SPATIAL ANALYSIS OF ETHNIC MIGRATION BEHAVIOR: A CASE STUDY OF CHINESE IMMIGRANTS IN THE NEW YORK-NEWARK-JERSEY CITY METROPOLITAN AREA

By
YANXIA WU
Bachelor of Science in Civil Engineering
Yanshan University
Qinhuangdao City, Hebei Province, China
2005

Master of Science in Civil Engineering
Beijing University of Technology
Beijing City, China
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2021

SPATIAL ANALYSIS OF ETHNIC MIGRATION BEHAVIOR:
A CASE STUDY OF CHINESE IMMIGRANTS IN THE NEW YORK-NEWARK-JERSEY CITY METROPOLITAN AREA

Dissertation Approved:

Dr. Jonathan C. Comer

Dissertation Adviser

Dr. Hongbo Yu

Dr. Thomas A. Wikle

Dr. Andrew S. Fullerton

ACKNOWLEDGEMENTS

This dissertation is a process I meet with amazing people. They have been sweetening up this research process. It is my honor to have their names in my dissertation.

Dr. Jon Comer, my advisor, is fast in statistical thinking, talking, walking, and doing things. His office was the place I went to for everything. I had to check if his office door was open before starting my daily work. It began to be my habit. After moving to Coeur d Alene, I receive such support from emails. I felt supported when my advisor said, "WE can do this… WE need to do that." Just like I had to run to catch up with his walking speed, I need to stay sharp in my field to keep up with him.

Dr. Hongbo Yu smiles with his eyebrows. His professional GIS training inspires me. Dr. Yu is such a warm person that I took advantage of his open-door policy so much. He pointed out the direction of light when all I saw was darkness in my transition to life in a foreign country. I am deeply appreciative of Dr. Thomas Wikle's push on demographics study. I am not only a tech person. I frequently refer back to my class note from Dr. Andrew Fullerton's quantitative analysis in social studies. I had a pretty pretty pretty good time in his class. Also, I would like to give lots of thanks to Gabe de la Cruz, an assistant professor at North Idaho College. You hear his laughter before seeing his face. Gabe is very good at explaining codes clearly and fun and making people addicted to coding.

I also got a lot of help and advice from friends at OSU and NIC. Many thanks to Michael Larson! He is as sweet as Santa (whose job is to bring people happiness). Many thanks to Clay Barrett! Nothing seems could bother him. Many thanks to my friends Arlene Pan and Er Yue! Many thanks to Emily Williams! Also, many thanks to those genius classmates in my C++ class, Logan Rostron and Joshua Bohannan.

I also want to thank my family. If I were a mathematical curve, my husband, Nick Rose, would be the graph's x-axis. He holds me up when my emotions go up and down. Thanks to my daughter Margaret. There were 1,000 times I said no to her request: "You wanna play?" But there was always a 1,001 time she came back to me. Thanks to my parents. They let me fly. Thanks to my brother, who spent time with my parents when I was not around.

Name: YANXIA WU

Date of Degree: JULY, 2021

SPATIAL ANALYSIS OF ETHNIC MIGRATION BEHAVIOR
A CASE STUDY OF CHINESE IMMIGRANTS IN THE NEW YORK-NEWARK-JERSEY CITY METROPOLITAN AREA

Major Field: GEOGRAPHY

Abstract

The population of Chinese immigrants in the United States has undergone progressive growth in the past 50 years and has reached an epidemic number. As minorities, the Chinese immigrants move into receiving places to adapt and succeed in a new social structure while not losing their own identity. Previous studies highlight the role of local contexts that lead to an internal moving decision. Most of these studies view local contexts as global factors assumed to apply equally over a study area. However, the contextual factors do not disperse evenly across space, nor their relationships with migration behavior. Understanding the spatial variability of factors related to Chinese people's migration in the study area is necessary. Therefore, this dissertation aims to explore the role in which neighborhood context may predict migration behavior, with particular attention to how migration factors and their effects vary across space. This research presents novel applications of two methods: clustering analysis (followed by regression models) and multiscale geographically weighted regression (MGWR) to the Chinese population in the New York-Newark-Jersey City metropolitan statistical area as a case study. Besides regression analysis, this research also provides a detailed examination of relationships between micro-level factors using decision tree analysis. Wages, education, English proficiency, and self-employment status are crucial variables in differentiating movers from non-movers. Having naturalized citizenship has a dual effect on migration behavior. Among the movers, stratifications exist in the immigrant Chinese population. Each subgroup has its particular migration pattern and significant indicators. Spatial variations exist in the study area. Neighborhood type 2 (low in socioeconomic and stable status) is the residential place for immigrants from other states. And neighborhood type 1 (high in socioeconomic and stable status) has more within-state immigrants. Regression models accounting for the population stratification and spatial variations have a vast improvement over the OLS model. Approaches considering data associations in both geographic dimension and non-geographic dimensions could be promising.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

