu, Olga V. Kalashnikova, Ralph A. Kahn, Colin Lee, Robert C. Levy, Alexei Lyapustin, Andrew M. Sayer and Randall V. Martin. 2021. Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology.* doi:10.1021/acs.est.1c05309.

Acemoglu, D. 2012. What Does Human Capital Do? A Review of Goldin and Katz's The Race Between Education And Technology. *Journal of Economic Literature.* 50(2): 426-463.

Anselin, L., Sridharan, S., & Gholston, S. 2007. Using Exploratory Spatial Data Analysis To Leverage Social Indicator Databases: The Discovery Of Interesting Patterns. *Social Indicators Research.* 82(2): 287-309.

Bai Jingfeng, Zhang Haijun. 2014. Spatio-temporal Analysis of Economic Growth in the Central Plains Economic Region Based on EOF and GWR Model[J]. *Geographical Research.* 33(07): 1230-1238.

Barro, R. J., Sala-I-Martin, X., Blanchard, O. J., & Hall, R. E. 1991. Convergence Across States and Regions. *Brookings Papers on Economic Activity.* 1991(1): 107–182. https://doi.org/10.2307/2534639

Becker, G. S. 1962. Investment In Human Capital: A Theoretical Analysis. *Journal Of Political Economy*. 70(5, Part 2), 9-49.

Bode, E., & Nunnenkamp, P. 2011. Does foreign direct investment promote regional development in developed countries? A Markov chain approach for US states. *Review of World Economics.* 147(2): 351-383.

Chen, Q. 2020. Machine Learning and Python Application. *Beijing: Higher Education Press.*

China Center for Human Capital and Labor Market Research (CHLR). 2020. Human Capital Index Project 2020. http://humancapital.cufe.edu.cn/en/Human_Capital_Index_Project/Project_2020/Human_Capital_Brief_Report.htm. Accessed on January 25, 2022.

Chu, E., & Cao, C. 2019. Does Talent Flow Narrow the Regional Economic Disparities? —— Empirical Evidence from Technology Transfer. *Financial Science.* (9): 99-112.

Danny T. Quah. 1996. Twin Peaks: Growth and Convergence in Models of Distribution Dynamics. *The Economic Journal.* 106(437): 1045–1055. https://doi.org/10.2307/2235377

Fang, C., & Luo, Y. 2016. A Study on The Impact of Educational Human Capital and Its Spillover Effects on China's Economic Growth - A Spatial Econometric Analysis Based on Lucas Model. *Education and Economics.* (4): 21-29.

Fisher, I. 1906. *The Nature of Capital and Income.* Macmillan and Cie.

Fraumeni, B. M., He, J., Li, H., & Liu, Q. 2019. Regional Distribution and Dynamics of Human Capital in China 1985–2014. *Journal of Comparative Economics.* 47(4): 853–866. https://doi.org/10.1016/j.jce.2019.06.003

Fotheringham, A. S., Yang, W., & Kang, W. 2017. Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*. 107(6), 1247-1265.

Fu, Y. 2014. *A Study on The Contribution of Human Capital and Its Structure to China's Economic Growth.* Changchun: Jilin University (Doctoral dissertation).

Hammond, G. W. 2004. Metropolitan/Non-Metropolitan Divergence: A Spatial Markov Chain Approach. *Papers in regional Science.* 83(3): 543-563.

He, X. G., Luo, Q., & Chen, J. L. 2020. High-Quality Human Capital and Upgrading of Urban Industrial Structure in China: Evidence from Enrollment Expansion. *Economic Review.* (4): 3-19.

Hugo Storm, Kathy Baylis, Thomas Heckelei. 2020. Machine Learning in Agricultural And Applied Economics. *European Review of Agricultural Economics.* 47(3): 849–892. https://doi.org/10.1093/erae/jbz033

Jorgenson, D., & Fraumeni, B. M. 1989. The Accumulation of Human and Nonhuman Capital, 1948-84. *In The Measurement of Saving, Investment, and Wealth* (pp. 227-286). National Bureau of Economic Research, Inc.

