2-20-2019

Analyzing the Population Density Pattern in China with a GIS-Automated Regionalization Method: Hu Line Revisited

Cuiling Liu
cliu36@lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

ANALYZING THE POPULATION DENSITY PATTERN IN CHINA WITH GIS-AUTOMATED REGIONALIZATION METHODS: HU LINE REVISITED

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Geography and Anthropology

by
Cuiling Liu
B.S., Shandong Jiaotong University , 2010
M.S., Capital Normal University, 2013
August 2019

ACKNOWLEDGEMENTS

The successful completion of my dissertation is a significant milestone for me. I would like to acknowledge the individuals who have provided major supports toward this process.

First, I would like to give tons of thanks to my advisor, Dr. Fahui Wang. He is a fountain of knowledge, and has generously given his time, expertise, and support to all my endeavors. Dr. Wang always gives his priority to students. He always tells his students: "Feel free to come to me if you need any help. No appointment needed." He has kept his promise. Every time I went to his office for help, he always put aside his work and attended to my questions immediately. Every time I sent him a draft paper, he always gave me feedback in a very short time, despite his busy schedules. He is a wonderful mentor.

Many thanks to my committee members, Dr. Nina Lam, Dr. Michael Leitner, and Dr. Gabriele Piccoli. I appreciate their invaluable services as my committee members. Their constructive advices in my general exam have helped me a great deal in improving my dissertation.

I would also like to thank my dear husband Yaping Xu. He is also my best friend and my "advisor". He encouraged me to pursue my academic career. He is always the first one, from whom I asked for help whenever I encountered any problems in life and research.

Last but not least, I wish to express my deepest gratitude to my parents, my grandparents, my brother, and my sister. They always encourage me to pursue my dreams. They have given me so much, but never asked for anything back.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The famous "Hu Line", proposed by Hu Huanyong in 1935, divided China into two regions of comparable area sizes that drastically differ in population: about 4% in the northwest part and 96% in the southeast. However, the Hu Line was proposed largely by visual examination of hand-made maps and arduous experiments of numerous configurations, and has been subject to criticism of lack of scientific rigor and accuracy. Furthermore, it has been over eight decades since the Hu Line was proposed. During the time, China sustained several major man-made and natural disasters (e.g., the World War II, the subsequent Civil War and the 1958-62 Great Famine), and also experienced some major government-sponsored migrations, economic growth and unprecedented urbanization. It is necessary to revisit the (in) stability of Hu Line. By using a GIS-automated regionalization method, termed REDCAP (Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning), this study re-visits the Hu Line in three aspects. First, by reconstructing the demarcation line based on the latest census of 2010 county-level population by REDCAP, this study largely validates and refines the classic Hu Line. Secondly, this research also seeks to uncover the underlying physical environment factors that shape such a contrast by proposing a habitation environment suitability index (HESI) model. In the third part, this study examines the population density change and disparity change over time by using all the six censuses (1953, 1964, 1982, 1990, 2000, and 2010) since the founding of the People's Republic of China. This study advances the methodological rigor in defining the Hu Line, solidifies the inherent connection between physical environment and population settlement, and strengthens the findings by extending the analysis across time epochs.

CHAPTER 1. INTRODUCTION

Population has profound impacts on natural resources, the environment and our social economy (Sutton et al., 1997). In turn, these factors also affect the population distribution. Therefore, the acquisition of precise spatial distributions of population and underlying driving factors is crucial for understanding their interactions, as well as for the management of these factors (Zhuo et al., 2009). The interlocking crises in population, economy, resources, and environment have always been the focus of countless papers, dozens of prestigious symposia, and a growing avalanche of books (e.g., Cassen, 1978; Wang et al., 2016).

As the world's most populous country, China has experienced formidable challenges related to its large population over the past 100 years. Based on the first census in 1953, the population was about 0.55 billion. When China conducted its latest population census in 2010, the population was more than 1.3 billion, one-fifth of the world's population. However, the numerous population was not evenly distributed. Most of the population live in the east part of China while a few of them live in the west. Since people are both labors and consumers, i.e., a vital element in the supply and the demand sides of an economic system, population distribution pattern is often used as a good proxy for measuring the economic development pattern. One of the best known studies on the uneven population distribution in China was the famous "Hu Huanyong Line" (simply referred to as "Hu Line" thereafter), which is proposed by Hu (1935, 1990). The Hu Line begins from Heihe in Heilongjiang Province in the northeast to Tengchong in Yunnan Province in the southwest of China (Figure 1), and it is considered as one of the great geographic discoveries of China (Shan, 2009). With similar area sizes, the northwest side of the line has only about 4% of the country's total population, and the southeast side has nearly 96% of the population (Hu,1935), displaying a stunning contrast in land use intensity on the two sides. Many studies attempted to

find possible reasons for the difference (e.g., Yuan, 1993; Shan, 2009; Qi et al., 2015).
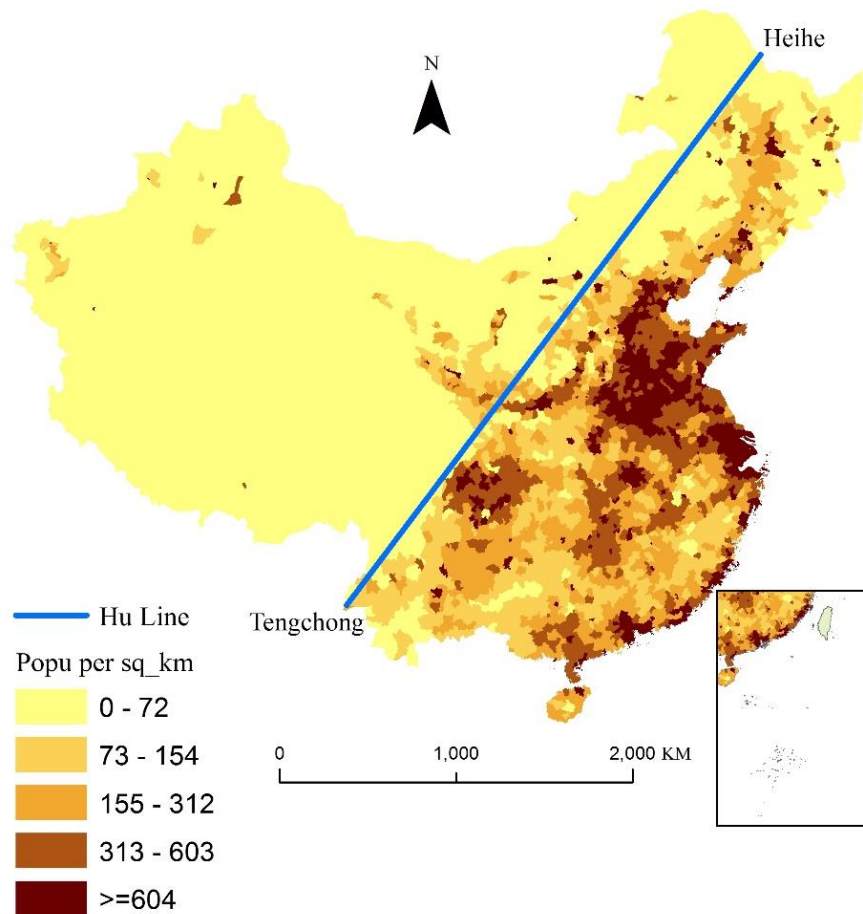
Figure 1. Hu Line superimposed on the population density pattern of China in 2010

More recently in the era of the reform and open door policy, the eastern coastal areas in China have benefited more from influx of major foreign investment, better physical environment, and more health care and education facilities, and generated more economic growth and attracted more population settlement, especially in the ever-expanding urban areas (Lu et al., 2005; He et al., 2016). In contrast, economic growth and population (permanent residents) have been staggering in the west of China. Uneven development on the two sides further widens the

disparities between them. The difference in population density elevates the public's concerns of crowded housing, environmental pollution, public resource shortage, traffic congestion, resource depletion, and other problems in high-density areas (The World bank, 2007; Yu et al., 2012; WHO, 2013; Chen et al., 2013). Understandably, the leadership in China was tempted to look into the plausibility of overcoming the invisible barrier of Hu Line and contemplating the possibility of balanced development for the whole China (Guo et al., 2016). It has generated some hotly-contested debate (Gao et al., 2016; Lu et al., 2016). Some support massive population relocation from the southeast to northwest by promoting more development in the northwest, especially after Chinese President Xi launched the Silk Road Economic Belt and the 21st-Century Maritime Silk Road strategy, better known as the One Belt and One Road Initiative (OBOR). Others, especially some geographers, suspect that any large-scale population relocation is limited by major environmental challenges in the northwest regions (Lu et al., 2016).

The Hu Line was proposed largely by visual examination of hand-made maps and arduous experiments of numerous configurations, and has been subject to criticism of lack of scientific rigor and accuracy. In the digital computational era, a school of Geographic Information Systems (GIS)-automated regionalization methods have recently been developed to delineate regions by grouping areas of similar attributes (e.g., population density) while attaining some properties in derived regions (e.g., a desirable number of regions, and each with a certain threshold value of, say area size) (Wang, 2015: 193-215). This technique, when carefully crafted, affords us an opportunity to validate the Hu Line scientifically. Furthermore, it has been over eight decades since the Hu Line was put forward. During the time, China sustained several major man-made and natural disasters (e.g., the World War II, the subsequent Civil War and the 1958-62 Great Famine), and also experienced some major government-sponsored migrations and aforementioned recent

economic growth and unprecedented urbanization. It is necessary to revisit the (in) stability of Hu Line over time. There is also value of detecting population variability across more regions than the Line's dichotomous division.

Specifically, this dissertation examine several scientific aspects surrounding the debate by (1) advancing the methodological rigor in defining the Hu Line in the first place, (2) solidifying the inherent connection between physical environment and population settlement, and (3) strengthening the findings by extensions to analyses across time.

The dissertation uses a three-paper format. Chapter 2 validates and refines the Hu Line by the rigorous scientific method of regionalization, REDCAP. It also extends to explore delineation of multiple regions. Chapter 3 seeks to uncover the underlying physical environment factors that shape such a contrast by proposing a Habitation Environment Suitability index (HESI) and then compare the HESI-based delineation line with the population density demarcation line. Chapter 4 analyzes the population density patterns and distribution disparity changes from the year 1953 to 2010. Chapter 5 concludes the dissertation with major findings of the study and discussed possible future work.

CHAPTER 2. ANALYZING THE POPULATION DENSITY PATTERNS IN CHINA WITH GIS-AUTOMATED REGIONALIZATION METHOD: THE HU LINE REVISITED

2.1 INTRODUCTION

As the world's most populous country, China has experienced formidable challenges related to its large population over the past 100 years. Adding to the challenges is the massive disparity of population density in the country. One of the best known studies on the issue was the famous "Hu Huanyong Line" (simply referred to as "Hu Line" thereafter), named after its discoverer (Hu, 1935). It is considered one of the greatest geographic discoveries of China (Shan, 2009). The Hu Line begins from Heihe (or Aihui) in Heilongjiang Province in the northeast to Tengchong in Yunnan Province in the southwest of China, as shown in Figure 1. It divides China into two parts of similar area sizes but a striking contrast in population: the northwest side with only about 4% of the country's total population and the southeast side with nearly 96% of the population. An early report of Hu's work on population distribution in China in the mainstream English outlet can be found in Alexander (1948). Considering the population growth and border changes over the years (e.g., excluding Taiwan and Mongolia), Hu (1990) made some adjustment of the line based on the 1982 Census of China: the southeast part with 94.4% population in 42.9% land, and the northwest part with 5.6% population lived in 57.1% land. The stunning difference in population density remained on the two sides of the Hu Line.

It has been over eight decades since the original Hu Line was proposed. During the time, China sustained several major man-made and natural disasters (e.g., the World War II, the subsequent Civil War and the 1958-62 Great Famine), and conducted some major government-sponsored migrations (1950s-70s). Since Hu's revision based on the 1982 data, China has experienced tremendous economic growth and unprecedented urbanization in the era of the reform

and open door policy. The eastern coastal areas in China have benefited more from influx of major foreign investment, favorable physical environments, and better infrastructures, and generated more economic growth and attracted more population settlement, especially in the ever-expanding urban areas (Lu et al., 2005; He et al., 2016). In contrast, economic growth and population have been stagnant in the west of China. Uneven development on the two sides may have further widened the disparities between them, and elevated the public's concerns of crowded housing, environmental pollution, public resource shortage, traffic congestion, resource depletion, and other problems in high-density areas (The World bank, 2007; Yu et al., 2012; WHO, 2013; Chen et al., 2013). Understandably, the leadership in China was tempted to look into the plausibility of overcoming the invisible barrier of Hu Line and contemplating the possibility of more balanced development for the whole China (Guo et al., 2016). It has generated some hotly-contested debates (Gao et al., 2016; Lu et al., 2016). Some support massive population relocation from the southeast to northwest by promoting more development in the northwest, especially after Chinese President Xi launched the Silk Road Economic Belt and the 21st-Century Maritime Silk Road strategy, better known as the One Belt and One Road Initiative (OBOR). Others, especially some geographers, suspect that any large-scale population relocation is limited by major environmental challenges in the northwest regions (Lu et al., 2016). It is necessary to revisit the (in) stability of Hu Line based on the most recent census in 2010. There is also value of detecting population variability across more regions than the Hu Line's dichotomous division.

Moreover, the classic Hu Line was derived by visual examination of hand-made maps and arduous experiments of numerous configurations, and thus has been subject to criticism of lack of scientific rigor and accuracy. Most of related studies were on either differences between the two sides or possible reasons behind the divisions (e.g., Yuan, 1993; Shan, 2009; Qi et al., 2015; Wang

and Pan, 2016; Wang and Deng, 2016). Seldom are researches on how to verify it, or whether one can find a better line. In the digital computational era, a school of Geographic Information Systems (GIS)-automated regionalization methods have recently been developed to delineate regions by grouping areas of similar attributes (e.g., population density) while attaining some properties in derived regions (e.g., a desirable number of regions, and each with a certain threshold value of, say area size) (Wang, 2015: 193-215). This technique, when carefully crafted, affords us an opportunity to validate the Hu Line scientifically.

The remainder of the chapter is organized as follows. The next section discusses the GIS-automated regionalization technique with a focus on the REDCAP method. Section 2.3 provides a brief description of the study area and data. Sections 2.4 and 2.5 use the method to divide China into two regions with comparable area size while maximizing the homogeneity within the derived regions, with the former measuring the attribute homogeneity in population density and the latter in the logarithm of population density. The paper is concluded with a brief summary.

2.2 METHOD

Regionalization is to divide a large set of spatial objects into a number of spatially contiguous regions while optimizing an objective function, which is normally a homogeneity (or heterogeneity) measure of the derived regions (Duque et al., 2007). There have been a rich set of automated regionalization methods in the literature. A popular approach is the agglomerative hierarchical methods (AHM), which calculate a distance (dissimilarity) matrix between objects, and then apply some agglomerative criterion for grouping those objects (Everitt et al., 2001). More recently, Guo (2008) followed the AHM approach and developed a family of regionalization methods, termed "regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)." REDCAP groups contiguous areas of similar attribute values to obtain a set of

7

regions while explicitly optimizing an overall homogeneity measure. The method has several advantages over existing regionalization methods such as spatial compactness, attribute homogeneity, and scale flexibility within derived regions. An improved version of the REDCAP further accommodated some desirable properties such as ensuring constructed regions above certain threshold sizes (Guo and Wang 2011). These features of REDCAP have made it a popular choice in various applications (e.g., Kupfer et al., 2012; Wang et al., 2012; Xu et al. 2014; Jin et al. 2015; Boluwade et al. 2016; Wang et al. 2018). This study also uses the REDCAP method to delineate regions.

Specifically, REDCAP extends four commonly used hierarchical clustering methods (the single-linkage (SLK) clustering, average-linkage (ALK) clustering, complete-linkage (CLK) clustering, and the Ward hierarchical clustering methods), and designs two different strategies to incorporate contiguity constraints in hierarchical clustering: the first-order constraining and the full-order constraining. As suggested by Guo (2008), in terms of the overall heterogeneity, Full-Order-ALK is a preferred choice for generating fewer than 12 regions, and thus is chosen for this research. Two polygons are considered contiguous in space if they share a segment of boundary (i.e., rook contiguity). (Dis)similarity between two neighboring areas i and j is measured by their attribute distance D_{ij} such as

$$D_{ij} = \sum_t (x_{it} - x_{jt})^2 \qquad (2\text{-}1)$$

where objects i and j have t-th ($t = 1, 2, \ldots, T$) attributes standardized as ($x_{i1}, \ldots, x_{iT}$) and ($x_{j1}, \ldots, x_{jT}$), respectively. In our case, simply $T = 1$ as the only attribute variable is population density. Therefore, the quality of a regionalization result is evaluated based on the overall heterogeneity, measured by the total sum of squared deviations (SSD) (Everitt, 2002), such as

$$SSD = \sum_{r=1}^{k} \sum_{i=1}^{nr} \sum_{t=1}^{T} \left(x_{it} - \overline{x_t} \right)^2 \qquad (2\text{-}2)$$

where, k is the number of regions; n_r is the number of small areas in region r; x_{it} is a variable value; and $\overline{x_t}$ is the regional mean for variable t. Note that x_{it} is the *normalized* value of the attributive variable. The regionalization algorithms seek to minimize the SSD (or maximize the homogeneity within regions). A lower value of SSD indicates more homogeneous regions and thus is preferred.

The regionalization procedure is composed of two steps. The first step is a bottom-up contiguity-constrained hierarchical clustering process, detailed as:

(1) Compute attribute similarity between each object and its adjacent object (i.e., "contiguity-constrained"), according to Equ.(2-1);

(2) Group two adjacent and most similar areas to form the first cluster, and mark a link/edge between them;

(3) Group two adjacent and most similar clusters to form a higher level cluster, and continue until the whole study area is one cluster; and

(4) Generate a clustering tree to fully represent the cluster hierarchy.

The second step is a top–down tree partitioning process, such as

(1) Remove the best edge to create two regions that optimizes the homogeneity according to Equ.(2-2);

(2) Continue the partitioning until the desired number of regions is reached and the threshold population in each region is met.

The REDCAP method is suitable for our task as it generates a given number of regions (e.g., two) for a specific objective (e.g., minimum difference in population density within derived

9

regions) with a certain constraint on the threshold population in regions (e.g., comparable area sizes) .

In addition to attributive homogeneity as defined in equ. (2-2), the results of regionalization can be assessed by spatial compactness in the derived regions. *Isoperimeter quotient* (IPQ) is a common spatial compactness index, and is defined as the ratio of a region's actual area size (A) over the area of a circle having the same perimeter (P) of the region (e.g., Wang and Robert 2015), simplified as

$$IPQ = 4\pi A/P^2. \qquad (2\text{-}3)$$

A higher average IPQ indicates more compact regions in shape and thus is preferred.