1.1 Introduction

Over the past 60 years, the size of the immigrant population[1] in the United States has increased dramatically, accompanied by changes in its origins and impacts. The total immigrant population grew steadily from 9.7 million in 1960 to 44.9 million in 2019, while its percentage of the total population grew from 5.4 to 13.7 (USAFACTS 2021a; USAFACTS 2021b). In the 1960s and 1970s, European countries were the primary origins for the immigrant population. Today, Mexico, China, and India are the top three origin countries of immigrants in the US (Budiman 2020). With a total population of 4.4 million in 2019, Chinese immigrants (including people from Hong Kong and Taiwan) to the United States ranked second in the immigrant population of this country, outnumbered immigrants from India (Budiman 2020; US Census Bureau 2019).

Chinese immigration to the US can be divided into three periods (Poston Jr and Luo 2007). The first period of immigration happened between 1849 and 1882, following the California Gold

[1] Foreign born is anyone "who is not a US citizen at birth. This includes naturalized US citizens, lawful permanent residents (immigrants), temporary migrants (such as foreign students), humanitarian migrants (such as refugees and asylees), and unauthorized migrants" (US Census Bureau 2020, para 3). Native and native born refers to "anyone born in the US, Puerto Rico, a US Island Area, or abroad of a US citizen parent or parents" (US Census Bureau 2020, para 5).

Rush. The Chinese were mainly employed in San Francisco as miners and railroad workers. Some Chinese still call San Francisco "Jiu Jin Shan" (Old Gold Mountain). In 1882, the passage of the Chinese Exclusion Act represented the beginning of the second immigration period. Concerned about the competition from the Chinese in the US workforce, the government banned laborers from China. During that period, the number of immigrants from China was very small. Diplomats, merchants, students, and their dependents were the main immigrant types (Poston Jr and Luo 2007). After the Immigration and Nationality Act of 1965, the US government passed reopening immigration from China (Poston Jr and Luo 2007). Since then, the Chinese population has undergone tremendous growth in the US, reaching 4.4 million in 2019. Its contribution to the US immigrant population rose from 1% in 1960 to 9.8 % in 2019 (Figure I-1)[2].