Jorgenson, D., & Fraumeni, B. M. 1992b. The Output of the Education Sector. *In Output Measurement in the Service Sectors* (pp. 303-341). National Bureau of Economic Research, Inc.

Kang, W., & Rey, S. J. 2018. Conditional And Joint Tests for Spatial Effects In Discrete Markov Chain Models Of Regional Income Distribution Dynamics. *The Annals of Regional Science.* 61(1): 73-93.

Le Gallo, J. 2004. Space-Time Analysis of GDP Disparities Among European Regions: A Markov Chains Approach. *International Regional Science Review.* 27(2): 138-163.

Li, H., Fraumeni, B. M., Liu, Z., & Wang, X. 2009. Human Capital in China. *Working Paper No. 15500; Working Paper Series.* National Bureau of Economic Research. https://doi.org/10.3386/w15500

Li, H., Jia, N., Zhang, X., & Fraumeni, B. M. 2013. Regional Distribution and Development Dynamics of Human Capital in China. *Economic Research.* 48(7): 49-62.

Li, H., Liang, Y., Fraumeni, B. M., Liu, Z., & Wang, X. 2013. Human Capital in China, 1985-2008. *Review of Income and Wealth.* 59(2): 212-234. https://doi.org/https://doi.org/10.1111/j.1475- 4991.2012.00517.x

Li, J., & Chen, Y. P. 2019. Regional Polarization and Dynamic Evolution of Human Capital Distribution: A Measure Based on The Provincial Dimension in China. *Statistics and Information Forum.* 34(6): 44-50.

Li Xiaofei, Zhao Lichen, Hou Fan, Lv Kewen. 2018. Spatial Knowledge Spillover and Regional Economic Growth: An Empirical Analysis Based on SDM and GWR Models[J]. *Soft Science.* 32(04): 16-19+30.

Lu, J., & Zhou, H. M. 2014. An Empirical Analysis of The Spatial Spillover Effects of Human Capital in Chinese Provinces: Based on ESDA Approach and Spatial Lucas Model. *Journal of Population.* 36(6): 48-61.

Lucas, R. E. 1988. On The Mechanics of Economic Development. *Journal of Monetary Economics.* 22(1): 3-42. https://doi.org/https://doi.org/10.1016/0304-3932(88) 90168-7

Ma, H. 2021. The Impact of Human Capital Distribution Structure on Technological Innovation: Also on The Mediating And Moderating Effects of Income Disparity. *Journal of Central China Normal University (Humanities and Social Sciences Edition).* 60(1): 34-44.

Mou Juan. 2010. Economic Spatial Analysis Based on GWR Model[D]. *Shandong University of Science and Technology.*

Nie, J. X., & Liu H. L. 2018. Spatial Pattern and the Resulting Characteristics of Talent Flows

in China. *Scientia Geographica Sinica.* 38(12): 1979-1987.

Otchia, C. and Asongu, S. 2021. "Industrial Growth in Sub-Saharan Africa: Evidence From Machine Learning With Insights From Nightlight Satellite Images". *Journal of Economic Studies*. 48(8): 1421-1441. https://doi.org/10.1108/JES-05-2020-0201

Peng, S. H. 2019. The Dynamic Evolution of Regional Human Capital Inequality and Its Spatial Distribution in China. *Journal of Central University of Finance and Economics.* (11): 115-128.

Qiu, J. P., & Wen F. F. 2010. Study on The Regional Distribution of Higher Education Resources in China. *China Higher Education Studies.* (7): 12-16.

Rey, S. J. 2001. Spatial Empirics for Economic Growth and Convergence. *Geographical analysis.* 33(3): 195-214.

Rey, S. J., Kang, W., & Wolf, L. 2016. The Properties of Tests For Spatial Effects in Discrete Markov Chain Models of Regional Income Distribution Dynamics. *Journal of Geographical Systems.* 18(4): 377-398.

Schultz, T. W. 1961. Investment in Human Capital. *The American Economic Review.* 51(1): 1-17.