2.3 STUDY AREA AND DATA

The study area is mainland China. Due to the lack of data, it does not include the Hong Kong special administrative region, the Macao special administrative region, Taiwan and South China Sea islands (Figure 2). The county-level population data are based on the China 2010 Census. The study area had 2341 county-level administrative units (hereafter simply referred to as "county"), total population of 1.327 billion and total area of about 9.45 million km^2 in 2010. Figure 2 shows that population density is high in the east, and sparse in the west.

In order to establish the baseline, the Hu Line is constructed by connecting the centroids of Heihe in Heilongjiang Province and Tengchong in Yunnan Province, and then extending to the borders. The Hu Line was originally called Aihui-Tengchong Line, and later renamed Heihe-Tengchong Line after Aihui was renamed to Heihe. As shown in Figure 2, counties along the Hu Line are split. To improve the accuracy for estimates of population and areas in the two regions, we use the areal weighting interpolator to estimate their population of these counties in each side of the line proportionally to the corresponding area sizes.
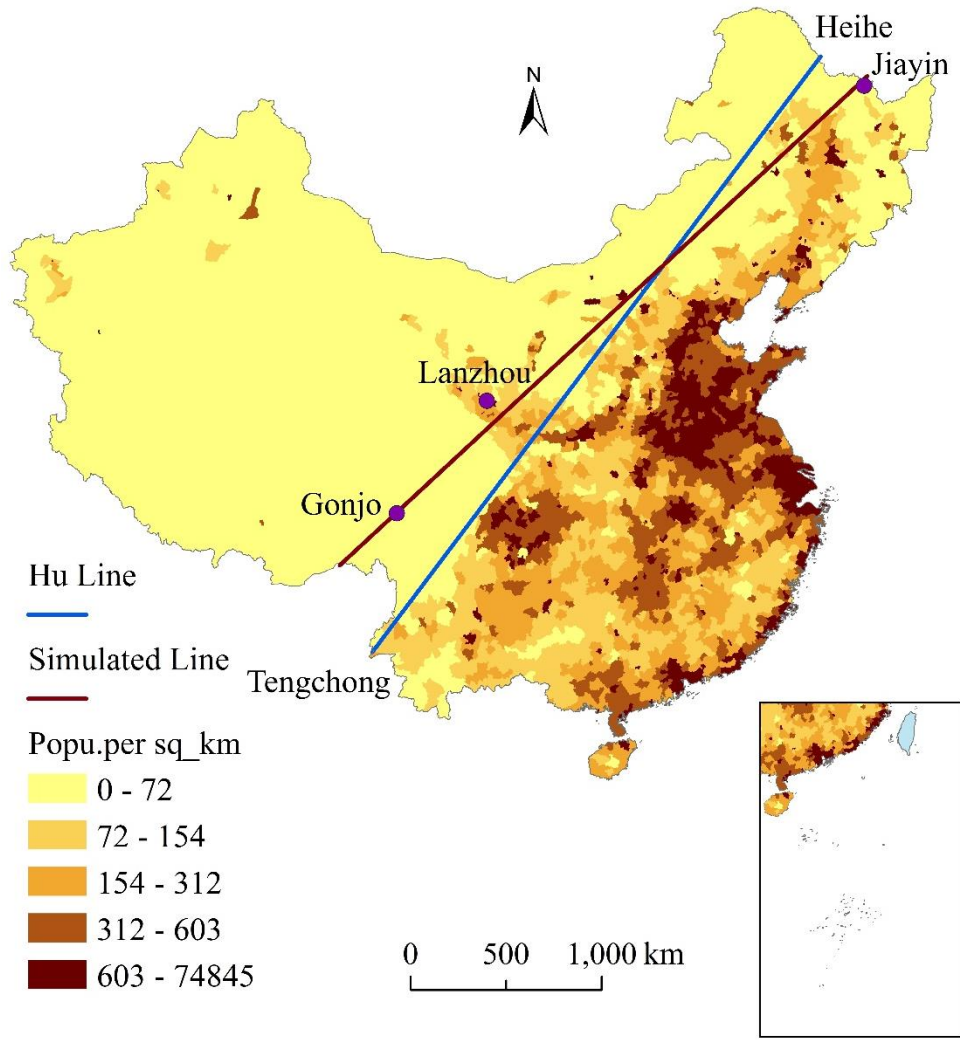
Figure 2. Population distribution, Hu Line and simulated line in China 2010

Table 1. Basic statistics for 2 regions divided by the Hu Line and simulated line

Demarcation Line	Variable	Southeast Region	Northwest Region	Ratio (SE/NW)
Hu Line	Population (mil) (%)	1248.6 (94.1%)	78.7 (5.9%)	15.86
	Area (mil km^2) (%)	4.17 (44.1%)	5.28 (55.9%)	0.79
	Population density (p/km^2)	299.3	14.9	20.06
Simulated Straight Line	Population (mil) (%)	1258.2 (94.8%)	69.2 (5.2%)	18.19
	Area (mil km^2) (%)	4.46 (47.1%)	5.00 (52.9%)	0.89
	Population density (p/km^2)	282.4	13.8	20.40

11

As reported in Table 1, the southeast region has 94.1% population in 44.1% area (and thus the northwest region has 5.9% population in 55.9% area) in 2010. This is very similar to the result reported by Hu's (1990) study using the 1982 Census: the southeast accounted for 94.4% population in 42.9% area. The minor discrepancy in area sizes in the two studies may be attributable to lack of GIS data for accurate measure of areas and the aforementioned estimation of areas in split counties in Hu's 1990 study. In other words, the population disparity between the two sides has remained astonishingly stable[1].

2.4 DELINEATING TWO REGIONS BASED ON POPULATION DENSITY

Several issues merit discussion in implementing the REDCAP method for regionalization in our case study. First, as stated previously, an important property of REDCAP is its flexibility in the number of regions generated. As our first objective is to validate the Hu Line, this study focuses on deriving two regions. In other words, the task of regionalization is to cluster the counties into two groups.

How do we define the attribute homogeneity for regionalization? As the purpose of Hu Line was to demonstrate the disparity in population distribution between the two sides divided by it, we began with setting "population density" as the variable for defining attribute homogeneity in the clustering process. Furthermore, REDCAP can also generate regions of desirable size. Recall that the two regions divided by the Hu Line have comparable area sizes (44% vs. 56%). In order to have two regions with "comparable area proportions," we experimented with various threshold area sizes within the range of 35%-49% at an interval of 1%. It is unlikely to derive two regions

---

[1] Note that the Census data in China reports population according to their registered residences (i.e., "hukou"). Conceivably, significantly more migrant workers with their registered residences in the northwest region actually lived in the southeast region in 2010 than in 1982, so the disparity according to actual residences between the two sides has probably become even higher. That is beyond of the scope of this work.

of the same area size, and thus no attempt was made to set a threshold at 50% of total area. Related statistics are reported in Table 2.

Table 2. Basic statistics for 2 REDCAP-derived regions (based on population density)

Area threshold in % of total area	Variable	Southeast Region (SE)	Northwest Region (NW)	Ratio (SE/NW)
	Population (mil) (%)	7.0 (0.5%)	1320.4 (99.5%)	0.005
	Area (mil km^2) (%)	0.0002 (0.002%)	9.45 (99.998%)	0.00002
(1) None	Population density (p/km^2)	39911.4	139.7	285.69
	Sum of squared deviations (SSD)		1273.1	
	Isoperimeter quotient (IPQ)		0.2691	
	Population (mil) (%)	1212.8 (91.4%)	114.6 (8.6%)	10.59
	Area (mil km^2) (%)	3.62 (38.3%)	5.83 (61.7%)	0.62
(2) 35%-38%	Population density (p/km^2)	335.3	19.6	17.08
	Sum of squared deviations (SSD)		2297.7	
	Isoperimeter quotient (IPQ)		0.0615	
	Population (mil) (%)	1214.2 (91.5%)	113.2 (8.5%)	10.72
	Area (mil km^2) (%)	3.71 (39.2%)	5.74 (60.8%)	0.65
(3) 39%	Population density (p/km^2)	327.4	19.7	16.61
	Sum of squared deviations (SSD)		2298.2	
	Isoperimeter quotient (IPQ)		0.0634	
	Population (mil) (%)	1287.9 (97.0%)	39.46 (3.0%)	32.64
	Area (mil km^2) (%)	5.16 (54.6%)	4.29 (45.4%)	1.20
(4) 40-45%	Population density (p/km^2)	249.4	9.2	27.10
	Sum of squared deviations (SSD)		2304.6	
	Isoperimeter quotient (IPQ)		0.1416	

(1) As a baseline, when no threshold area size was defined, the resulting two regions included a very small area for a minuscule 0.002% of total area (composed of the central nine districts of Shanghai). Its extent can be seen in Figure 6.

(2) For threshold area sizes in the range of 35%-38% of the total area, the results were the same. As shown in Figure 3 (line a), the derived southeast region was smaller than the northwest region and accounted for 38.3% of the total area.

(3) When the threshold area size was set at 39% of the total area, the result was very similar to the above scenario and thus its map is not shown. The derived southeast region was also smaller than the northwest region, and accounted for 39.2% of the total area.

(4) For threshold area sizes in the range of 40%-45% of the total area, all scenarios yielded the same result[2]. As shown in Figure 3 (line b), in contrast, the derived southeast region was larger than the northwest region, and accounted for 54.6% of the total area.

(5) When threshold area sizes were set in the range of 46%-49% of the total area, it is infeasible for the REDCAP to generate two regions. In other words, no links in step 2 of REDCAP could be removed to yield two regions with areas above the threshold.

---

[2] In fact, as shown in Table 2, the resulting smaller region (northwest) had an area size 45.37% of the total area. In other words, one might increase the threshold area up to 45.37% and obtain the same result.
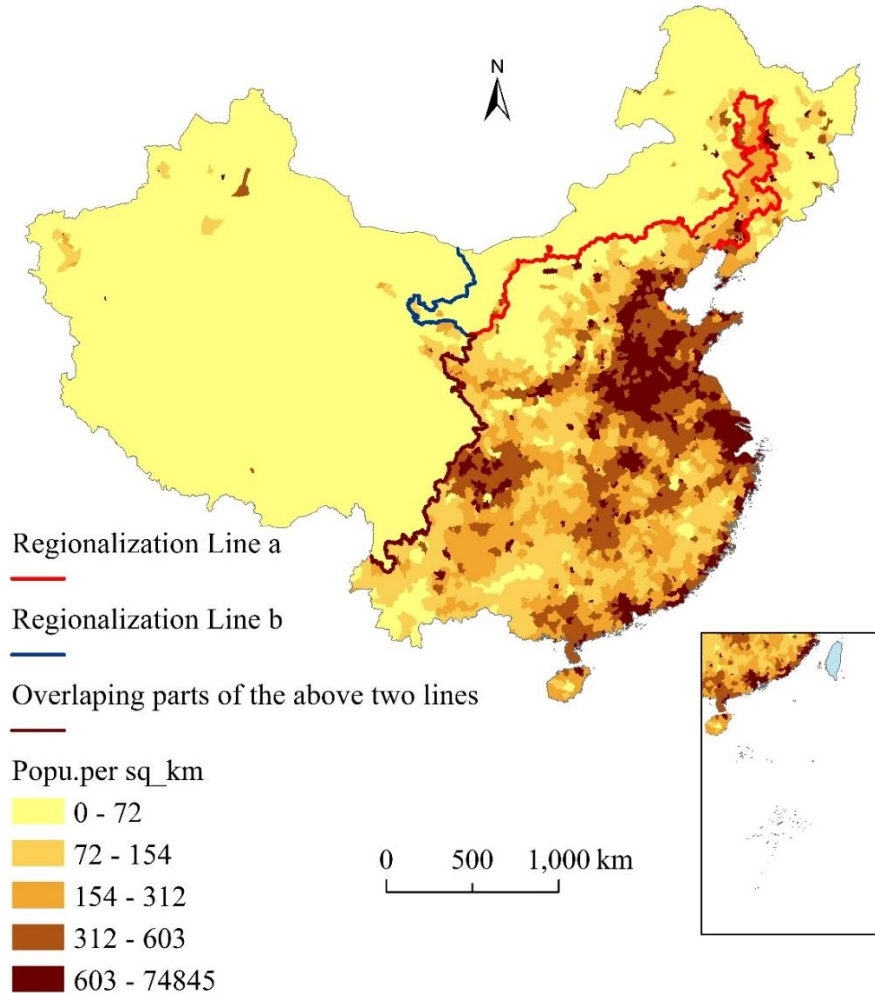
Figure 3. Population-density based regionalization (two regions) in China 2010

Since the population densities in urbanized areas of Shanghai were so high (average of about 40000 p/km^2 and about 284 times the national average), the algorithm yielded a very small region there and the rest forming another region. Therefore, when no minimum area was imposed for regions, obviously the REDCAP yielded the smallest SSD value 1273.1 (Table 2). When the threshold area was raised, the SSD value increased and it came at the cost of less homogeneity. However, as stated previously, the purpose of this study was to generate two regions with comparable area sizes. For this reason, the 1st scenario is not considered desirable.

Scenarios (2) and (3) yielded similar results as the southeast region remained smaller than the northwest. When the threshold area size was raised to 40%, scenario (4) yielded a very different result as the southeast region expanded so much that it outsized the northwest region. Despite a slightly higher SSD value, at least three reasons make scenario (4) a preferable choice over the other two:

(1) As shown in Figure 3, regions in scenario (4) are more compact in shape than scenarios (2) or (3). This is evident a larger average IPQ value in scenario (4), and a much more zigzag delineation line in the other two scenarios (particularly the northeast area).

(2) The two regions generated in scenario (4) are more comparable or balanced in area size (54.6% vs. 45.4%) than the other two.

(3) The contrast in population density between the two derived regions is stronger in scenario (4) (a ratio of 27) than the other two (ratios of about 17).

However, the above regionalization experiments based on population density raised some concerns. The homogeneity measure, SSD, for the study area prior to the regionalization was 2315, and its reduction was small and ranged from 10.4 to 17.3 (i.e., 0.45-0.75%) in scenarios (2)-(4). While the optimal outcome of regionalization without any constraint in threshold area size had the highest reduction in SSD by 1041.9 (or 45%) to 1273.1, the resulting regions were massively-unbalanced in area size. Indeed, because of this scenario, we looked for measures of attributive homogeneity beyond population density.

2.5 DELINEATING TWO REGIONS BASED ON LOGARITHM OF POPULATION DENSITY

Empirical studies on regional population density patterns including China (e.g., Wang 2001) suggest that population densities usually decline exponentially with distances from cities. Therefore, the disparity in density between urban and rural area is by several orders of magnitude.

This leads to the calibration of SSD dominated by observations in urban areas, especially those cities with the highest density such as Shanghai. The aforementioned scenario (1) illustrated the shortfall of using population density to measure attributive homogeneity in regionalization. This section explores regionalization based on the logarithmic transform of population density.

Recall that SSD was based on the normalized values of the attributive variable. As shown in Figure 4a, the normalized values of population density were highly skewed (more than 85% counties had their normalized values below the mean 0). On the contrary, the normalized values of logarithm of population density closely resembled a normal distribution (Figure 4b). This lends another strong support to the proposition that population density in logarithm is a more suitable measure of spatial variability of population distribution than merely population density itself. Note that population density breaks for legends used geometrical interval in Figure 2 (the same in Figures 3 and 5).

(a)   (b)
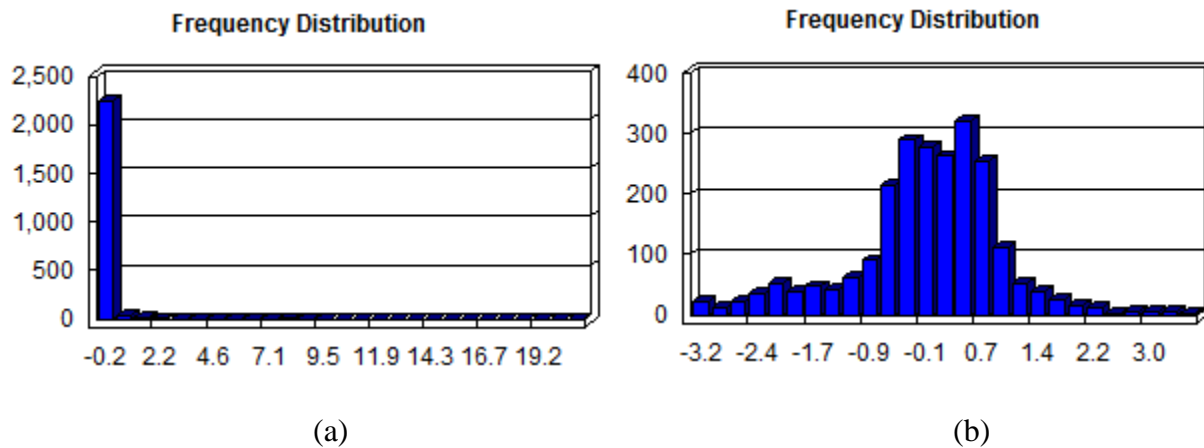
Figure 4. Distribution of normalized attribute values of: (a) population density, (b) logarithm of population density

In implementation, the logarithm of density was used to replace density as the variable for measuring attributive homogeneity in REDCAP. Similarly, we began the analysis by imposing no threshold constraint for region size (in area). Experiments with any other threshold area sizes led