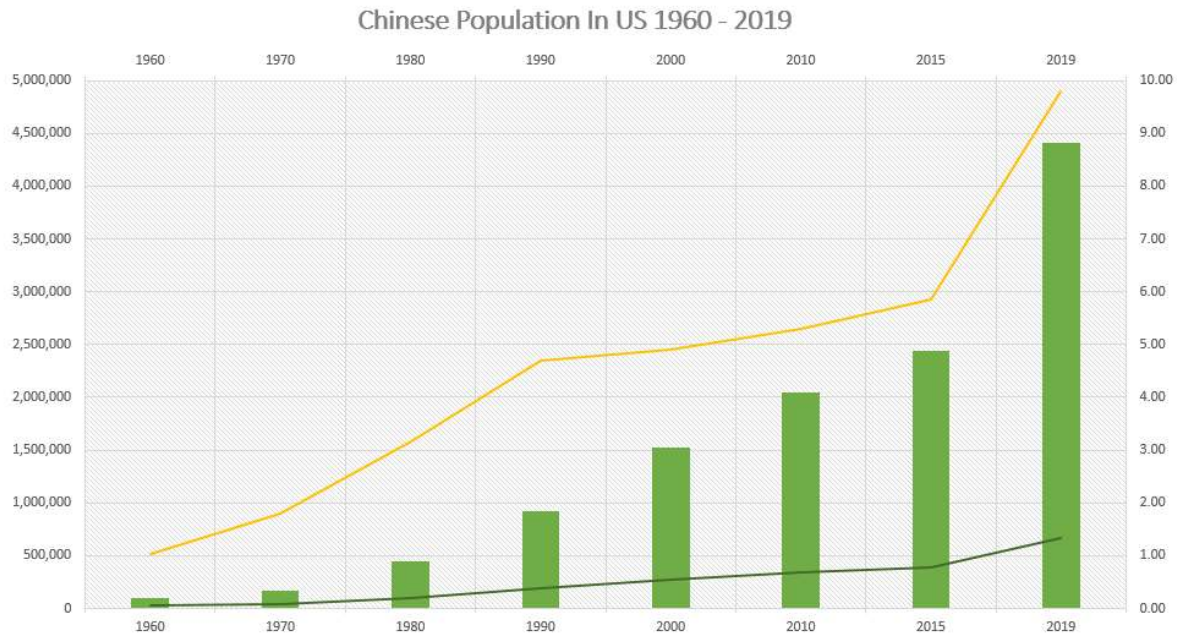


Figure I-1 Chinese population in the US (1960 - 2019)

The percentage changes of the Chinese population within the total US immigrant population and the total US population can be seen on the two polylines. The Chinese population in the US is denoted by

[2] Data between 1960 to 2000 are from Gibson and Jung 2006. Data from 2000 to 2015 are from US Census Bureau: ACS 5-year estimates. 2019 Data are from ACS 1-year estimates, USAFACTS 2021a and 2021b.

Pc. The percentages of the Chinese population within the total US immigrant population and the total US population are denoted by *Pc/Pi* (*%*) and *Pc/Pt*(*%*).

With the growth of the Chinese immigrant population in the US, different cultures, religions, and lifestyles were introduced to the US. Successful cultural integration is vital in keeping the social cohesion of receiving places, especially large port-cities at the macro-level (Kaplan and Douzet 2011; Penninx 2003). Government policies and planning strategies towards immigrants ground their experiences (inclusion or exclusion) at the micro-level (Chacko and Price 2020). The migration of the Chinese population is a dynamic process involving interactions of factors at diverse scales. To fully understand the experiences of the immigrants, one needs to thoroughly investigate the resources of ethnic migration.

Due to the complex nature of ethnic migration, it is difficult to identify all factors related to migration behavior. However, there are common factors that lead to an internal moving decision. Previous studies highlight the role of neighborhood contexts in a migration decision. Location-specific amenities include economic conditions (e.g., unemployment rate and employment growth rate) (Gurak and Kritz 2000; Kritz et al. 2011; Kritz and Nogle 1994; Liu and Painter 2012), urban infrastructure (e.g., transportation) (Yu 2018; Zhou 1998), social programs (e.g., daycare and senior citizen services) (Zhou 1998), and housing market (Lee 2018; Zhou and Logan 1991). Chinese immigrants have added ethnic economy and social structure to the neighborhood contexts in their integration process into a new place (Chen 2017; Ling 2005; Zhou 1998; Zhou 2004; Zhou and Lin 2005; Zhou and Logan 1991).

Ethnic migration displays non-random geographic distributions across a study area. The contextual factors do not disperse evenly across space, nor their relationships with migration behavior. Research highlights the significance of locality in a migration study in terms of the immigrant geography of residence (Allen and Turner 2009; Logan et al. 2011; Walton 2017; Yu 2018) and workplace (Liu and

van Holm 2019; Schuch and Wang 2015; Wang 2010; Zhou 1998). With these factors changing across space, their impacts on the migration behavior of the Chinese population vary from one place to another. The assimilation process could be different for people with similar demographic profiles if their local contexts are different (Newbold 2010; Newbold and Foulkes 2004).

However, most studies only account for the spatial nature of the research problems or view the contexts as global factors applied equally over space. They mask the spatial variations of relationships or ignore the scale dimension of processes in conceptualizing a space. Geographic methods which account for spatial variations will reveal patterns and relationships that are undetectable with other tools. Therefore, it is necessary to add a geographic perspective to the migration study of Chinese immigrants.