Smith, A. 1776. *The Wealth of Nations.*

Su Fanglin. 2005. Spatial Statistical Analysis of R&D and Economic Growth in China[D]. *East China Normal University*.

Su Fanglin. 2010. GWR Empirical Analysis of R&D Knowledge Spillovers in Prefecture-level Cities[J]. *Mathematics in Economics*. 29(01): 41-51.

Sun, H. B. 2017. *Research on The Impact of Human Capital and Its Spatial Distribution on Industrial Structure Upgrading in China.* Changchun: Jilin University (Doctoral dissertation).

Tao, X. H., & Qi Y. W. 2013. Spatial-temporal Evolution Analysis of China's Regional Economy with Weighted Spatial Markov Chain Approach. *China Industrial Economics.* (5): 31-43.

Thünen, J. H. v. 1875. *Der Isolierte Staat in Beziehung Auf Landwirthschaft und Nationalökonomie (Ed. 3).* Wiegandt, Berlin.

Verginer, L., Riccaboni, M. 2020. Cities and Countries in The Global Scientist Mobility Network[J]. *Applied Network Science.* 5(1): 1-16.

Wan, S., Chen, Y., Xiao, Y. *et al.* 2021. Spatial Analysis and Evaluation of Medical Resource Allocation in China Based on Geographic Big Data. *BMC Health Serv Res* ***21****.* 1084. https://doi.org/10.1186/s12913-021-07119-3

Wang, W. J. 2013. *Study on The Role and Convergence of Human Capital on Regional Economic Growth.* Changchun: Northeast Normal University (Doctoral dissertation).

Wang, Q. 2014. *An Empirical Study on Human Capital and Economic Growth in Northeast China.* Changchun: Jilin University (Doctoral dissertation).

Wang, Y., Cui, C., Wang, Q., Ning, Y., & Yang, Z. 2021. Migration of Human Capital in The Context of Vying For Talent Competition: A Case Study of China's "First-Class" University Graduates. *Geographic Research.* 40(3): 743-761.

Wenxuan, Y. 2023. Human Capital Dynamics across Provinces in China: A Spatial Markov Chain Approach. *Forum of International Development Studies.* 53(8), 1-17.

World Bank. 2018. The Human Capital Project. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/30498. Accessed on January 28, 2022.

World Bank. 2020. The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/34432.

Xia, C. Y., & Zhang, X. 2021. In-depth Analysis Report on the Comparison and Benefit Evaluation of Talent Policies in 31 Provinces and Cities in China in 2021. https://blog.csdn.net/kymdidicom/article/details/117309723. Accessed on January 28, 2022.

Xu, Y., & Li, A. 2020. The Relationship between Innovative Human Capital and Interprovincial Economic Growth Based on Panel Data Model and Spatial Econometrics. *Journal of Computational and Applied Mathematics.* 365:112381.

Yan, L. G., & Zeng, S. M. 2020. Why Eastern Industries Have Difficulty Moving to The Midwest: An Explanation Based on Spatial Differences in Human Capital. *Economic Geography.* 40(1): 125-131.

Young, A. T., Higgins, M. J., & Levy, D. 2008. Sigma Convergence Versus Beta Convergence: Evidence from US County-Level Data. *Journal of Money, Credit and Banking.* 40(5): 1083-1093.

Zhang, J. H., & Gao, J. 2019. Study on the Problems and the Countermeasures of the Unbalanced Development of Regional Economy in the 40 Years of China's Reform and Opening up. *Contemporary Economic Management.* 41(2): 9-14.

Zhang, K., & Huang, L. Y. 2020. A Study on The Spatio-Temporal Evolution Characteristics of Human Capital Structure in China. *Research in Quantitative Economics and Technology Economics.* 37(12): 66-88.

Zhou, M., Li, Y. N., Yao, X., & Lu, Y. 2019. Human Capital Accumulation and Urban Manufacturing Export Upgrading in China: Evidence from Higher Education Expansion. *Management World.* (5): 64-77.

Zhang, Y. 2022. Estimating the Trend in the Returns to Education in China: Evidence from Longitudinal Data. *Forum of International Development Studies.* 52(10), 1-17.