17

the same result. The delineation line is shown in Figure 5, and related statistics is reported in Table 3.
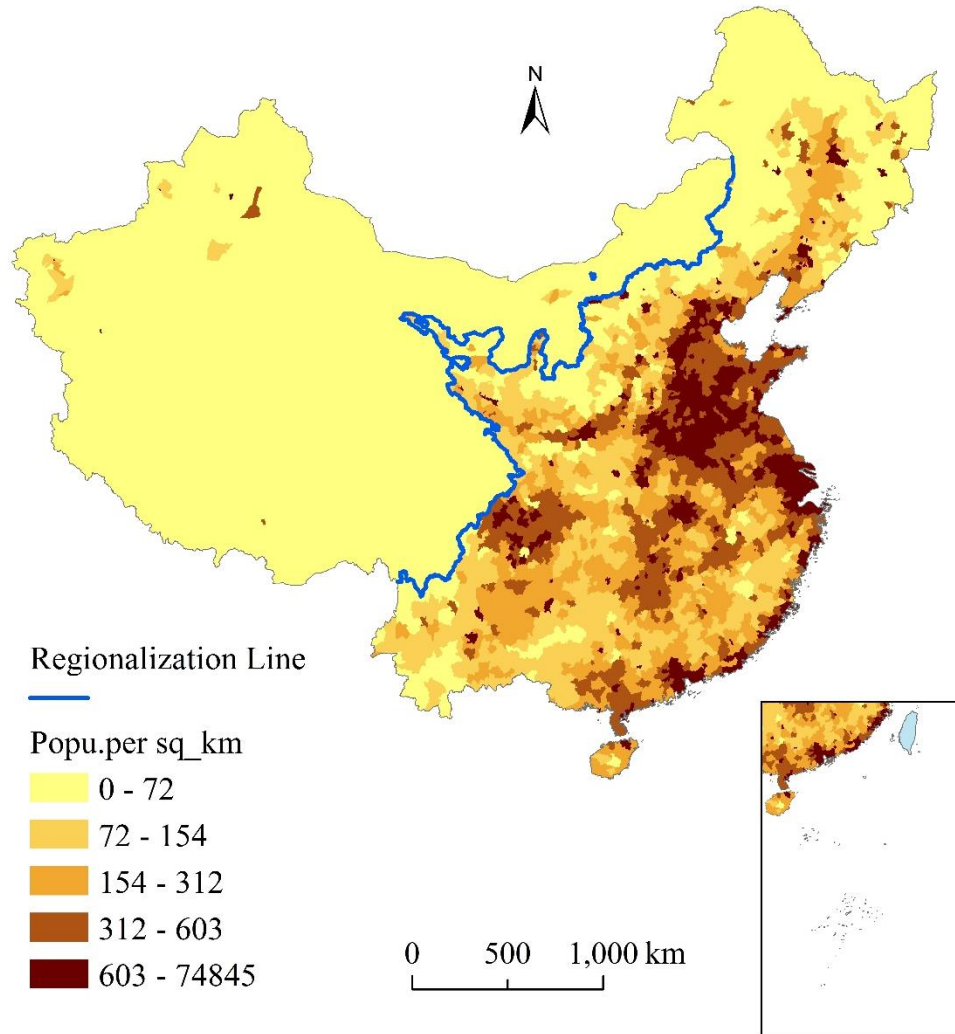
Figure 5. Population-density-logarithm based regionalization (two regions) in China 2010

Table 3. Basic statistics for 2 REDCAP-derived regions (based on logarithm of population density)

Area threshold in % of total area	Variable	Southeast Region (SE)	Northwest Region (NW)	Ratio (SE/NW)
None	Population (mil) (%)	1292.9 (97.4%)	34.5 (2.6%)	37.48
	Area (mil km^2) (%)	4.80 (50.8%)	4.65 (49.2%)	1.03
	Population density (p/km^2)	269.4	7.4	36.3
	Sum of squared deviations (SSD)		1338.6	
	Isoperimeter quotient (IPQ)		0.0925	

18

A close examination of the result revealed that the two derived regions were very close in area size (50.8% in southeast region vs. 49.2% in northwest region). It is helpful to revisit step 2 of REDCAP as the constraint is enforced. Specifically, this step, termed "tree partitioning", begins to generate two regions by removing the best edge that minimizes the SSD. When a constraint such a minimum area size for derived regions is defined, a potential cut (i.e., removal of an edge) would not be considered if its produced regions do not meet this constraint. The algorithm moves on to the next best cut, and so on until a feasible cut is made (here, both generated regions with their area sizes above the threshold). If there is no candidate cut, the study area will not be partitioned further (as it was the case in Scenario (5) in the previous section). In other words, any threshold area below 49.2% would not take effect as the optimal solution without any constraint already yielded two regions equal or larger than 49.2%.

That is to say, when based on logarithm of population density, the two regions derived by REDCAP were "natural regions" that enjoyed the optimal within-region homogeneity. The SSD value (1338.6, representing 42% reduction from the SSD=2315 prior to regionalization) (Table 3) was also much smaller than any of the realistic partition scenarios (2)-(4) based on population density without any constraint (2297.7-2304.6) (Table 2). It was even more stunning that this optimal partition yielded two regions of near-identical area size. The density contrast on the two sides was also sharper (36.3 times, Table 3) than density-based scenarios (2)-(4) (16.6-27.1 times, Table 2). In summary, China is naturally divided in its settlement pattern China with the southeast side commanding more than 97% population in about the same area as the northwest side. The only less desirable property in this logarithmic-density-based result was a smaller IPQ value (0.0925) than that in density-based scenario (4) (0.1416), and thus less compact regions.

As discussed above, the logarithmic-density-based delineation line marks a natural division with several favorable properties over the Hu Line: (1) foremost, derived scientifically with a clear objective of maximizing regional homogeneity, (2) resulting in two regions of near-equal area sizes, and (3) the greatest disparity in population density between the regions. However, one may consider it desirable to draw a straight line like the Hu Line so comparison can be made. For this purpose, we developed the following procedure to simulate such a straight line:

(1) extract the turning points in the REDCAP-derived line;

(2) add the X and Y coordinates for the turning points;

(3) run a regression line $y = a+bx$, where x and y are the X and Y coordinates; and

(4) generate the line in GIS as the simulated straight line.

See Figure 2 for the simulated line.

In comparison to the Hu Line, it is more tilted southeastward by rotating the Hu Line about 10^0 clockwise (or linking Jiayin of Heilongjiang in the northeast and Gonjo of Tibet in the southwest). Such an adjustment enables the inclusion of some populous settlement area around Lanzhou (i.e., Longzhong Plateau) in the southeast region. As reported in Table 1, the simulated line expands the southeast region from 44.1% to 47.1% of the total area and added 0.7% of total population. It is slightly favored over the Hu Line as the two regions are more comparable in area size with a bit stronger contrast in density (20.40 times vs. 20.06). It may be summarized as "the southeast side of the simulated line accounts for about 95% population in only 47% area of China." Understandably, straight lines are more limited to account for the variability of population density, and both create regions with less disparity in population density on the two sides than the REDCAP line shown in Figure 5.

A major feature of the REDCAP method is its scale flexibility. In other words, one may generate any number of regions up to the total number of counties in this case. This section explores the scenarios up to eight regions, as an example, to further advance our understanding of population density patterns in China. While eight seems to be an arbitrary choice, it has been a popular number for delineating regions by the central government in various stages of development, most recently adopted by the State Council of China (2008).

Here the eight regions are derived in terms of homogeneity of population density, as explained in the previous Section 2.5, defined by its logarithm. As shown in Figure 6, the order of regionalization indicates how more regions are gradually created by carving one region at a time from a previous round. For instance, regions 1 and 2 are first defined, then region 3 is carved out of region 1, and so on. Note that region 5 is composed of only a small area of highly-urbanized central Shanghai, identical to scenario (1) in Section 4. According to Table 4, these eight regions may be classified as six types: from the low density west (7.4 p/km^2), to medium-low density northeast periphery (37.6 p/km^2), medium density inland (156.6 p/km^2), 3 medium-high density regions (North China Plain, Sichuan Basin and south Guangdong) (507.8-566.7 p/km^2), high density southeast coastal (979.2 p/km^2), and very high density urban Shanghai (39911.4 p/km^2). These eight regions coincide well with natural geographic divisions predominately shaped by physical environment.
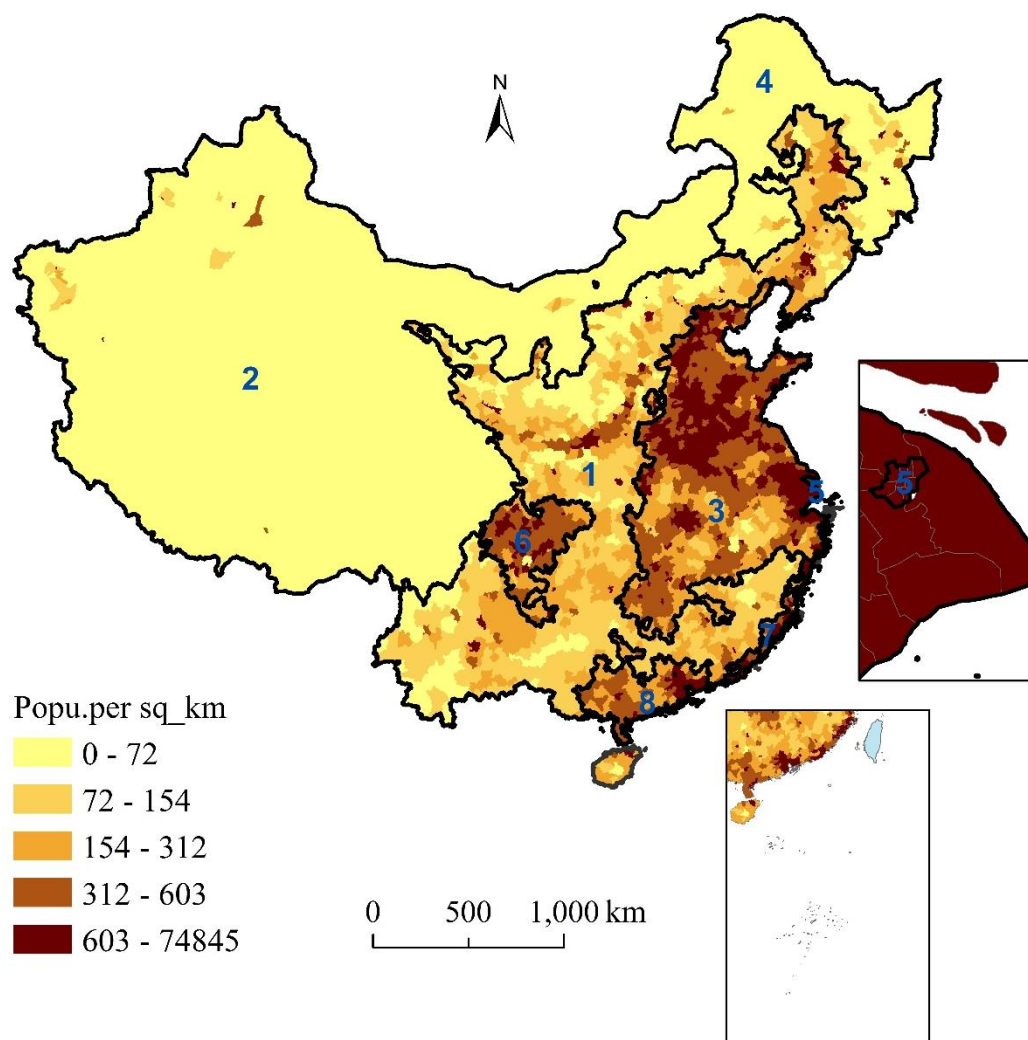
Figure 6. Population-density-logarithm based regionalization (8 regions) in China 2010

Table 4. Basic statistics for 8 REDCAP-derived regions (based on logarithm of population density)

Order of regionalization	Population (mil)	Area (mil km^2)	Population density (p/km^2)	SSD	Note
1	377.2	2.41	156.6	2315.0	Medium density inland
2	34.5	4.65	7.4	1338.6	Low density west
3	619.2	1.09	566.7	1052.4	Medium-high density N China Plain
4	32.1	0.85	37.6	977.2	Medium-low density NE periphery
5	7.0	0.00	39911.4	918.0	Very high density urban Shanghai
6	98.0	0.19	507.8	867.3	Medium-high density Sichuan Basin
7	48.4	0.05	979.2	826.0	High-density SE coastal
8	111.0	0.20	550.9	789.4	Medium-high density S Guangdong

Figure 7 shows how the SSD value declines with an increasing number of regions. Clearly the steepest drop is the delineation of two regions (42.2%), followed by three regions (12.4%), and the remaining trend line levels off. This exploratory analysis once again supports the duality in China's population density pattern, and validates the wisdom of Hu Line (1935) as an attempt to highlight the major geographic division in China.
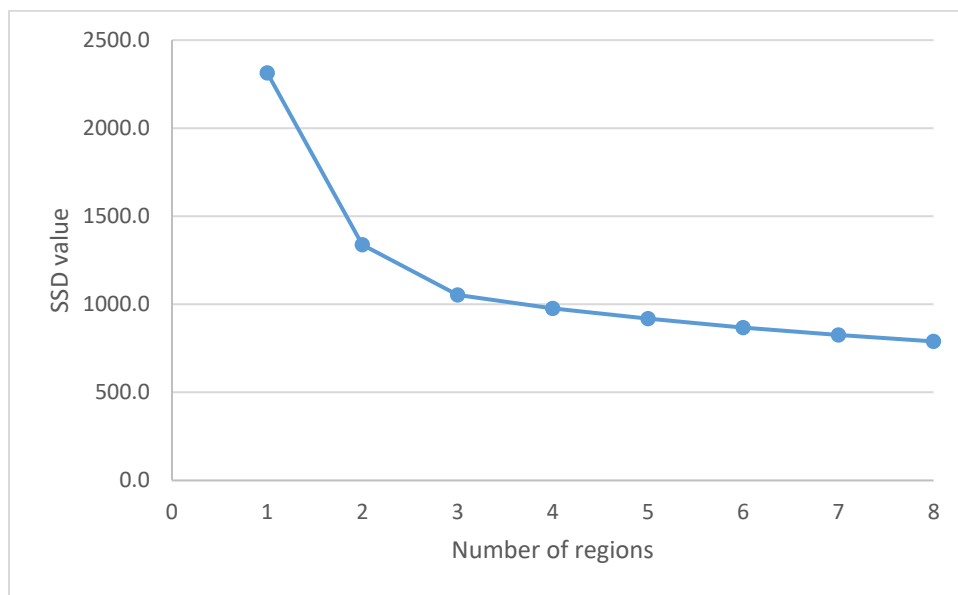
Figure 7. SSD values in various regionalization scenarios

2.7 CONCLUSION

Researchers have long been fascinated by the Hu Line that separates China into two parts of similar area but stunningly different population sizes. This research revisits the Hu Line by employing a scientific method of regionalization, termed "REDCAP", to derive two regions with desirable properties. The method generates regions that maximize the homogeneity within them. Our attempt of regionalization by defining attributive homogeneity as population density has only yielded moderate success. By imposing a threshold area size for derived regions, the density-based

method generates two regions that are comparable and yet still significant different area sizes (about 10% difference). The best result is produced by using the logarithmic transformation of population density to define attribute homogeneity. Even without any input of a threshold area size for regions, the method delineates two regions that are nearly-identical in area size and with stronger contrast in population density. It suggests that derived regions reflect truly the natural division of China. Further exploratory analysis of delineating larger numbers of regions validates the duality in China's population density. For those who desire a straight demarcation line, this research proposes a regression method to simulate a line. In comparison to the Hu Line, the simulated line is tilted about 10^0 more clockwise, and divides China into two regions that are more comparable in area size with slightly stronger contrast in density.

Future work will be extended in several directions. This paper is based on the 2010 population census data. Analysis of historical population data over a period of time helps us examine the (in)stability of the delineation line. Application of the method in different countries (or regions) may reveal similar or other interesting settlement patterns. Perhaps the most valuable and also challenging task is to identify underlying forces that shape this division, and examine the consistencies and inconsistences between them. The latter (i.e., inconsistences) may hold the key to possibility of overcoming the barrier – an elusive goal that has enticed generations of policy makers and researchers in China.

CHAPTER 3. HABITATION ENVIRONMENT SUITABILITY AND POPULATION DENSITY PATTERNS IN CHINA: A REGIONALIZATION APPROACH

3.1 INTRODUCTION

Population has profound impacts on natural resources, the environment and our economy (Sutton et al., 1997). In turn, these factors also affect the population distribution. As the world's most populous country, China has experienced formidable challenges related to its large population over the past 100 years. Adding the challenges is its uneven distribution. One of the best-known studies on the topic is Hu (1935, 1990) by proposing the famous "Hu Huanyong Line" (simply referred to as "*Hu Line*" thereafter). The Hu Line begins from Heihe in Heilongjiang Province in the northeast to Tengchong in Yunnan Province in the southwest of China (Figure 1), and divides China into two parts: the northwest side has 56% of total area but only about 4% of total population, and the southeast side has 44% of total area but nearly 96% of total population. Figure 1 shows the Hu Line on the background of population density pattern in 2010. According to the 2010 Census of China, the population proportion remains about 94% in the southeast of the Line, and about 6% in the northwest (Qi et al., 2015).

With an increasing demand of balanced development for the whole China, the leadership in China has been contemplating the possibility of overcoming the invisible barrier of the Hu Line (Guo et al., 2016; Lu et al., 2016). Examining the feasibility of such a proposition is a valuable scientific inquiry. Researchers have long sought to find reasons for the variability of population distribution (Stutz and Warf, 2007; Kumar, 2015). The factors affecting population distribution at the global scale consist of physical and human factors. Several important physical factors include elevation, relief, climate, water supply, soil, and natural resources, etc. Elevation has a significant influence on population distribution. High elevation in general imposes a physiological limit upon human existence due to reduced atmospheric pressure and low oxygen content. High elevation is

also unfavorable to crop growth because of the low temperature. Staszewski (1967) found that both the number and density of population decline with increasing altitude across the world. Very few permanent settlements can be seen at a height above 5,200 meters (Cohen and Small, 1998). Topographical relief refers to the vertical elevation change within a given area, and reflects regional terrain. Flatter, alluvial plains tend to have better farming soils than steeper, rockier uplands. Rugged and undulating topography restricts the condensation of human population in any area. Temperatures that are too high cause people discomfort, and temperatures that are too low make a short growing season, so temperature is an important factor influencing the population settlement. Water is a basic necessity for many purposes such as irrigation, industries, transport and domestic affairs. The quality of soils also exerts a conspicuous influence on the distribution of world population. Areas with good quality soil, such as the fertile alluvial plain and deltas, can support dense population. In addition, areas with a wealth of natural resources such as oil, coal or minerals may also support higher population densities than areas without. Human factors mainly include economic, political, historical, and cultural, etc. The impact of economic conditions on population distribution is much straightforward. Better economy can attract population concentration for more job opportunities, better public services, and more convenient transportation. Population density in an area depends to a large extent on the type and scale of economic activities. Policies and political factors are also an important determinant of population patterns in many parts of the world. Duration of human settlements often correlates with the magnitude of population in many regions. Most of the densely populated areas of the world have a very long history of human habitation.