1.2 Research Questions

Although migration behavior is an important topic in human geography and social studies, a universal definition of migration behavior that could be applied to various places and populations does not exist (Sinha 2005). With the background described in the previous section, I will use immigrant Chinese in the New York-Newark-Jersey City metropolitan statistical areas as a case study to help define migration behavior. Migration behavior is a change of residence (permanent or semi-permanent), through which people adapt to changes or strive to succeed in their lives. In this study, a migration pattern specifically refers to whether a Chinese immigrant was from a different country, a different state, or within the state. The movement is a dynamic process including factors in demographics, social and economic conditions. These factors constitute the contexts of migration behavior. In this study, I will examine for each migration pattern where Chinese immigrants move to (in terms of neighborhood types), what the migration-related factors are, and how these factors impact or indicate migration behavior.

Two classic theories explain migration patterns and contextual factors of ethnic populations: spatial assimilation theory and ethnic enclave theory. From the perspective of spatial assimilation theory, though ethnic minorities enter into a receiving country at a disadvantage, with the increase of their socioeconomic status, they eventually move to residential locations that match their social levels (Massey and Mullan 1984; Yu 2018). Based on ethnic enclave theory, immigrants tend to live in their ethnically distinct community in the host country due to the drawing from ethnic resources (Forbes 1984; Fang and Brown 1999; Lobo and Mellander 2020; Zhou 2010). There are also other theories, such as resurgent ethnicity and heterolocalism. Resurgent ethnicity theory re-evaluates the role of ethnic enclaves. People chose to stay in the ethnic concentrations even when they can move to areas matching their improved socioeconomic status because they may lose more than gain if moved (Walton 2012; Wen et al. 2009). Heterolocalism suggests that ethnic populations chose a dispersed residential pattern with their co-ethnics while staying in touch in various ways, such as through ethnic associations (Mukherjee and Pattnaik 2020; Zelinsky and Lee 1998).

In efforts to test the above theories, this study aims to explore the role in which neighborhood contexts may predict migration behavior, with particular attention to how migration factors and their effects vary across space. To achieve the goal, I will address three research questions.

Question 1. What contextual factors may affect a Chinese immigrant's migration behavior in the New York-Newark-Jersey City MSA?

Question 2. What are the distribution patterns and characteristics of each neighborhood type in the New York-Newark-Jersey City MSA?

Question 3. How do local factors in the New York-Newark-Jersey City MSA impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially?

The three research questions build upon each other to help guide my research. This study will provide a systematic evaluation of factors related to the ethnic migration of Chinese immigrants.

Neighborhoods are the spatial units of migration contexts. I will identify the main neighborhood types in the study area from the eyes of Chinese immigrants. Different neighborhood types reflect the spatial variability of locational attributes formed by the constant inflow of immigrants over the long run. Nevertheless, I will model the processes or relationships between diverse migration patterns and neighborhood contexts, in which the scale of each process will be determined. In summary, this study will broaden and deepen the application of geography in ethnic migration studies by identifying critical processes, their scales, and spatial variations in Chinese people's migration.

1.3 Chinese immigrants in the New York-Newark-Jersey City MSA

To examine the role of geography in migration behavior, I focus on the Chinese immigrants in the New York-Newark-Jersey City (NY-N-JC) metropolitan statistical area (MSA). It is one of the oldest and well-established ethnic Chinese concentrations in the US (Fang and Brown 1999). Chinatowns of different sizes have developed in US history. Manhattan's Chinatown was the first Chinese community on the east coast of the US, started in the late 1800s. After the passing of the Immigration and Nationality Act of 1965, Manhattan's Chinatown grew speedily, expanding north into the Little Italy area and southeast into the Lower East Side. It was the largest Chinatown in the US from the 1980s until recently replaced by Brooklyn Chinatown (Ostrow and Ostrow 2008). Although the first groups of Chinese people were Cantonese and Fujianese, Manhattan's Chinatown is more diverse today, with people from different Chinese provinces and other Asian countries. Mandarin became the dominant language in the area, replacing Cantonese, a dialect from southern China (Semple 2009). Because Manhattan's Chinatown is next to the World Trade Center, its development slowed down significantly after the September 11 attacks in 2001 (CGTN America 2019).

Flushing's Chinatown in the borough of Queens in New York City was first a satellite community of Manhattan's Chinatown. In the 1970s, a wave of Taiwanese arrived in New York City. They spoke Mandarin and had a relatively high educational attainment and socioeconomic status. They chose to

reside in the Flushing area several miles east of Manhattan instead of the increasingly crowded Chinatown in Manhattan. Moreover, the dominant language in Manhattan's Chinatown was Cantonese, a language obstacle to the Taiwanese (Julia 2010; Min 2006). Compared to Manhattan, houses were more affordable in the boroughs of Queens and Brooklyn. Flushing's Chinatown has attracted new immigrants from different Chinese provinces and even multiple ethnic groups (Zhou 2010; Chen 2017). The landscape in Flushing New York City has become more diverse and is still in constant transformation (Yu 2018).

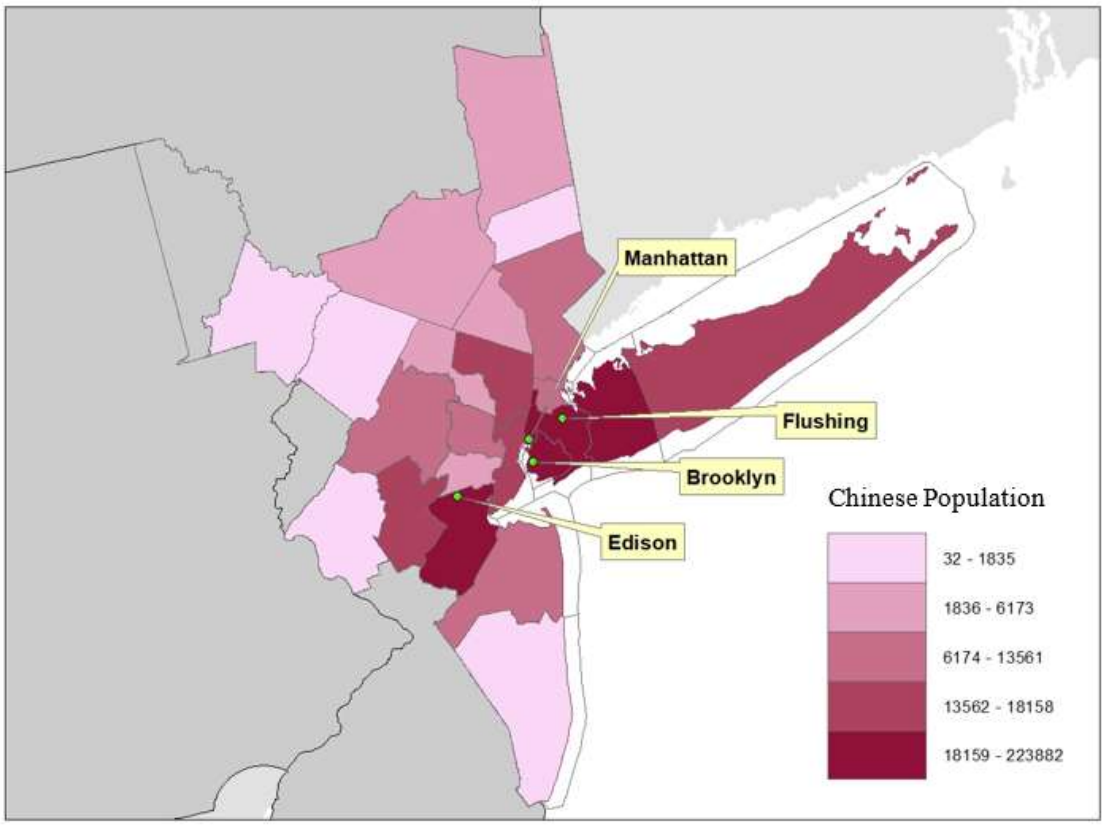


Figure I-2 Chinese population by county

Today, the biggest Chinese enclave is on Eighth Avenue in the borough of Brooklyn. The community began to grow in the 1980s at Sunset Park. Because of its extensive mass public transportation and affordable storefronts and houses, this area skyrocketed to the top populated Chinese enclave along the east coast, surpassing Manhattan's Chinatown (Beekman 2011). Besides the above big three Chinatowns, Chinese hubs continue to form around the old Chinatowns. East Harlem is one in

Manhattan (McGlinn 2002). Several places started to grow as Flushing's satellite Chinatowns, such as Elmhurst, Corona, Whitestone, and Nassau County (McGlinn 2002; Roleke 2019). In Brooklyn, new stores continue to grow in neighboring areas, such as Bensonhurst and Sheepshead Bay, forming dispersed Chinese communities (Robbins 2015).