However, the most important factors influencing the population distribution on a macro scale in China are physical environment and land productivity (Yuan, 1993), and land productivity

is also largely determined by physical environment. As summarized by Gao et al. (1999), the important natural factors related to human settlement are hydrothermal conditions. Similarly, Wang and Na (2008) indicated that temperature, precipitation and sunlight conditions are the main factors for agricultural production and thus population. More recently, Yang and Ma (2009) constructed a Natural Environment Suitability Index (NESI) model of China that included variables such as climate, hydrology, surface configuration and ecological conditions, and used the composite index to evaluate the relationship between population density and the natural environment. Similarly, Feng et al. (2009) analyzed the ecological environment suitability for human settlements in China by considering terrain, climate, water, and land use and cover change. Li et al. (2011) evaluated the human settlements environment suitability in the Three Gorges Reservoir Area of Chongqing, and considered factors such as terrain, climate, hydrology, vegetation and other natural factors. Most of the variables considered in the aforementioned studies overlap in terrain, climate and hydrological conditions.

The main objectives of this paper are threefold. Foremost, we examine whether there is a line for the natural environment, similar to the Hu Line, in China. Secondly, is the environment-based line consistent with the population-based Hu Line? Thirdly, where do the two elements (physical environment vs. population settlement) coincide and differ on a regional scale? Our approach goes beyond the traditional regression model that focuses on explaining the variability of population density by environmental factors. The emphasis centers on the scientific foundation for the validity of the Hu Line and possible consistency and inconsistency between population-based and environment-based divisions. This leads us to tap into the long tradition of regionalization in geographic analysis, which has been revitalized by some recent advancement of GIS automated regionalization methods. Specifically, GIS-based regionalization groups similar

and adjacent small areas into larger regions while optimizing a given objective (e.g., maximum homogeneity within derived regions), and enables us to derive the "natural" divisions on the two elements with scientific rigor.

3.2 DATA AND VARIABLE DEFINITIONS

There are two types of data for this research, namely population data and physical environment data. The population data is straightforward, and is extracted from the latest (i.e., 2010) census data of China at the county level. The following discusses the variables used to define physical environment, and corresponding data sources and related processing (see Table 5).

Table 5. Summary of data sets

Data type		Spatial unit/ resolution	Year	Data provider
Demography	Population	County	2010	National Bureau of Census, China
	Elevation	1 km	Contemporary	DIVA-GIS
Physical environment	Temperature	1 km	1970–2000	Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC)
	Relative humidity	752 meteorological Stations	2001–2010	
	precipitation	1 km	1970–2000	
	Water body	County	2010	

Physical environment factors are complex and diverse. Many factors can influence the suitability for human settlement. Based on literature review, we strive to define four indices: elevation index (EI), topographical relief index (TRI), climatic suitability index (CSI), and hydrology index (HI).

Elevation is considered as an important index. The elevation data is extracted from the DEM data with 1 kilometer spatial resolution, originally produced by the NASA, available from DIVA-GIS (http://www.diva-gis.org/). The average elevation of each county is used as the *Elevation index (EI)* in this research.

28

Topographical relief index (TRI) is calculated based on the aforementioned DEM data, such as (Feng et al., 2007):

$$TRI = [Max\,(H) - Min(H)] * [1 - P(A)/A] \qquad (3\text{-}1)$$

where *Max* (*H*) and *Min* (*H*) are the highest and lowest elevations, respectively, *P* (*A*) is the flatland area size in km^2, *A* is total area in a county. *Max(H)* and *Min(H)* can be directly extracted from the elevation data in each county. Flatland area is defined as follows.

An area is defined as flatland when the maximum difference of elevation in a certain area is less than 30 meters (Feng et al., 2007). We utilize a 5 km×5 km moving window to extract the maximum and minimum elevations. The detailed procedure is:

1) Extract the maximum and minimum elevation values within a 5 km×5 km filtering window and use them to represent the value at the center of the window. The window moves across the study area until values at all locations are obtained. Then two new images are created, maximum elevation value image and minimum elevation value image.

2) Compute a new image by map algebra subtraction operator of the maximum elevation value image and minimum elevation value image.

3) Extract grids with elevation difference values less than 30 meters as flatland area.

4) Calculate the total flatland area in each county.

Three variables are considered for climate. *Temperature* and *precipitation* are defined as the averages based on the monthly average data for 12 months over 30 years for 1970-2000. *Relative humidity* data is based on 752 meteorological stations across China from 2001 to 2010. All the above three types of data are provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn). Based on the average relative humidity at each station over the period (2001-2010), ordinary kriging interpolation is

used to generate a surface of estimated relative humidity for the whole study. Specifically, the spherical semivariogram model is used to perform the interpolation. The output cell size is 5 km×5 km, and the number of nearest input sample points is set at 12. Then averaging the relative humidity values across all interpolated points within a county yields relative humidity for the county.

A combination of temperature and relative humidity are often used to measure climatic suitability. This research defines the *climatic suitability index* (CSI) based on temperature-humidity index proposed by Thom (1959) such as:

$$CSI = T – 0.55(1–RH) (T–14.5) \qquad (3\text{-}2)$$

where *T* is average annual temperature (℃) and *RH* is relative air humidity (%).

Water resource in an area can be evaluated from two aspects, precipitation and water storage in various water bodies. Together they are captured by a *hydrology index (HI)*. According to Li et al. (2011), HI is calibrated as

$$HI = \alpha\, P + \beta W \qquad (3\text{-}3)$$

where *P* is normalized annual precipitation, *W* is normalized water body, and α and β are coefficients for weighting the two factors. As suggested by Li et al. (2011), $\alpha = 0.8$ and $\beta = 0.2$. The precipitation measure *P* is defined from the aforementioned climate data. The water body measure *W* (including rivers, lakes, reservoirs, permanent glaciers, beach areas, and tidal flats) is extracted from land cover and land use data in 2010, downloaded from the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn).

3.3 METHODOLOGY

As stated previously, the physical environment factors act in totality and not individually in influencing population settlement. After these factors are defined, the first task is to integrate these factors into a comprehensive *habitation environment suitability index (HESI)*.

Factor analysis is used here to identify latent constructs or factors by condensing original variables that may contain some duplicated information (e.g., some variables may be correlated with each other) into a smaller set of independent factors (Wang, 2009). After several experiments, two factors are selected as the scenario balances the amount of information retained and simplicity of resulting factorial structure, and more importantly yields meaningful interpretations. The diagnosis procedures for factor analysis yield: KMO measure of sampling adequacy = 0.593, and Bartlett's test of sphericity with Chi-Square = 5263.1 and very significant ($p<0.001$). That is to say, the relationship among variables is strong and factor analysis is appropriate for consolidating the data.

Table 6. Factor loadings of four environmental variables on two factors

	Climatic factor	Terrain factor
NEI	-.385	**.854**
NTRI	.-.011	**.952**
NCSI	**.879**	-.361
NHI	**.958**	-.018
Eigenvalue	2.478	1.128
% of variance explained by each factor	62.0	28.2

Note: number in bold indicates the largest loading of a variable on one factor

The result on factor structure is shown in Table 6. As all original variables are *normalized* in the process of factor analysis, these four variables, namely elevation index (EI), topographical relief index (TRI), climatic suitability index (CSI) and hydrology index (HI), are renamed NEI, NTRI, NCSI and NHI, respectively, to indicate corresponding normalized values. The variables

31

elevation and topographic relief (NEI and NTRI) are mostly captured by factor 2, and thus labeled

"terrain factor". The variables climatic suitability and hydrology (NCSI and NHI) are mostly

captured by factor 1, and labeled "climatic factor". The climatic factor accounts for 62.0% of total

variance, and is far more important than the terrain factor accounting for 28.2% of total variance.

The two factors together account for 90.2%, i.e., a very high ratio, of total variance. In other words,

the two factors are able to effectively capture the lion's share of the original four variables.
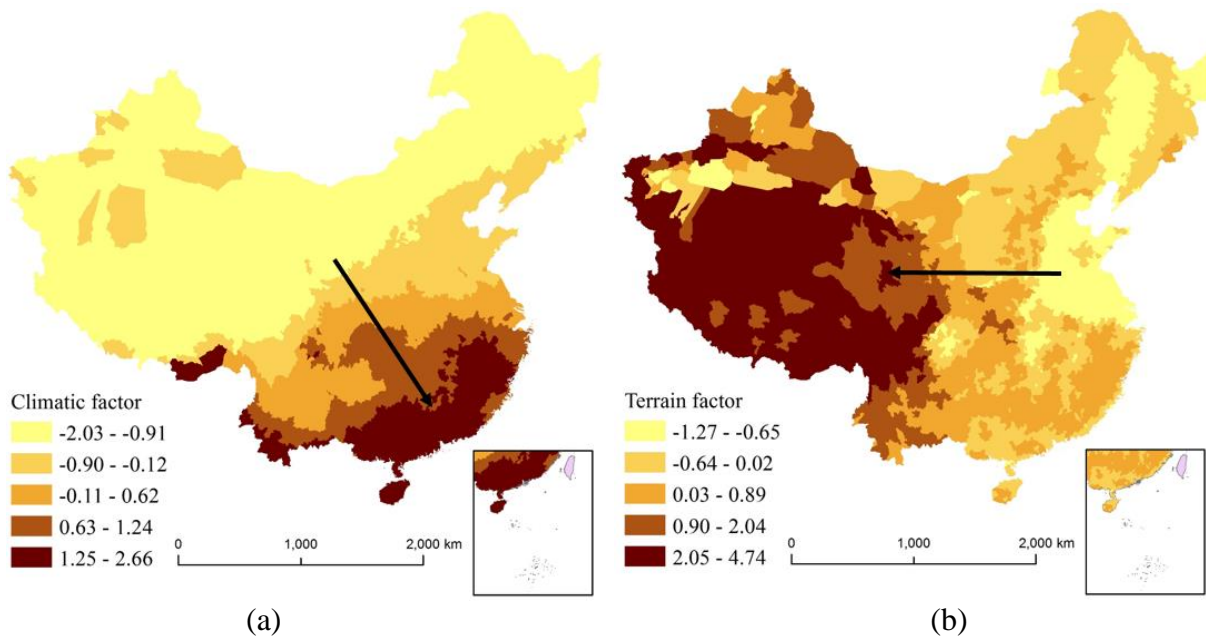
Figure 8. Consolidated environmental factors in China: a. climatic factor; b. terrain factor

Figure 8a and 8b show how the climatic factor and the terrain factor vary across China,

respectively. The climatic factor has a decreasing trend from southeast to northwest. Higher

climatic value signifies higher temperature and more abundant water, and thus a climatic condition

preferable for crop growth and favorable for human settlement. The terrain factor distribution map

suggests an opposite trend, i.e., increasing trend from east to west. A higher value means a rougher

and higher-altitude terrain and rather a more challenging environment for human settlement. As

the corresponding eigenvalues for the factors indicate their relative importance, the composite habitation environment suitability index (HESI) is calculated as

$$HESI = 2.478 \times Climatic\ Factor + 1.128 \times Terrain\ factor \qquad (3\text{-}4)$$

Note the opposite effects of a higher value of climatic factor and a higher value of terrain factor, and thus the terrain factor score is multiplied by -1. Both factors are standardized.

Figure 9 shows the HESI distribution pattern. The general pattern of HESI is a decreasing trend counterclockwise from the southeast to west. The highest HESI values are observed in the southeastern region, and extend to north and northeast, and to a less extent toward northwest. The lowest values are in the west, especially in the Tibetan Plateau.

Figure 9. Spatial distribution pattern of habitation environment suitability index (HESI)

As our main objective is to explore the influence of physical environment on the disparity in population distribution divided by the Hu Line, this study focuses on deriving two regions.
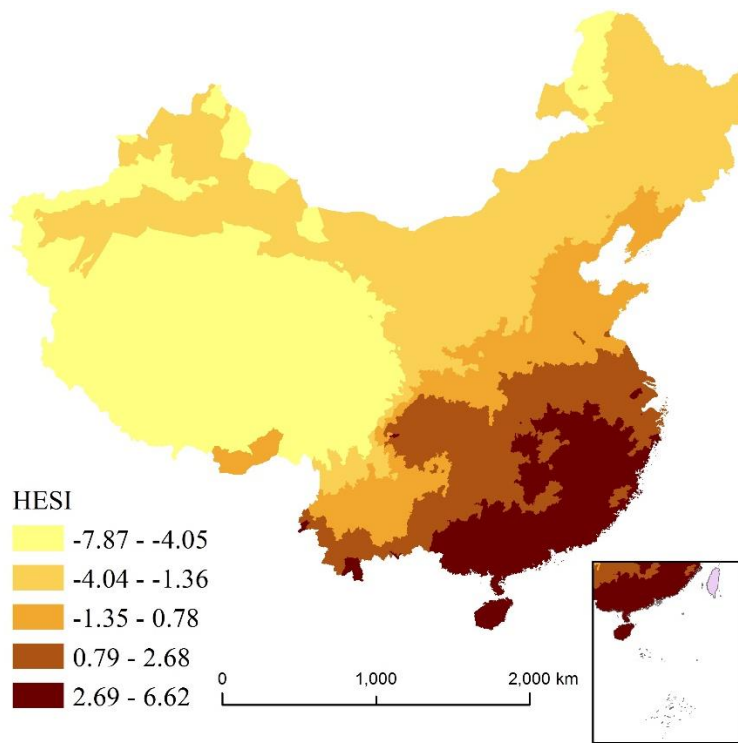
We begin with using the HESI values to define the attribute homogeneity in regionalization by REDCAP. One enhanced feature of the refined REDCAP by Guo and Wang (2011) is to accommodate the constraint that derived regions have a threshold size measured in a variable (e.g., area size). Recall that the Hu Line divides China into two regions with comparable areas but significantly different population sizes. Therefore it is desirable to derive regions with similar area sizes.

As a baseline, when no threshold area size is imposed, the REDCAP yields the overall heterogeneity SSD = 807.20, marked as the "Regionalization Line with No threshold" in Figure 10a. However, the resulting southeast region only accounts for 22.4% of the total area. See Table 7. In order to derive regions with comparable areas, we have experimented with various regionalization scenarios by gradually increasing the threshold area size. The algorithm does not converge when threshold areas are set above 48%. When the threshold area is set at 48% of the total area (i.e., 4.60 mi km2), it yields SSD = 1303.42. The resulting southeast region accounts for 48.7% area and the northwest for 51.3% area, marked as the "Regionalization Line with 48% threshold area" in Figure 10a. Understandably, its SSD value is higher, and yields two regions less homogeneous, than the baseline scenario without any constraint. Hereafter, HESI-based regionalization line refers to the one with a 48%-area threshold as it yields comparable area sizes in the two regions.
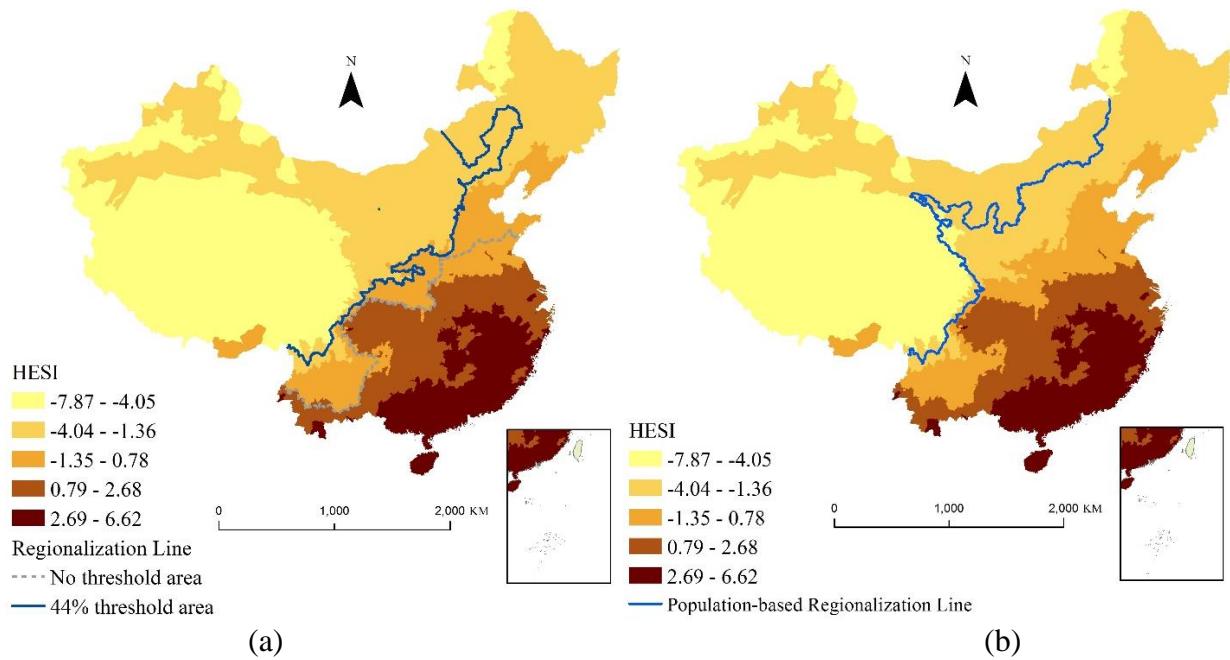
Figure 10. Two regions derived by REDCAP: (a) HESI-based regionalization, (b) population-based regionalization

Table 7. Basic statistics for 2 REDCAP-derived regions based on HESI and population

Regionalization variable	Area threshold in % of total area	Variable	Southeast Region (SE)	Northwest Region (NW)	Ratio (SE/NW)	Difference (SE-NW)
		Population (mil) (%)	818.86(61.7%)	508.52(38.3%)	1.61	310.34
		Area (mil km^2) (%)	2.12(22.4%)	7.33(77.6%)	0.29	-5.21
HESI	no threshold	Mean HESI	2.38	-3.91		6.29
		Sum of squared deviations (SSD)		807.20		
		Population (mil) (%)	1197.29 (90.2)	130.09 (9.8%)	9.20	1067.2
		Area (mil km^2) (%)	4.17 (44.1%)	5.28 (55.9%)	0.79	-1.11
HESI	44%	Mean HESI	0.28	-4.70		4.98
		Sum of squared deviations (SSD)		1332.86		

(Table 7 continued)

Regionalization variable	Area threshold in % of total area	Variable	Southeast Region (SE)	Northwest Region (NW)	Ratio (SE/NW)	Difference (SE-NW)
Logarithm of population density	no threshold	Population (mil) (%)	1292.9 (97.4%)	34.5 (2.6%)	37.48	1258.4
		Area (mil km^2) (%)	4.80 (50.8%)	4.65 (49.2%)	1.03	0.15
		Mean HESI	-2.0	-9.38		9.18
		Sum of squared deviations (SSD)		1338.6		

For comparison, the REDCAP regionalization method is also applied to the distribution of population. Empirical studies on regional population density patterns such as in China (e.g., Wang 2001) suggest that population densities decline exponentially with distances from major cities. In other words, the spatial variability of population density is better captured by the logarithm of density than the density itself. Therefore, population density in logarithm is used to measure attributive homogeneity in REDCAP. The regionalization result with no threshold constraint for region size in area is presented in Table 7 and Figure 10b. The two derived regions are almost identical in area size (50.8% in southeast region and 49.2% in northwest region), but stunningly contrasting in population (population in the southeast is 37.48 times of that in the northwest).
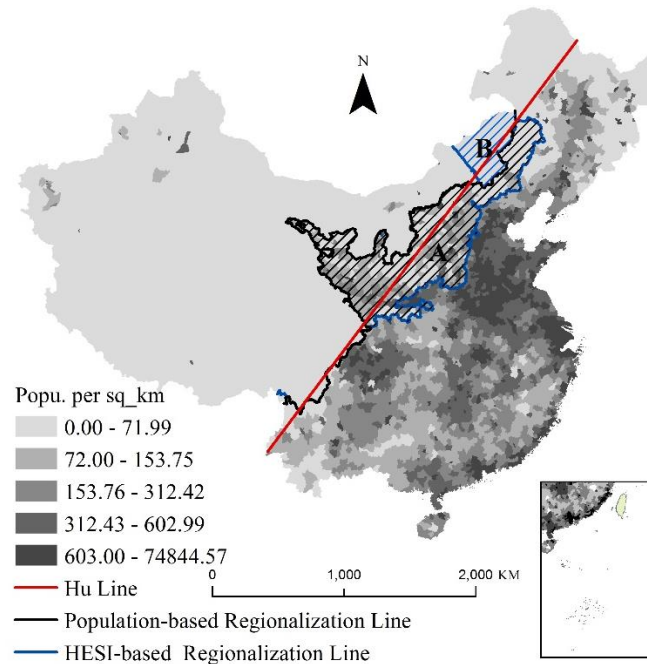
Figure 11. Comparison of Hu Line and two REDCAP regionalization lines

Figure 11 shows the discrepancies between the HESI-based and population-based regions derived from REDCAP, with the Hu Line on the background. For the most part, the two regionalization lines are largely consistent, and divide China into the northeast and southwest parts with contrasting population density and physical environment patterns. The two types of regionalization yield similar SSD values (1303.42 for the HESI-based one and 1338.6 for the population-based one). The areas in disagreement are highlighted in shade: Region A in Loess Plateau region, and Region B in Xilin Gol high plain. Region A is grouped in the higher-population-density southeast by the population-based regionalization, but in the more-challenging-environment northwest by the HESI-based regionalization. One may interpret that Region A currently is inhabited by more population than its environment suggests, and thus is likely to experience more challenges on sustainable development. On the contrary, Region B represents areas grouped into the more-favorable-environment southeast but lower-population-density

northwest, and suggests a physical environment perhaps capable of accommodating more population growth.