Compared to their counterpart in New York City, Chinese neighborhoods are loosely dispersed in the suburban areas around Edison in New Jersey (McGlinn 2002). These areas attracted the Chinese for good educational systems, high ratios of professionals, and safe and affluent neighborhoods. Commuter rail offers easy access to Manhattan, which is a benefit of the area around Edison as well. Ethnic-related businesses moved in quickly with the growth of the population (McGlinn 2002). All these elements stimulate the continued formation of new Chinese concentrations.

Public use microdata areas (PUMAs) offer the primary geographic information of people in the NY-N-JC Metropolitan statistical area (MSA) (Figure I-3). An MSA is an urbanized area of at least 50,000 people and surrounding areas with strong social and economic ties with the core (US Census Bureau 2016). The definitions of their physical boundaries are based on counties or county equivalents. This study evaluates contextual effects on the Public Use Micro Area (PUMA) geographical scale. The New York-Newark-Jersey City metropolitan area encompasses 160 PUMAs. Among them, 98 are in New York, 2 in Pennsylvania, and 60 in New Jersey. The boundaries of PUMAs are constructed based on census tracts and counties. Each PUMA has at least 100,000 people (US Census Bureau 2021).

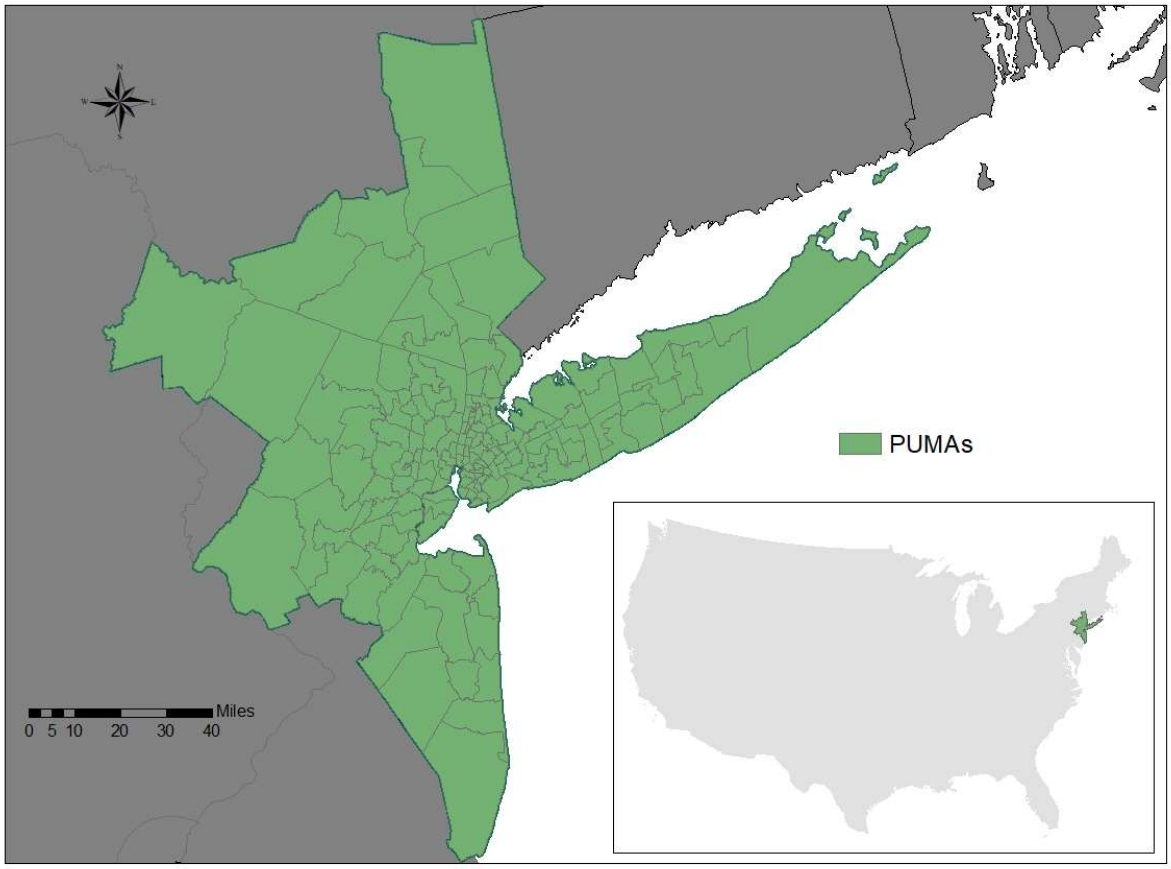


Figure I-3 PUMAs in New York-Newark-Jersey City MSA

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Migration studies examine the interactions between locational attributes and an individual's behavior around the issue of migration. To gain a complete picture of migration behavior, one needs to study the formation of locations and residents' migration behavior. In the following sections, I will first review the historical formation of Chinese ethnic enclaves in the study area. They are the key to understanding the ethnic enclaves. Next, I will provide a review of migration-related factors of Chinese immigrants in previous research. By viewing these factors, particularly those pertinent to Chinese immigrants, I can then continue with a systematic evaluation of the factors and their influences on migration in the subsequent analysis. After summarizing the historical formation of the Chinese enclaves and migration-related factors, I will review the spatial dimension pointed out in previous research and show how this dissertation can deepen the application of geography in ethnic migration study.

2.2 The formation and characteristics of Chinese ethnic enclaves

Ethnic enclaves are residential concentrations and sometimes business districts for ethnic populations (Barabantseva 2016; Li 1998). In the US, Chinese ethnic enclaves are commonly referred to as Chinatowns. Residents in Chinatowns have easy access to their home-country food from Chinese supermarkets. Many of them obtain their daily information via newspapers, radio stations, and TV stations in Chinese languages (Chen 2017; Li 1998; Shircliff 2020). The formation and evolvement of Chinese ethnic enclaves are greatly influenced by political policies and economic events in the US history of Chinese immigrants. While those significant historical events constitute contextual factors of the migration decision of a Chinese immigrant, his/her demographic profile plays a crucial role in the process as well.

Chinese immigrants in a city generally follow the trend of moving from city centers to peripheries and then into suburbs (Alba et al. 1999; Liu and Painter 2012). In the 19th century, early Chinese immigrants settled in inner cities where traditional Chinatowns started to form. On average, those Chinese immigrants were poor and less-educated (Li 1998; Bai 2015; Hooper and Batalova 2015). Since the 1960s, some Chinese moved out of downtown areas and into suburbs as they sought better housing and communities that matched their increased socioeconomic status. Around the same time, some new Chinese immigrants moved directly into the suburbs. Those new immigrants had distinct characteristics from early Chinese immigrants. Instead of being poor and less educated, many were well-educated and wealthy (Li 1998; Bai 2015; Hooper and Batalova 2015). With the growth of the Chinese population in those suburban areas, new ethnic communities formed, which were labeled ethnic suburbs (Li 1997; Li 1998; Lin and Robinson 2005; Zhou and Lin 2005).

This new type of ethnic community has formed as the interplay between economic and political contexts at international, national, and local scales (Chacko and Price 2020; Li 1998; Yu 2018). Global economic restructuring bred new configurations for national economies. Some major shifts

were: reindustrialization of craft business, the rapid expansion of service business, and the formation of multinational corporations (Scott 1988; Storper and Walker 1989; Davis 1992; Dymski and Veitch 1996). The above changes generated labor demands for entrepreneurs and investors, high-skilled professionals, and low-skilled laborers. On the national level, the changes in US economic and immigration policies and quotas allowed immigrant streams to flow into the US and became a part of its social and economic contexts on a local scale (Light and Bonacich 1988; Grayson 1995).