3.5 EXPLORATORY ANALYSIS ON (IN)CONSISTENCY BETWEEN POPULATION AND HESI PATTERNS

As stated in the previous section, the geographic distributions of population density and HESI are mostly consistent with each other at the macro scale. This section takes a closer look at the consistency and inconsistency between them. Considering possible scaling effects in the measurement of variables, we have experimented with four correlation models at the county level between (1) density vs. HESI, (2) ln(density) vs. HESI, (3) density vs. ln(HESI), and (4) ln(density) vs. ln(HESI), and found that the fourth model yielded the highest correlation coefficient (0.67).
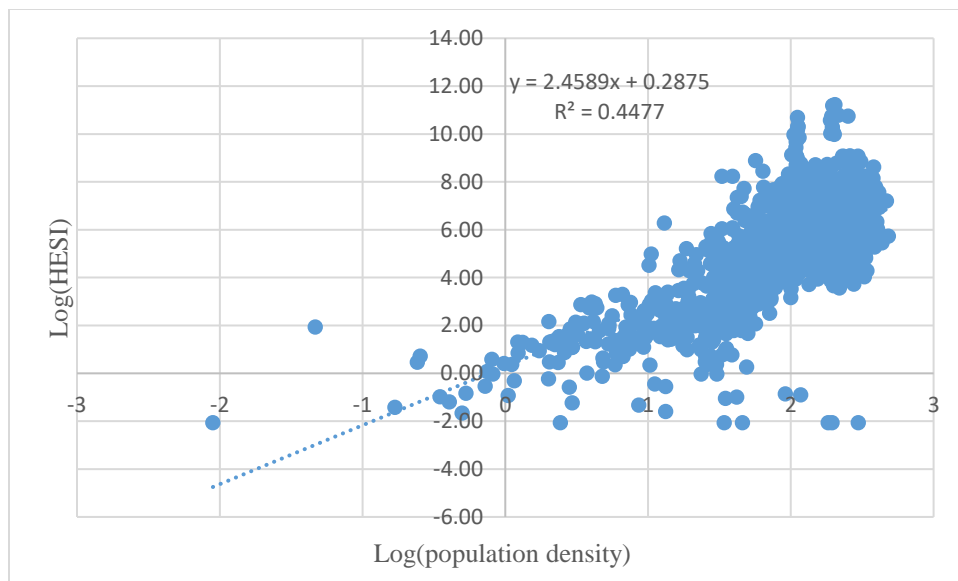
Figure 12. Regression analysis of log(population density) and log(HESI)

Where do the two variables coincide and divert? As shown in Figure 12, there is a positive correlation between them, and the OLS regression yields $R^2 = 0.4477$. Note that the HESI values

vary from -7.87 to 6.62, and thus they are uniformly inflated by adding 8 to avoid taking logarithm of negative values. In other words, the regression model yields:

$$ln(density) = 0.2875 + 2.4589\ ln(HESI+8) \qquad (3\text{-}5)$$

The residuals from the above regression model indicate whether the HESI over-predicts or under-predicts the population density in a particular area. Figure 13 shows the geographic variability of the regression residuals across the study area. It clearly exhibits some spatial clustering. With the Moran's I = 0.55 and Z-score = 149.22 (statistically highly significant), the residuals are spatially autocorrelated. This suggests the potential benefit of grouping areas of similar attribute (here "regression residuals") in understanding the regional pattern of (in) consistencies between population density and HESI.
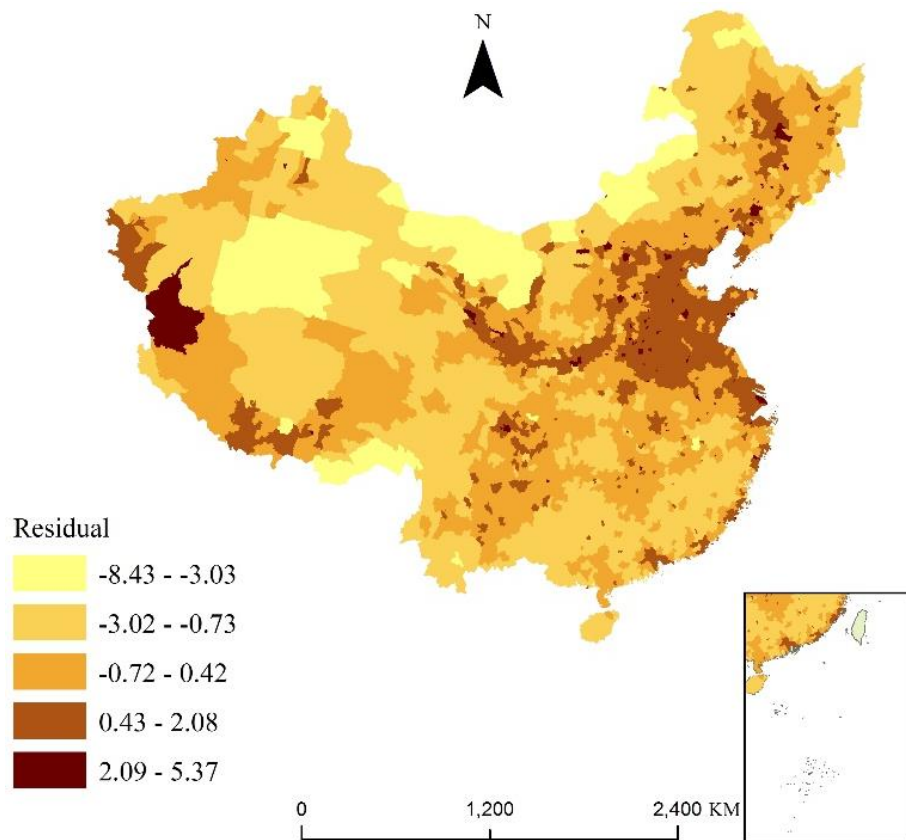
Figure 13. Spatial distribution pattern of residual

39

Here the REDCAP regionalization method is used once again to consolidate the spatially-correlated areas and explore the regional pattern in the residuals. Specifically, the regression residuals are used to define the variable measuring homogeneity of derived regions by REDCAP. The analysis begins with determining the number of regions. Figures 14a and 14b show how SSD and reduced SSD decline with an increasing number of regions, respectively. The two-region scenario has clearly the most noticeable reduction of the SSD, and produces Regions I and II as shown in Figure 15a. Also see Table 7 for related statistics (1st scenario when HESI is the attribute variable without any constraint for area threshold). Region I, accounting for 22.4 % total area, includes the three major plains in China (the North China Plain, Northeast Plain, Lower-Middle Yangtze River Plain), and is composed of areas with positive residuals, suggesting higher than predicted population densities by HESI. Region II claims the remaining part for 77.6% area, mostly with negative residuals and thus lower than predicted densities.

(a)   (b)
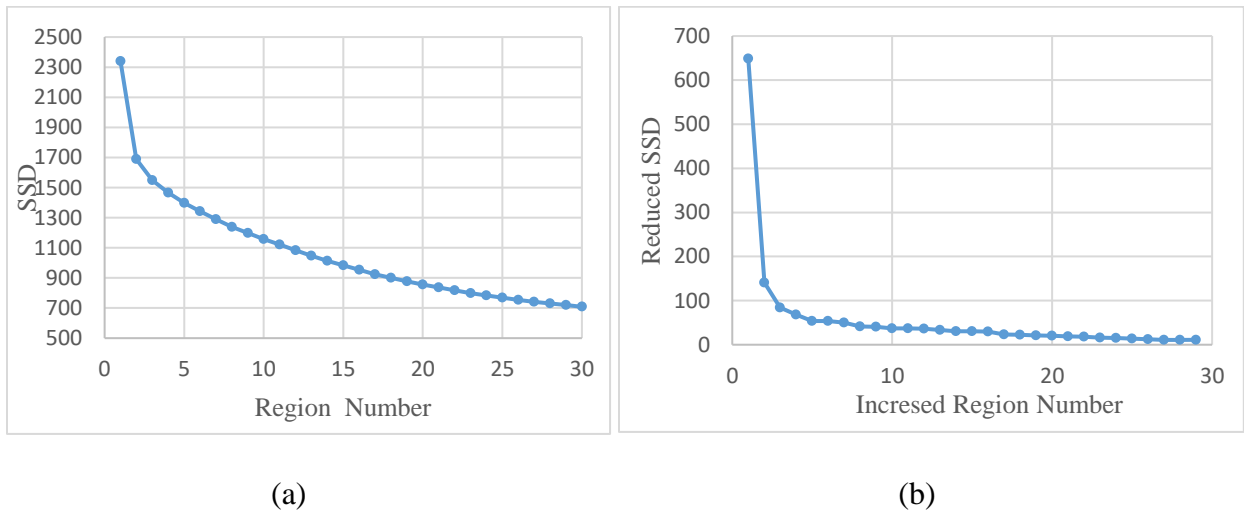
Figure 14. SSD values and reduced SSD in various regionalization scenarios
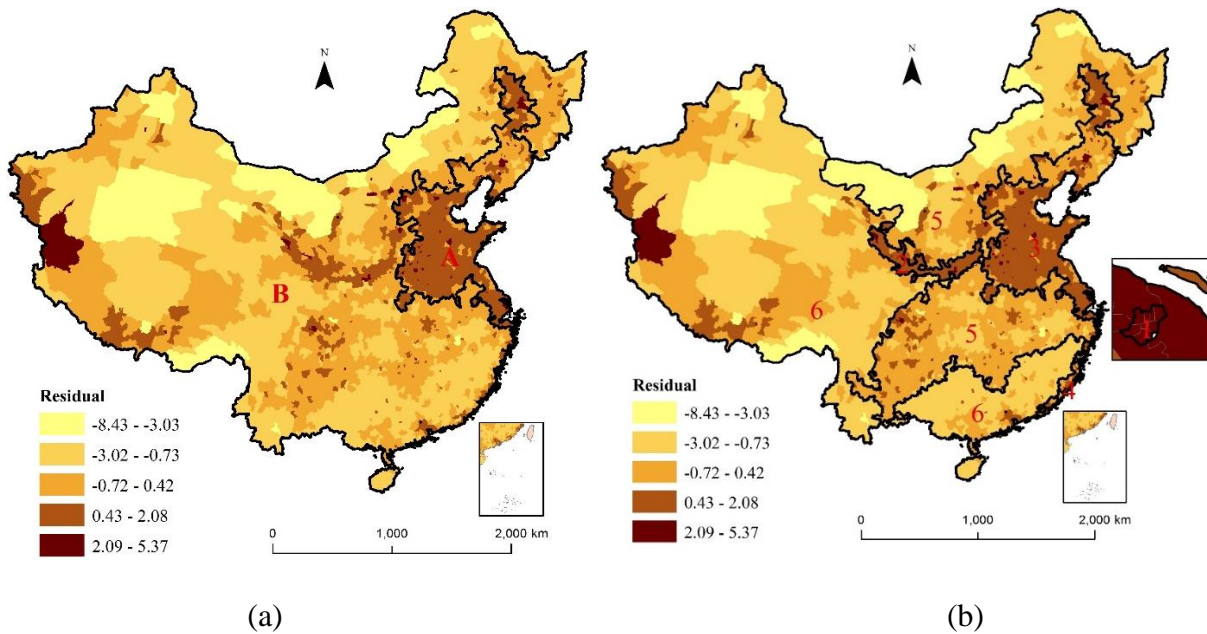
Figure 15. Two regionalization scenarios based on residuals

After the number of regions goes beyond six, the reduced SSD values level off, and suggest a possible turning point at 6 regions. As stated previously, REDCAP generates a hierarchical system of regions, and gradually creates more regions by carving one region at a time from a previous round. As shown in Figure 15b, Regions 1 and 3 are carved out of Region I from the two-region scenario, and Regions 2, 4, 5 and 6 are carved out of Region II (hereafter these regions are simply referred to as R1, R2 …). Note that R5 is composed of two large sub-regions (North China and South-central China) connecting by a narrow band in the middle and thus geographically contiguous, and R6 is also made of two large contiguous sub-regions (West China and South China). Table 8 reports the average residual, population density and HESI and provides a brief description for each region.

Table 8. Basic statistics for 6 REDCAP-derived regions based on the regression residual

Order of regionalization	Mean Residual	Population Density (Popu./km^2)	Mean HESI	Region
1	4.58	39911.44	1.94	Over-populated urban Shanghai
2	1.09	338.45	-2.55	Over-populated Guanzhong Plain, south of Loess Plateau
3	0.73	547.27	-0.55	North China Plain, Northeast plain, Lower Yangtze Plain
4	0.31	979.20	3.18	Southeast coastal area
5	-1.21	143.32	-1.32	Less-populated south-central China and North China
6	-1.31	46.77	-3.72	Less-populated South China and West China

Based on Table 8, the orders of mean population density and mean HESI across the six regions are largely consistent. The mean of population density declines from urban Shanghai (R1), to southeast coastal region (R4), three major plains (R3), Guanzhong Plain and south Loess Plateau (R2), North China and South-central China (R5), and South China and West China (R6); and the order of HESI mean values is similar with two minor discrepancies (R4>R1, and R5>R2). The discrepancies in the two variables are better captured by the regression residuals. As a baseline, we consider (1) regions with residuals ranging (-1, 1) as "population-environment consistent", (2) regions with residuals higher than 1 as "environmentally-stressed", and (3) regions with residuals lower than -1 as "less-populated." More discussion for the three types of regions is as follows.

1) The highest value in residuals in urban Shanghai (R1) is understandable as urban development is less dependent on physical environment. However, Guanzhong Plain and its adjacency (R2) has a moderately-high population density and a highly-negative HESI and thus a relatively-high residual (1.09). It remains mostly rural and suggests a more stressed physical environment.

2) The three major plains (North China Plain, Northeast Plain, and Middle-Lower Yangtze River Plain) (R3) have very favorable physical environments for agriculture and also support moderately-high population density. A narrow stretch on the southeast coast (R4) enjoys the highest HESI (attributable to its most favorable climatic factor and reasonably amenable terrain) and is able to accommodate about twice as much population density as the three major plains. This region is also where several large coastal cities reside, which help explain its high density. The population density and physical environment are relatively consistent in these two regions.

3) Regions 5 and 6 make up Region II in the two-region regionalization scenario for about 78% of territory of the entire study area, and both have negative mean residuals. Either region is composed of two large sub-regions with very different underlying factors. Region 5 includes North China and South-central China. The North-China sub-region has below-average HESI (poor climatic factor and average terrain factor) and low population density, and the South-central China has moderate HESI value (favorable climatic factor but average terrain factor) and above-average population density. As a result, the two sub-regions in Region 5 share similar residuals. Region 6 also includes the very different South China and West China. The South China sub-region has the highest HESI value and also a very high population density, while West China sub-region has the lowest HESI value and the lowest density. The two sub-regions end with the most negative residuals.

It is important to bear in mind a few concerns for the above exploratory analysis. First, the development of the aforementioned HESI is intended to capture major physical factors that may influence human settlement, and it inevitably leaves other factors. The consolidation process by factor analysis also trims out part of the variability of original factors, which may be valuable in

43

explaining population density. Secondly, the human settlement pattern, especially today with an increasingly higher urbanized society and much of the economy composed of non-agricultural activities, is less dependent on or confined by physical environments. The association between the two has weakened over time and the trend may well continue. Thirdly, regression itself is designed to capture the trend, which may be heavily influenced by a relatively small number of outliers. While our study employed logarithmic transformation on both variables to mitigate the effect, the dominant effect of very high density areas in Shanghai and very low density areas in the west remains evident. In short, one should interpret the results with caution and refrain from suggesting which areas be "over-populated" and which areas be "under-populated."

3.6 CONCLUSION

This research examines the variability of physical environment and its relationship with the population distribution pattern in China. It employs a unique approach by GIS-automated regionalization, namely the REDCAP method, to group areas into a small number of regions that are relatively homogenous. By defining regions based a composite physical environmental index (HESI) and population density and comparing the two types of regions, we are able to solidify the connection between them at a regional level. The results indicate that the sharp population disparity between the northwest and the southeast in China is largely supported by a similar contrast in physical environment. In addition, this research also attempts to identify the consistency and inconsistency between the two in more geographic details, and proposes three types of regions, namely environmentally-stressed, population-environment consistent, and less-populated regions.

An obvious lesson from the study is that one cannot assert whether the population distribution in a country or a region is unbalanced and needs adjustment by merely focusing on its uneven population density pattern. The overall consistency between the HESI and population

delineation lines suggests that the disparity in population distribution be largely attributable to underlying environmental factors. Policy makers over several generations in China have been tempted to overcome the invisible barrier of Hu Line by relocating population from the higher-density southeast to the lower-density northwest. Such a notion needs to be carefully examined at finer scales before it triggers massive ecological disasters beyond repairable. As discussed previously, when a more refined and comprehensive physical environment indicator is available, regional planners may follow the methodology developed in this research to identify different types of regions and devise corresponding developmental policies.

There are also some limitations in this research. The physical environment system is much more complicated than our HESI can capture. There are other important factors, such as the quality of soil, light, wind etc., which also influence population distribution. Each factor could be measured more accurately with better data. Better methods may also help integrate the individual factors together. Furthermore, the effect of urbanization on population distribution needs to be better modeled. Urban areas, though not completely free from environmental constraint, are increasingly less reliant on it. The research methodology is valuable for others to refine and improve for future work.

CHAPTER 4. DISPARITY ANALYSIS OF POPULATION DENSITY IN CHINA, 1953-2010

4.1 INTRODUCTION

China has been the most populous country in the world for centuries, and has recently experienced formidable challenges related to its large population growth. When China carried out its first census in 1953, the population was about 582 million. By the sixth census in 2010, the population had doubled to 1.327 billion. Knowing how the population changed (e.g. its size, composition, growth rate, and spatial distribution) is important to understand and anticipate population changes in the future. Furthermore, since people are both labors and consumers, i.e., a vital element in the supply and the demand sides of an economic system, the population distribution and its temporal variation are an intensive embodiment of socioeconomic activities (Huang, 2005). In addition, as a core element of urbanization, population distribution pattern and change trend can effectively reflect the urbanization process (Mao et al., 2016). This study focuses on its spatial distribution pattern, which is the basic information to project future needs for infrastructure and services, including housing, transportation, human services, and community facilities.

Due to the physical environment, human settlement history, and economic development policy, China's population distribution is astonishingly unbalanced with most of the population concentrated in the eastern half of the country and a small proportion in the west. A large body of academic literature attempts to describe and explain the pattern of spatial disparity. A classic study was the proposition of "Hu Huanyong Line" (simply referred to as "*Hu Line*" thereafter) by Huanyong Hu (1935, 1990). The Hu Line begins from Heihe in Heilongjiang Province in the northeast to Tengchong in Yunnan Province in the southwest (see Figure 1 and related details in Section 4.4). With similar area sizes, Hu (1935) estimated that the northwest side of the line had only about 4% of the country's total population, and the southeast side had nearly 96% of the

population, a strong contrast in population density on the two sides. Since the inception of Hu Line, researchers have largely confirmed its brilliance and persistency in characterizing China's population disparity ever since. Considered one of the greatest geographic discoveries of China (Shan, 2009), it is forever tied to Hu's legacy.

The uneven population distribution pattern is also linked to inequality in economic development and other problems in China. For example, most of the western region has not benefited as much from China's recent spectacular economic growth, and continued to be plagued by a high poverty rate. On the other hand, the high-density population in the southeastern region elevates the public's concerns of crowded housing, environmental pollution, public resource shortage, traffic congestion, resource depletion, and other problems (The World bank, 2007; Yu et al., 2012; WHO, 2013; Chen et al., 2013). An increased regional disparity may lead to serious social and political problems, and "negatively influence China's economic and social stability" (Xue, 1997, p. 46). Reducing the disparity in population distribution between the two sides has been explicitly stated as a desirable development goal by the central government of China (Guo et al., 2016).

Data for this study include all the six censuses (1953, 1964, 1982, 1990, 2000, and 2010) since the establishment of People's Republic of China in 1949. During the study period, China sustained several major natural and man-made disasters (e.g., the 1958-62 Great Famine, the 1966-76 Cultural Revolution), and also experienced some major government-sponsored migrations, economic growth, and unprecedented urbanization. It is necessary to evaluate whether and how these influenced the spatial distribution pattern of population in China, especially the footprints on the disparity of population density pattern. There are at least three motivations for this study.

1) We need a scientifically driven and rigorous method of deriving a delineation line that

supports (validates) or raises suspicion (rejects) for the Hu Line, which was largely drawn manually in an era of poor data quality and compute capability.

2) As stated above, it is of great value to examine possible changes of this delineation line over time and contemplate the underlying forces.

3) The results on (in)stability of the Hu Line may shed light on whether it is feasible and prudent to guide public policy toward overcoming the "barrier" of Hu Line in population settlement and related economic development.

This study examines how the population pattern changed in the last six decades or so, and whether the disparity in population density has reduced or execrated over time. The reminder of this paper is organized as follows. Section 4.2 discusses the data and data analysis methods. Section 4.3 examines the variability and disparity of population density at the county level. Section 4.4 uses a GIS-automated regionalization method to derive a delineation line that divides China into two regions, compares it to the Hu Line, and analyzes its changes over time. The final section concludes the paper by highlighting the major findings and discussing future work to extend this study.

4.2 DATA AND METHODS

4.2.1 Data

The study area is mainland China with a total area of about 945 million km^2. Due to lack of data, Taiwan, Hong Kong, and Macao are beyond the scope of this research. All the six census data sets are available at the county level, which is our basic research unit. There are 2340 counties in the study area. For the first three censuses (1953, 1964 and 1982), the population was based on household registration status (*hukou*). That was relatively reliable since the *hukou* system at the time strictly required people to live and work where they were registered. It was not until 1978

with the Open Door and Economic Reform Policy that people were permitted to seek work away from their registered locations. For the fourth census in 1990, the census population included those who lived in the place more than one year. For the fifth and sixth censuses, the census population included those who lived in the place more than half year (including foreigners).

Over the study period 1953-2010, a small number of counties experienced minor changes in their boundaries. In order to examine the changes over time consistently, this study uses the areal weighting interpolator (Goodchild and Lam, 1980) to integrate all census data in one unified unit, i.e., the 2010 county unit. In other words, population is interpolated proportionally to corresponding area sizes when counties are split into multiple parts by overlaying the county boundaries in different years.

4.2.2 Assessing county-level population density disparity by Gini coefficient

Gini coefficient is derived from the Lorenz curve, a graphical representation of income inequality or wealth inequality developed by American economist Max Lorenz (1905). This research uses Gini coefficient to measure the disparity of population density that reflects the discord between population and land area across counties.
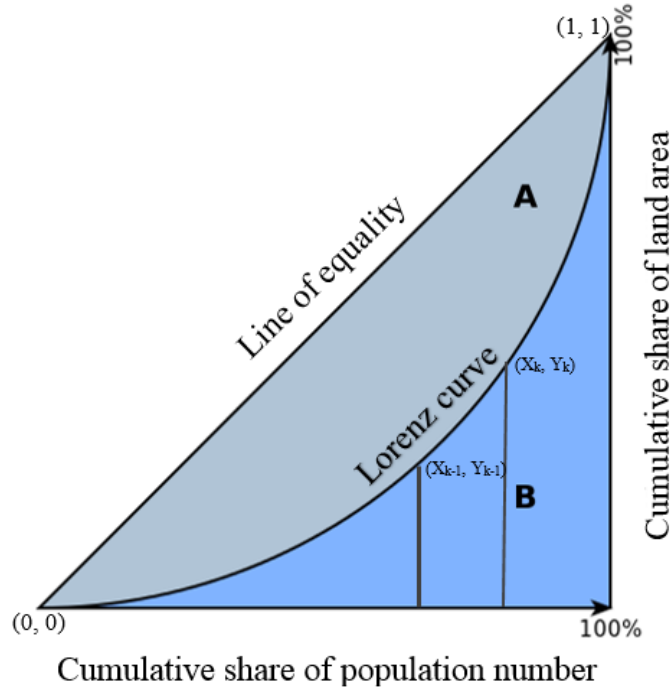
Figure 16. Illustration of Gini coefficient

As shown in Figure 16, the horizontal line is the cumulative share of population, and the vertical line is the cumulative share of land area. The Gini coefficient is the ratio of A (i.e., the area between the Lorenz curve and the line of equality) out of the total area of A and B (i.e., the triangle below the line of equality), written as

$$\text{Gini} = A/(A + B) = 2A \qquad\qquad (4\text{-}1)$$

where A+B is always 0.5, and thus Gini = 2A.

While the entire Lorenz curve is unknown, the values at certain intervals (here, by 2,340 counties) are given. In this case, the Gini coefficient is approximated as

$$\text{Gini} = \text{abs}(1 - \sum_{1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1})) \qquad\qquad (4\text{-}2)$$

where n is the total number of counties, X_k and Y_k are the cumulated percentages of population and land area for each county, respectively, and "abs" stands for absolute value as the Gini value is always positive.

The Gini coefficient value ranges 0-1 with 0 for minimum inequality and 1 for maximal inequality. A larger Gini coefficient indicates higher inequality.

4.3 ANALYSIS OF CHANGES IN POPULATION DENSITY AND DISPARITY AT THE COUNTY LEVEL

4.3.1 Examining population density changes

Table 9. Population changes between two census years

Census year	Annual population growth rate (‰)	Number of counties	
		With Population Decline	With Population Growth
1953-1964	16	408	1,932
1964-1982	21	104	2,236
1982-1990	15	143	2,197
1990-2000	10	543	1,797
2000-2010	7	957	1,383

As shown in Table 9, the annual population growth rate has fluctuated since the first census in 1953. It increased from the year 1953 to 1982, and then began to decline from 1982. The rapid population growth from 1953 to 1982 benefited from the economic progress and social stability after the foundation of PRC. In addition, the family planning policy (or birth control campaign) was voluntary before 1970s. Therefore, China experienced a rapid population growth period from the first census to the third one. From 1970s, the family plan policy transformed from voluntarily birth planning campaigns to mandatory policy campaigns, and from late marriage and childbearing, birth spacing, and fertility limitation policy in 1970s to one child policy in 1980s (Attane,2002). It resulted in a sharp descent in fertility rates, which was the reason why the annual growth rate showed a continuous decreasing trend from 1982.
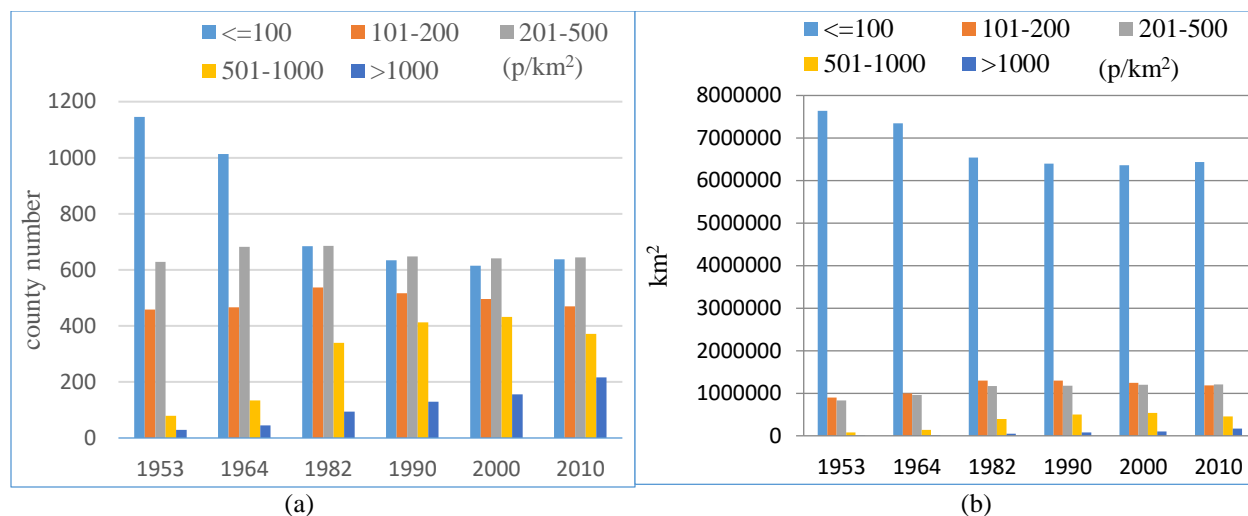
Figure 17. Distributions of population density at the county level 1953–2010: (a) number of counties, (b) area size

As for the population change in each county, 2,340 counties are classified into five categories: very low density (0-100 persons/km^2), low density (101-200 persons/km^2), median density (201-500 persons/km^2), high density (501-1000 persons/km^2), and very high density (>1000 persons/km^2). As shown in Figures 17a and 17b, in 1953, more than 1,100 counties had population density less or equal to 100 per/km^2, covering more than 80 percent of the total land area. Both the number of counties and total area size for this very-low-density category declined significantly from 1953 to 1982, and then became relatively stable after that. This was consistent with the trend of population growth in general as the earlier part of the study period experienced rapid population growth and the later period was fairly stable. The numbers and areas for low-density and median-density counties kept relatively stable for the entire study period. The number and land area of high-density counties had a noticeable and consistent growth trend from the year 1953 to 2000, and then began to decline from 2000 to 2010.The very-high-density counties, likely representing urban areas, accounted for small proportions in their number and land area. Both the number and land have been growing consistently over the entire period, and reached 216 counties and accounted for 1.8 percent of the total land area in 2010, reflecting the process of urbanization.
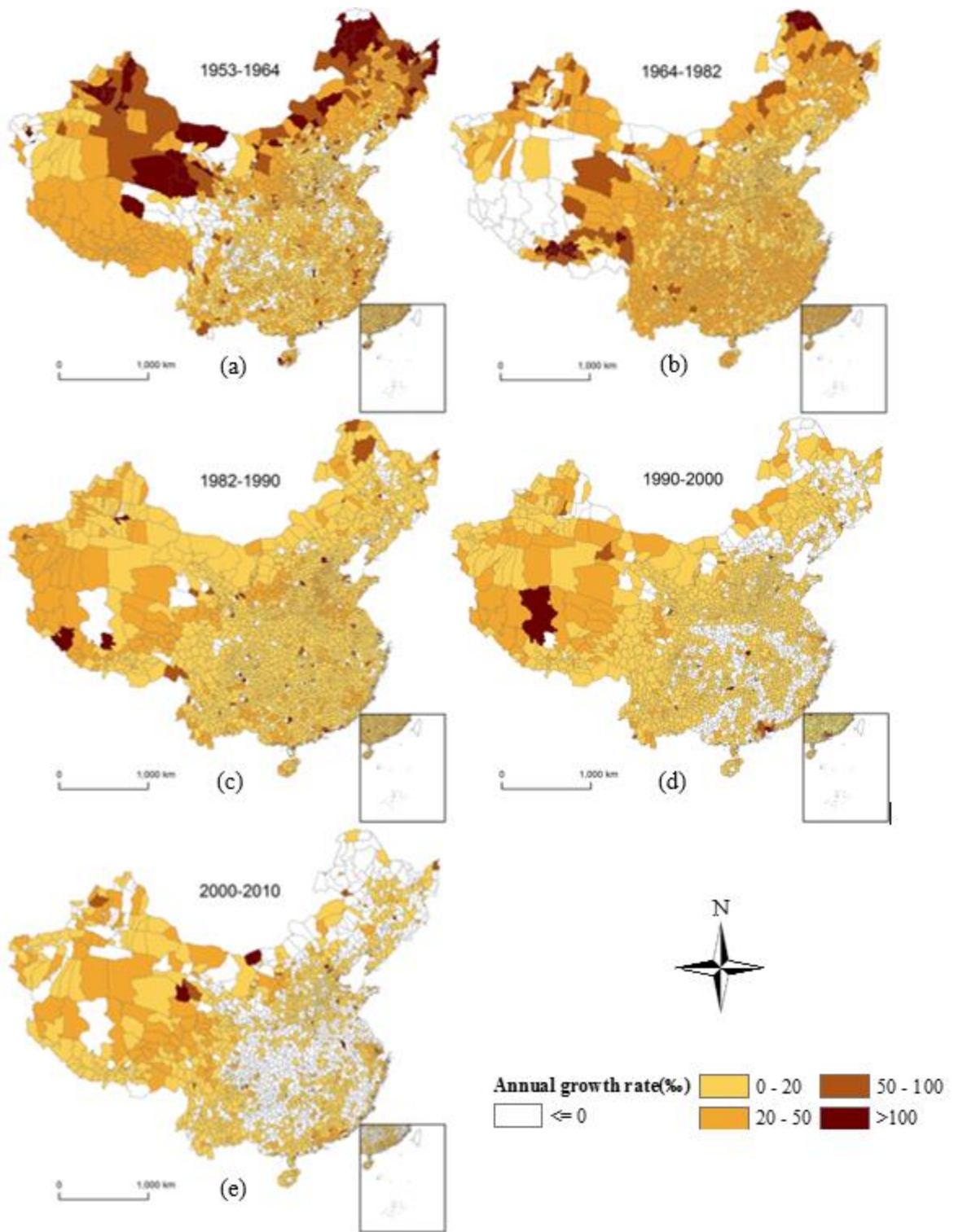
Figure 18. Population (density) growth rates at the county level
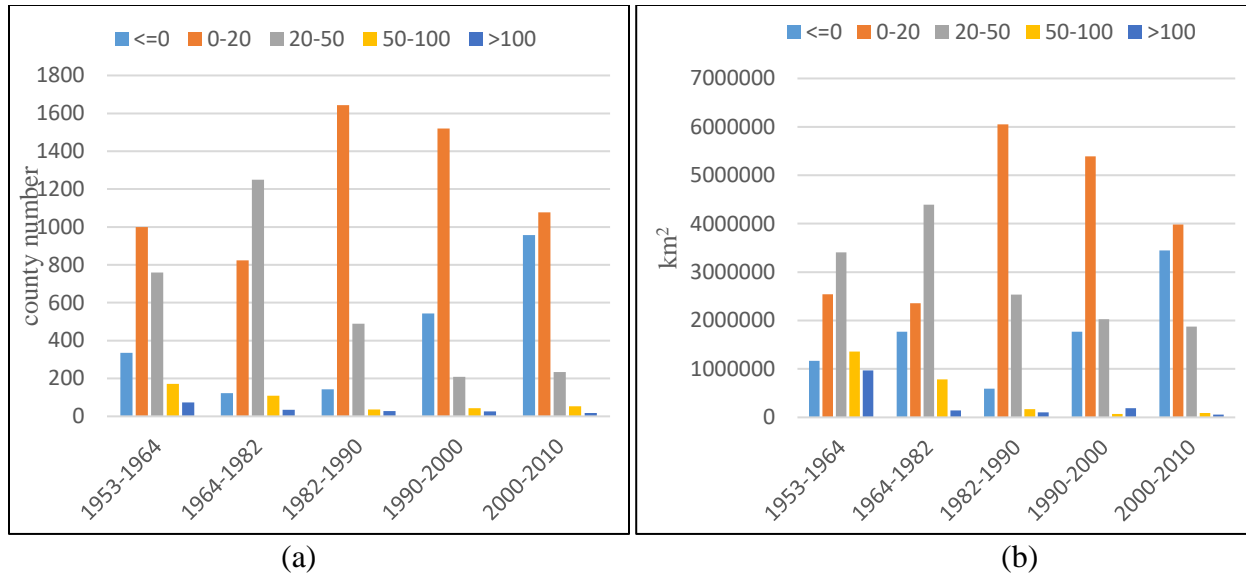
Figure 19. Distributions of population density growth rates at the county level 1953–2010: (a) number of counties, (b) area size

Based on the per thousand (‰) annual population growth rates, the counties were classified into five grades: negative growth (≤ 0), minimal growth ($0 < R \leq 20$), slow growth ($20 < R \leq 50$), moderate growth ($50 < R \leq 100$), and fast growth ($R > 100$). Since the area size for each county remained constant over time, population growth rates were equivalent to population density growth rates. As shown in Figures 18a-18b and Figure 19a-19b, for both periods of 1953-1964 and 1964-1982, most of the counties fell into the categories of minimal growth and slow growth. The moderate growth and fast growth counties also accounted for considerable proportion in both county number and area size as the country experienced rapid population growth overall, as stated previously. For 1982-1990 and 1990-2000, both the county number and size of minimal-growth counties were more than all the other combined, reflecting the national trend of gradually slowing down in population growth. For the most recent period 2000-2010, the minimal growth counties and negative growth counties accounted for the most proportion. The negative growth rate counties

54

reached the largest number (957) among the five periods, as these counties lost population to urbanization. This was especially pronounced in large rural counties in east part of China (Liu et al., 2009), and left many villages in these counties depleted in population (Long et al., 2012) and some cases even extended to townships and cities observed (Long and Wu, 2016). Also noticeable were many counties in west China with moderate and fast growth rates in 1990-2000 and 2000-2010 (Figures 18d-18e). These low-density counties were rather small in population but large in area size, and the actual gains in population were modest.

In short, over the study period, the number of fast-growth counties and their areas increased, so did the number and areas of negative-growth counties. In other words, the country became increasingly polarized in a contrast of population density as it underwent urbanization.

4.3.2 Analysis of population density disparity by Gini coefficient

Gini coefficient is used here to further assess the outcome of aforementioned increasing gaps in population (density) growth rates. As shown in Table 10, the Gini values for the entire study period stayed above 0.65, indicating highly unequal population distributions in China. As illustrated in Figure 20a, the Gini coefficient for the entire country started out high in 1953, declined until reaching the valley bottom in 1982, and began an upward trend until 2010. The trend is characterized as a U-shape. The variation was confined to a narrow range 0.650-0.670, reflecting that the uneven distribution pattern in population was persistent.

Table 10. Gini coefficients for disparity in population distribution

Census year	Whole country	Regions	
		Southeast	Northwest
1953	0.661	0.316	0.467
1964	0.653	0.380	0.428
1982	0.650	0.294	0.563
1990	0.652	0.299	0.579
2000	0.657	0.382	0.573
2010	0.669	0.403	0.641
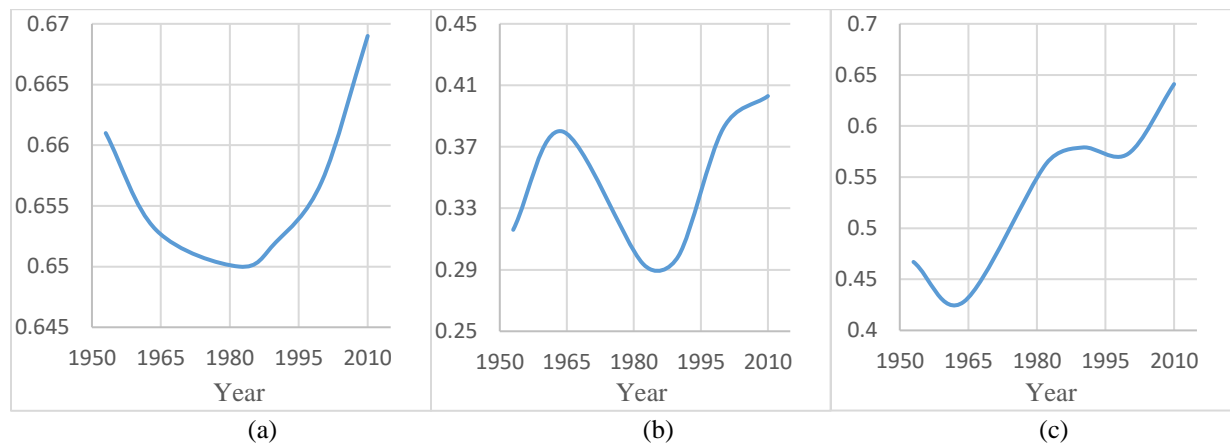
Figure 20. Gini coefficients for population density at the county level 1953-2010: (a) the whole country; (b) southeast; (c) northwest

Three possible forces help us understand the declining Gini values from 1953 to 1982. One was the "Third Front Movement", a massive industrial development by the central government in its interior from 1964 to 1976. During the period, national investment in industries and infrastructures focused in the mountainous inlands, considered safer than the east coastal areas for national defense. This movement supported significant government-sponsored migration in labor force. From late 1968 onward till 1976, another movement just as significant was the campaign termed "Up to the mountains, down to the villages." Millions of urban youth were mobilized and sent to rural villages and frontiers for "reeducation." Similarly, this brought much migration from more-urbanized and higher-density east coast to more primitive and lower-density inlands.

Thirdly, when the draconian birth control policy was launched in the early 1970s, it was much harsher and more strictly enforced in the urban areas and residents were subject to the one-child policy. In rural areas, couples were allowed to have two children under certain conditions. Moreover, minority couples could have two or more children (Gu et al., 2007), and disproportionally high percentages of minorities are distributed in the southwest and west of China.

The reversed trend of increasing disparity after 1982 was attributable to the unpresented economic development in the era of economic reform and open-door policy. During these periods, the east coastal regions experienced faster economic growth and urbanization than the rest of the country, and attracted massive migrant workers (Sun, 2013).

4.4. ANALYSIS OF POPULATION DENSITY DISPARITY BY REGIONALIZATION

The previous section examines population density disparity and changes at the county level. While the analysis is at a fine geographic resolution, the variability is detailed, fragmented, and sometimes hard to detect patterns or trend. Geography is best learned at various scales. This section moves to analyze the same theme in a different scale. Instead of an analysis at the prefecture or provincial level (both administrative units are large than county), we seek to construct "organic" regions by a GIS-automated regionalization method, specifically REDCAP, and examine the disparity in population distribution by the derived regions. For the interest of validating the classic and massively influential Hu Line, this study is limited to delineate one demarcation line and generate two regions. Based on the six censuses, we are also interested in probing possible shifts of the demarcation line over time.

4.4.1 Population (density) disparity in regions divided by the Hu Line

As a baseline, Table 11 reports the population sizes and population densities in the two regions divided by the Hu Line in the six census years. Given similar area sizes (5.28 million km^2 in the northwest and 4.17 million km^2 in the southeast with a ratio of 55.9%:44.1%), the contrast or disparity in population between the two regions divided by the Hu Line was very evident. The population ratio (southeast vs. northwest) declined from 95.78:4.22 in 1953 to 94.07:5.93 in 2010. The declining trend was consistent with the only exception that the ratio increased slightly from 1964 to 1982. In other words, disparity in population density between the southeast and northwest regions has been reduced. Given the large population size in the country, the 1.71 percent change in population proportion was very significant.

Table 11. Population and population densities in regions divided by the Hu Line

Census year	Whole country		Southeast		Northwest		Population % ratio (Southeast: Northwest)
	Population (mil)	Population density (p/km^2)	Population (mil)	Population density (p/km^2)	Population (mil)	Population density (p/km^2)	
1953	578.7	61.23	554,2	132.84	24.4	4.63	95.78 : 4.22
1964	689.7	72.98	658,4	157.80	31.3	5.93	95.46 : 4.54
1982	1,003.9	106.22	949,4	227.56	54.5	10.32	94.57 : 5.43
1990	1,130.5	119.61	1,067,8	255.93	62.7	11.88	94.45 : 5.55
2000	1,242.6	131.47	1,170,6	280.57	72.0	13.64	94.21 : 5.79
2010	1,327.4	140.44	1,248,6	299.27	78.7	14.92	94.07 : 5.93

4.4.2 Population (density) disparity in REDCAP-derived regions

The Hu Line divides China into two regions with comparable areas sizes but significantly different population number. As stated previously, this study attempts to derive a demarcation line that divides China into two regions with maximum contrast in population, and examines the (in)consistency between this simulated line and Hu Line.
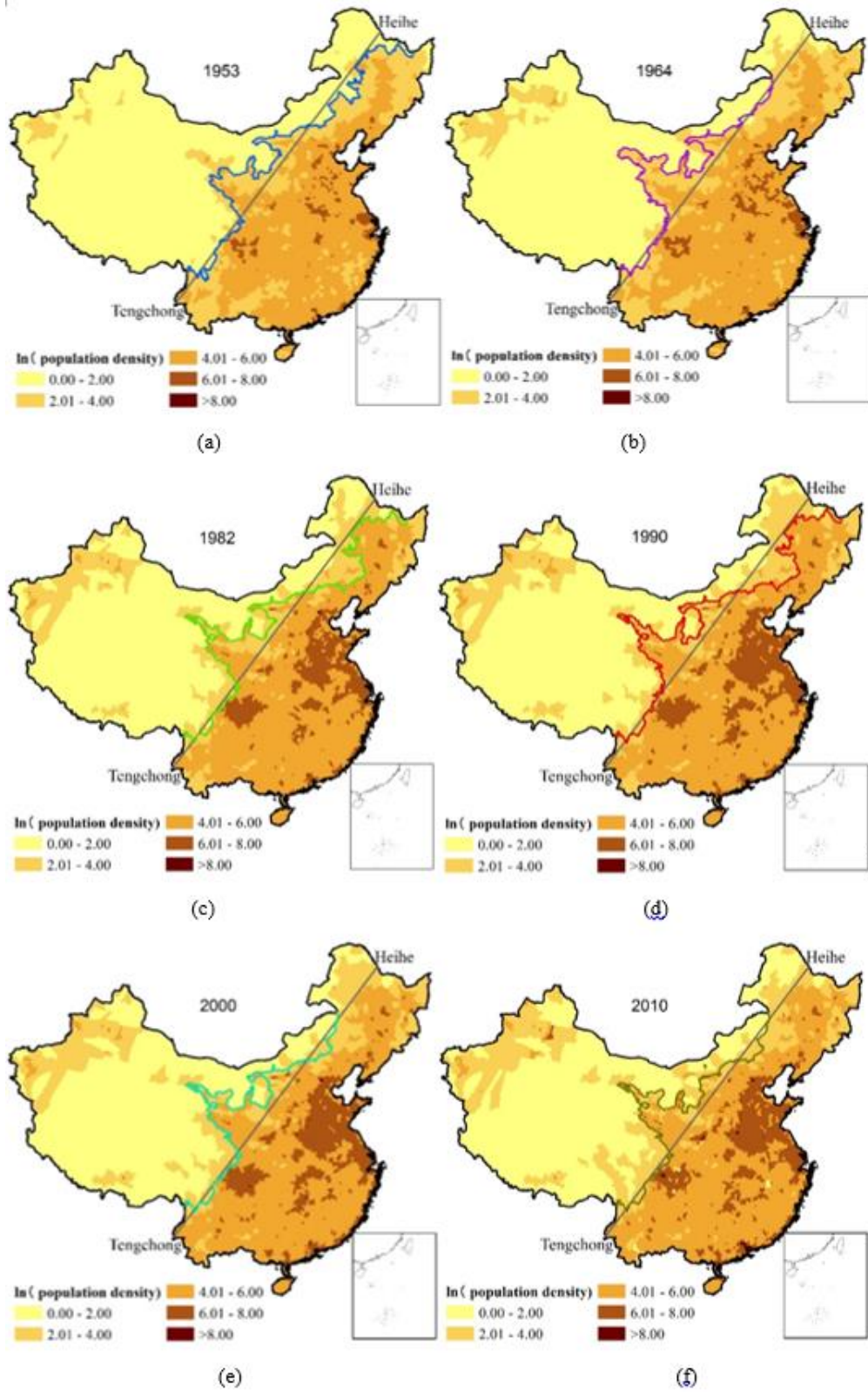
Figure 21. Two regions derived by REDCAP

Empirical studies on regional population density patterns including China (e.g., Wang 2001) suggest that population densities usually decline exponentially with distances from cities. By extension, population density varies exponentially across space. Therefore, population density in logarithm was used to measure attributive homogeneity in implementing the REDCAP regionalization method. As shown in Figure 21a-21f, all six REDCAP-derived lines were largely consistent with the Hu Line and relative stable over time. One noticeable discrepancy is an area of relative higher density near the midpoint of the Hu Line (Guanzhong Basin) that falls in the northwest region by the Hu Line but the southeast region by the REDCAP line.

Table 12. Comparison of two REDCAP-derived regions

Census years		Southeast region	Northwest region	Difference
1953	Population(mil)	567.3	11.4	555.8
	Population proportion (%)	98.03	1.97	96.06
	Area(km^2)	4308719.82	5142719.71	-833999.89
	Area proportion (%)	45.59	54.41	-8.82
	Population density(p/km^2)	131.65	2.22	129.43
	SSD		948.50	
1964	Population(mil)	679.1	11.9	667.2
	Population proportion (%)	98.3	1.7	96.6
	Area(km^2)	4921555.01	4529884.53	391670.48
	Area proportion (%)	52.07	47.93	4.14
	Population density(p/km^2)	137.98	2.63	135.35
	SSD		1104.71	
1982	Population(mil)	973.0	31.0	942.0
	Population proportion (%)	96.91	3.09	93.82
	Area(km^2)	4251784.34	5199655.19	-947870.85
	Area proportion (%)	44.99	55.01	-10.02
	Population density(p/km^2)	228.84	5.96	222.88
	SSD		1029.58	
1990	Population(mil)	1096.3	34.2	1062.1
	Population proportion (%)	96.97	3.03	93.94
	Area(km^2)	4276517.16	5174922.37	-898405.21
	Area proportion (%)	45.25	54.75	-9.5
	Population density(p/km^2)	256.35	6.61	249.74
	SSD		1034.87	

Census years		Southeast region	Northwest region	Difference
2000	Population(mil)	1214.0	28.4	1185.6
	Population proportion (%)	97.71	2.29	95.42
	Area(km^2)	4839160.69	4612278.85	226881.84
	Area proportion (%)	51.20	48.80	2.4
	Population density(p/km^2)	250.87	6.16	244.71
	SSD		1219.10	
2010	Population(mil)	1292.9	34.5	1258.4
	Population proportion (%)	97.40	2.60	94.8
	Area(km^2)	4799650.18	4651789.36	147860.82
	Area proportion (%)	50.79	49.22	1.57
	Population density(p/km^2)	269.37	7.42	261.95
	SSD		1351.82	

Table 12 summarizes the population and land area proportions of two REDCAP-derived regions over time. All the simulated lines divided the country into two regions with comparable area sizes but a striking difference in population. The regionalization for 1953, 1982, and 1990 yielded similar results as the southeast region smaller than the northwest (about 45% vs 55%). The regionalization for 1964, 2000, and 2010 yielded of near-identical area size. The expansion of area size in the southeast region was largely attributable to the significant increase in population density in the northwest corner in the Northeast Plain (credited to the land cultivation campaign by massive state-owned farms and more recently turned agricorporations in the region) (Zuo, 2014), and pushed the simulated lines toward the west. The population ratios in the southeast for all the six censuses were between 97:3 and 98:2. In other words, the two regions by the REDCAP method had a much stronger contrast in population than those by the Hu Line. As the simulated lines can better capture the detailed variability in population density than the straight Hu Line, they revealed the disparity of population density pattern than the Hu Line.

As measured in equation (2), the total sum of squared deviations (SSD) captures the overall heterogeneity in the derived regions. Overall, the SSD value increased from 1953 to 2010 with a

small peak in 1964. The increasing SSD value indicated enlarging disparity in population density across counties within the two regions (southeast and northwest regions) defined by the REDCAP method. This observation echoes the finding of a more polarized density pattern in China based the increasing Gini coefficients reported previously in Table 10.
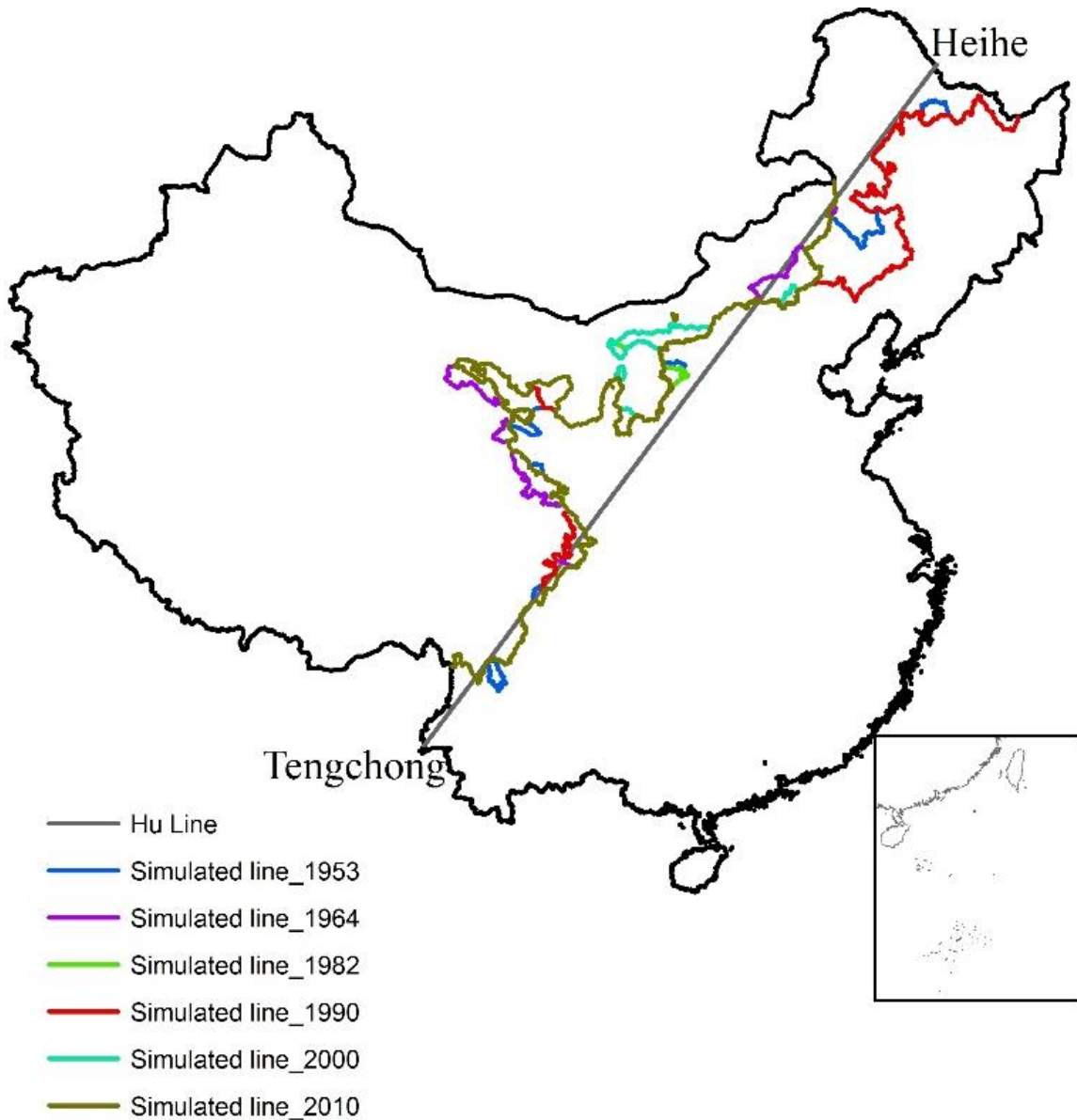

Figure 22. The overlapping of six simulated lines and the Hu Line

A closer examination on the movement of the simulated line over time, as shown in Figures

21a-21f and Figure 22, reveals changes that are more detailed. Instead of toward one direction, it moved back and forth. The following attempts to provide some insight on the underlying forces for the changes.

(1) For 1953-1964, the simulated line moved toward west, especially the northeast part. One likely cause was that immediately after the establishment of P. R. China, the central government followed the Soviet model in building up the heavy industry capacity. Much of the investment of industrialization in that era was concentrated in the northeast region, which brought a sizable migration there. The faster population growth rate in that area raised the population density there (also see Figure 18a), and helped push the demarcation line toward west.

(2) For 1964-1982, the simulated line moved to east, back to where the line stood in 1953. This period witnessed the fastest annual population growth in China (Table 9), but the uneven spatial distribution in growth rates (Figure 18b) in the northeast region increased the density gaps between the northwest corner and the rest of the region and shifted the delineation line eastward.

(3) For 1982-1990, the simulated line remained stable. After the economy reform and open door policy just took effect, the eastern coastal areas in China experienced benefit first and its economic growth (mainly in urban areas) attracted some population (Lu et al., 2005; He et al., 2016). However, population migration was at its infancy stage, and the change was not enough to move the demarcation line.

(4) For 1990-2000, the line moved to west in the northeast part. One likely cause was the success of economic reform became more pronounced in the coastal provinces, and the magnitude of migrant workers increased a great deal towards large urban centers in the

southeast for more job opportunities. This may help the expansion of higher-density areas in the southeast and push the line westward.

(5) For 2000- 2010, the demarcation line remained the same as the same trend in the previous period continued.

In addition, it is worthwhile to discuss the change of disparity within the southeast and northwest. As shown in Table 10 and Figure 20b-20c, the disparities within the two regions were much smaller than the whole country after much of the variability was already captured by the division between the two regions. Note that the southeast region always had a lower Gini value than the northwest. Within the southeast, the disparity exhibited a wavy variation (i.e., increasing in the beginning, then declining, finally increasing again). Within the northwest, the disparity showed an upward trend overall after 1963.

4.5 DISCUSSION AND CONCLUSION

This research examines the spatial disparity in population distribution in China since the first official census in 1953. The traditional Gini coefficient is used to examine the variability and disparity of population density at the county level. The GIS-automated regionalization method, REDCAP, is used to divide China into two regions by a line similar to the classic Hu Line, and examines the disparity of population patterns on the two sides. Both approaches are applied to the analysis of census data over time to detect temporal trends.

Based on the Gini values derived from the county-level population data, the study finds that the disparity in population density declined from 1953 to 1982, but the trend was reserved with increasing Gini value from the year 1982 to 2010. The former was largely attributable to major political movements that emphasized equal economic development during the period, and

the latter was a result of the economic reform and open door policy that led to the polarizing population settlement. The more recent trend since 1982 reflects the impact of urbanization that creates fast-growing urban areas on the one side and declining rural areas on the other side. The dualism of increasingly crowded cities and more deserted countryside is likely to stay and further exacerbate in the foreseeable future.

Our examination of population distribution indicates that the disparity in population between the northwest and southeast regions separated by the Hu Line shrunk from 1953 to 2010. The northwest region accounted for slight over 4% of the country's population in 1953 and increased to about 6% in 2010. By maximizing the difference in population density between two derived regions, we adopted the REDCAP method as a rigorous computation process to identify a demarcation (simulated) line in order to verify the classic Hu Line. This study has largely validated the Hu Line. The stability of the simulated line over time further supports the notion that this invisible barrier is tied to underlying factors such as physical environments suitable for human settlement, and is here to stay. The increasing SSD over time indicates an enlarged disparity within the southeast and northwest regions. In short, the trend in population distribution in China since 1953 has experienced a minor reduction in disparity between the two mega-regions (southeast and northwest) while the heterogeneity within each increased.

CHAPTER 5. CONCLUSION

The dissertation analyzes the population density pattern in China with the GIS-automated regionalization method, REDCAP. The research advances our understanding of the population distribution pattern in China, the relationship between the population distribution pattern and the physical environment pattern, and the population density disparity changes from 1953 to 2010.

5.1 SUMMARY OF MAJOR FINDINGS

The REDCAP regionalization is first used to reconstruct a demarcation line based on the 2010 county-level population density values. The results show that (1) the manually-sketched Hu Line is largely validated and refined by the rigorous scientific method of regionalization, and (2) the population disparity pattern in China has been relatively stable over the years despite some major historical events affecting population settlement. Furthermore, more regions are delineated to further advance our understanding of population distribution patterns in China.

In order to uncover the underlying physical environment factors that shape the aforementioned contrast in population density between the southeast and northwest of China, the dissertation proposes a habitation environment suitability index (HESI) model. The model integrates topographic factors, climatic suitability and hydrological condition into one comprehensive index, and then uses the REDCAP method again to derive another demarcation line based on the HESI values. The delineation lines on population density and HESI divide China into two regions that are largely consistent with each other. The result indicates that the population distribution disparity between the southeast and northwest is largely attributable to the difference in physical environments, and the barrier defined by the Hu Line is here to stay.

Finally, the dissertation uses the six censuses (1953, 1964, 1982, 1990, 2000, and 2010) at

the county level since the foundation of the People's Republic of China (PRC) to examine the changes of population density pattern in China over time. Based on the Gini coefficient, the change of disparity in population density at the county level followed a U-shape trend, i.e., decreasing from 1953-1982 and increasing from 1982-2010. The shrinking disparity in the pre-reform eras was largely attributable to various ill-conceived political movements, and the enlarging gaps in population growth rates in the post-reform era reflected a natural outcome of urbanization, which is ongoing and will continue in the near future.

5.2 FUTURE WORK

This dissertation mainly focuses on the population distribution disparity across different regions and over different time epochs in China. Future work will extend the research in at least three aspects.

(1) One major focus of this study is on linking the disparity in population distribution to underlying environmental factors in the HESI model. Future work will improve the HESI model by including more variables, such as land use type, economy, transportation etc., so the index will be more comprehensive.

(2) This study demonstrates the value of using the regionalization approach (here, REDCAP method) in examining disparities in population density in China. Future work can use other regionalization methods and examine the differences and possible causes.

(3) The methodology can be applied to different application domains (e.g., economic disparity) and to other countries (e.g., the U.S.).

REFERENCES

Alexander, J. W. (1948). The prewar population of China: distribution and density. *Annals of the Association of American Geographers,* 38(1), 1-5.

Attane, I. (2002). China's Family Planning Policy: An Overview of Its Past and Future. *Studies in Family Planning*, 33(1), 103-113.

Boluwade, A., Madramootoo, C., and Yari, A. (2016). Application of Unsupervised Clustering Techniques for Management Zone Delineation: Case Study of Variable Rate Irrigation in Southern Alberta, Canada. *Journal of Irrigation and Drainage Engineering,* 142 (1), Article Number 05015007.

Cassen, R. H. (1978). *India: Population, Economy and Society*. London: Macmillan Press, Ltd.

Chen, Z., Wang, J., Ma, G., Zhang, Y. (2013). China tackles the health effects of air pollution. *The Lancet*, 382(9909), 1959-60.

Cohen, J.E. and Small,C.(1998). Hypsographic demography: The distribution of human population by altitude. *Proc Natl Acad Sci USA*, 95(24), 14009–14014

Duque, J. C., Ramos, R., and Surinach, J. (2007). Supervised regionalization methods: a survey. *International Regional Science Review*, 30(3), 195–220.

Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics*. Cambridge University Press.

Everitt, B.S., Landau, S., and Leese, M. (2001).*Cluster Analysis*. 4th edn, London Arnold.

Feng Z., Tang Y., Yang Y., et al. (2007). The relief degree of land surface in China and its correlation with population distribution. *Acta Geographica Sinica*, 62(10), 1069–1082.

Feng, Z., Yang, Y., Zhang, D., et al. (2009).Natural environment suitability for human settlements in China based on GIS. *J. Geogr. Sci.* 19, 437-446

Gao, X, Wang, X, and Zhu, B. (2016). Distribution and Change of the Ethnic Minority Population in China: A "Hu Line" Perspective. *Population Research*, 40(3), 3-17.

Gao, Z, Liu, J,and Zhuang, D. (1999).The Relations Analysis between Ecological Environmental Quality of Chinese Land Resources and Population. *J. Remote Sens*, 3, 66–70.

Goodchild, M. F. and Lam, N. S.-N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geoprocessing,* 1, 297–331.

Gu, B., Wang, F., Guo, Z., and Zhang, E. (2007). China's Local and National Fertility Policies at the End of the Twentieth Century. *Population and Development Review,*33(1),129-148

Guo, D. (2008).Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science archive,* 22(7), 801-823.

Guo, D. and Wang, H. (2011). Automatic region building for spatial analysis. *Transactions in GIS,* 15(s1), 29–45.

Guo, H., Wang, X., Wu, B., Li, X. (2016). Cognizing Population Density Demarcative Line (Hu Huanyong-Line) Based on Space Technology. *S&T and Society 31(12)*, 1385-1394 (in Chinese).

He, C., Chen, T., Mao, X., et al. (2016). Economic transition, urbanization and population redistribution in China. *Habitat International*, 51, 39-47.

Hu, H. (1935). Distribution of China's population: Accompanying charts and density map. *Acta Geographica Sinica*, (2), 33-74 (in Chinese).

Hu, H. (1990). The distribution, regionalization and prospect of China's population. *Acta Geographica Sinica*, 45(2), 139-145(in Chinese).

Huang, R. (2005). Population Redistribution in the Context of Urbanization in Beijing Since 1980s. *Population Research*. (5), 19- 26.

Jin, X., Zhang, L., Gu, J., Zhao, C., Tian, J., He,C. (2015). Modelling the impacts of spatial heterogeneity in soil hydraulic properties on hydrological process in the upper reach of the Heihe River in the Qilian Mountains, Northwest China. *Hydrological Processes*, 29(15), 3318–3327.

Kumar, M. (2015). A Study of Population Distribution. *IJLTEMAS*, 4(3),24-28

Kupfer, J. A., Gao, P., and Guo, D. (2012). Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning. *Ecological Informatics,* 9, 11-18.

Li, Y., Liu C., Zhang, H., et al. (2011). Evaluation on the human settlements environment suitability in the Three Gorges Reservoir Area of Chongqing based on RS and GIS. *J. Geogr. Sci,.*, 21(2), 346-358

Liu, Y., Liu, Y., and Zhai, R. (2009). Geographical Research and Optimizing Practice of Rural Hollowing in China. *Acta Geographica Sinica*. 64 (10), 1193 – 1202.

Long, H., Li, Y., Liu, Y., Woods, M. and Zou, J. (2012) Accelerated restructuring in rural China fueled by'increasing vs. decreasing balance' land-use policy for dealing with hollowed villages. *Land Use Policy*, 29, 11–22.

Long, Y. and Wu, K. (2016). Shrinking cities in a rapidly urbanizing China. *Environment and Planning A,* 48(2) 220–222

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9 (70),209–219

Lu, D., Wang, Z., Feng, Z., et al. (2016). Academic debates on Hu Huanyong population line. *Geographical Research*, 35 (5), 805-824

Lu, Z., Huang, Q. H., Lu, T., et al. (2005). Process and problems of industrialization and urbanization in China. *China Industrial Economy*, 12, 5-13

Mao, Q., Long, Y.,Wu, K. (2016). Spatio-Temporal Changes of Population Density and Urbanization Pattern in China (2000 – 2010). *China City Planning Review,* 25(4),8-14

Qi, W., Liu, S.,and Zhao, M. (2015). Study on the stability of Hu Line and different spatial patterns of population growth on its both sides. *Acta Geographica Sinica*, 70(4), 551-566

Shan, Z. (2009).China's mutant line: Hu Line. *Chinese National Geography*, (10), 288-301. (In Chinese)

Staszewski, J. 1967. *Vertical Distribution of World Population*. Scientific Publishing House

State Council of China. (2008). Coordinated Regional Development Strategy and Policy (in Chinese). Report by the Development Research Center of the State Council, Beijing

Stutz, F.P.,and Warf, B. (2007). *The World Economy: Resources, Location, Trade, and Development*. Pearson Education Incorporated

Sun, Z. (2013). Explaining regional disparities of China's economic growth: Geography, policy and infrastructure. Berkeley, CA: University of California.

Sutton, P., Roberts, D., Elvidge, C.D. and Meij, H. (1997). A comparison of nighttime satellite imagery and population density for the continental United States. Photogrammetric Engineering and Remote Sensing, 63, 1303–1313

The World Bank. (2007).Cost of pollution in China: economic estimates of physical damages. The World Bank, Washington, DC

Thom, E. C. (1959). The discomfort index. *Weatherwise,* 12: 57–60

Wang, F. (2001). Regional density functions and growth patterns in major plains of China, 1982-90. *Papers in Regional Science,* 80, 231-240

Wang, F. 2009. Factor analysis and principal-components analysis, in *International Encyclopedia of Human Geography*, Vol. 4 (editors-in-chief: Rob Kitchin and Nigel Thrift), Oxford, UK:

Elsevier, 1-7

Wang, F. 2015. *Quantitative Methods and Socioeconomic Applications in GIS* (2nd ed.) Boca Raton, FL: CRC Press

Wang, F. and Robert, L. K. (2015). Constructing Geographic Areas by REDCAP and MLR for Analysis of Homicide Rates: A Case Study of New Orleans, Louisiana. *Papers in Applied Geography,* 1, 295-306

Wang, F., Guo, D., McLafferty, S. (2012).Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois. *Applied Geography*, 35(1-2),1-11.

Wang, G. X. and Pan, Z. H. (2016). The robustness of China's migration and Heihe-Tengchong line. *Chinese Journal of Population Science*, (1), 2-13, 126. (In Chinese)

Wang, K. and Deng, Y. (2016). Can new urbanization break through the Hu Huanyong Line? Further discussion on the geographical connotations of the Hu Huanyong Line. *Geographical Research*, 35 (5), 825-835

Wang, K., Li, Y., and Ding, J. (2016). The influence of Chinese population policy change on resources and the environment. *Chinese Journal of Population Resources and Environment*, 14(4), 227-234

Wang, M., Dong, Z., Cheng, Y., et al. (2018). Optimal Segmentation of High-Resolution Remote Sensing Image by Combining Superpixels With the Minimum Spanning Tree. *IEEE Transactions on Geoscience and Remote Sensing,* 56, 228-238

Wang, P., and Na, J. (2008). Analysis of Yield Components and Meteorological Factors Affecting Yields in Heilongjiang Province. *Meteorol. Sci. Tech*, *36*, 449–452

WHO. (2013). Country profile of environmental burden of disease: China

Xu, P., Huang, H., Dong, N., et al. (2014). Sensitivity analysis in the context of regional safety modeling: Identifying and assessing the modifiable areal unit problem. *Accident Analysis and Prevention,* 70, 110-120

Xue, J. (1997).Urban-rural income disparity and its significance in China. *Hitotsubashi Journal of Economics*, 38(1), 45-59

Yang, X., and Ma, H. (2009). Natural Environment Suitability of China and Its Relationship with Population Distributions. *Int. J. Environ. Res. Public Health*, 6(12), 3025-3039

Yu, F., Ma, G., Qi, J., et al. (2012). *Report of China's environmental-economic accounting in 2007–2008*. China Environmental Science Press, Beijing

Yuan, H. (1993).Study on the rationality of Chinese population distribution. *Geographical*

Research, 12(3), 64-69

Zhuo, L., Ichinose, T., Zheng, J., Chen, J., Shi, P. J., and Li, X. (2009). Modelling the population density of China at the pixel level based on DMSP/OLS non-radiance calibrated night-time light. *International Journal of Remote Sensing*, 30(4), 1003-1018

Zuo, M. (2014). China turning state-owned farms into agricorporations to take on world players. South China Morning Post (dated Aug 14, 2014). Available https://www.scmp.com/news/china/article/1573077/china-turning-state-owned-farms-agricorporations-take-world-players (last accessed 9-17-2018).

VITA

Cuiling Liu was born in Handan, Hebei province, China. She received her B.S. degree in GIS from Shandong Jiaotong University, Jinan, Shandong, in June 2010 and her M.S. degree in Cartography and GIS from Capital Normal University (CNU), Beijing, in June 2013. After that, she worked as an assistant engineer in the Ministry of Housing and Urban-Rural Development, China, for one year. In 2015, she came to Louisiana State University (LSU) to pursue her Ph.D. degree in Geography. Her research interests include GIS and spatial analysis of population, urban simulation, and social disparities.