Along with these economic moves were political changes. Before 1965, Chinese immigration to the US was minimal. After the passage of the 1965 Immigration Act, many Chinese immigrants started to move into the US due to relaxed immigration standards (Bai 2015). In 1972, President Nixon visited China, which brought about another increase in Chinese immigration to the US from mainland China and Taiwan. The period between 1970 and 1979 saw tremendous growth of immigrants from Taiwan into the US. In 1984, the United Kingdom (the UK) agreed to return Hong Kong to China, which formally took place in 1997. After that, Chinese immigration from Hong Kong to the US expanded rapidly (Li 1998). Ethnic Chinese emigrating from Indochina to the US increased during and after the Vietnam War, with most immigrants leaving their home countries to flee the war. Migrants from Indochina were of low socioeconomic status and mainly had little education (Liu and Cheng 1994; Ong and Liu 1994; Allen and Turner 2005). Coming from diverse origins and migration statuses, Chinese immigrants in the US have created layers and stratifications among Chinese ethnic communities while building their ethnic identity as a whole.

Compared to their predecessors, modern Chinese immigrants generally have a relatively high socioeconomic status. Most of them are well-educated, reside in good housing conditions, and have a high salary (Lee 2018; Li 1998; Bai 2015; Hooper and Batalova 2015; Kadarik 2019). Education has enormous weight in Chinese traditional value systems. It is the common goal for parents, teachers, and students to be accepted into a top university. Many parents are willing to sacrifice their lifestyles for their children's education. A good education, according to traditional Chinese value systems, leads

to a bright future. Chinese immigrants bring this traditional value about education to the US. Getting access to a good education becomes one main driving force of Chinese migration (Li 1998; Bai 2015; Hooper and Batalova 2015; Kadarik 2019). To own a house is viewed as a sign of achieving success, especially for immigrants. A majority of Chinese immigrants in ethnic communities own their houses. For example, in Los Angeles, Chinese homeownership is higher than average at the county level and above all other immigrant groups (Li 1998; Bai 2015; Hooper and Batalova 2015; Emeka 2020).

Income levels of modern Chinese immigrants in ethnic communities are higher than the average or other ethnic populations (Lee 2018; Li 1998; Bai 2015; Hooper and Batalova 2015). Within the Chinese immigrant group, income levels vary by occupation. Self-employed people have the highest income, followed by professionals. The high income is due to the job industries. For example, in Los Angeles County, many Chinese immigrant workers are employed in the finance, insurance, and real estate industries (FIRE). At the same time, lower-income labor is involved in industries such as furniture, food processing, and especially the garment industries (Light and Bonacich 1988; Li 1998).

Workplace concentrations exist in immigrant populations (Andersson et al. 2014; Bagchi-Sen et al. 2020; Lobo and Mellander 2020). Previous studies suggest strong ties between ethnic residential patterns and business concentrations (Ellis et al. 2007). Wang (2006) also suggested that certain residential choices, such as living in areas with a large ethnic population, promote niche employment. Li (1998) observed that two-thirds of the Chinese population in ethnic communities of Los Angeles were professionals or managers. Self-employed people usually assigned themselves as managers. Scholars also showed that many Chinese immigrants work in ethnic service jobs, such as restaurants and grocery stores (Bagchi-Sen et al. 2020; Ellis et al. 2007; Li 1998). Ethnic business concentrations play an active role in energizing ethnic neighborhoods (Schuch and Wang 2015). Conversely, ethnic residence clustering has the potential to promote ethnic business development (Kaplan 1998). However, Liu and Painter (2012) observed that most job growth happened in native-born white

concentrations. The employment opportunities attract new immigrants away from areas with high minority populations.

In summary, concerning demographic and socioeconomic characteristics, spatial variations exist across neighborhoods. Compared to the native population, Chinese immigrants have a unique population structure regarding age, sex, marital status, and education levels. These elements differentiate Chinese neighborhoods from other neighborhoods without an apparent ethnic identity. Moreover, the Chinese population is heavy in some industries, featuring both high-skilled and low-skilled labor. With ethnic-specific characteristics such as English proficiency and residing years in the US, all the above factors lead to a population profile specific to Chinese immigrants. Nevertheless, stratifications exist within the subgroups of Chinese immigrants of different origins. Ethnic stratifications lead to different residential choices (Kadarik 2019). When the Chinese population migrates into the US, their characteristics gradually shape the neighborhoods where they live. On the other hand, neighborhoods provide the environment in which macro factors (social and economic trends) and micro factors (demographic characteristics) interact (Wang 2006).

2.3 Migration factors of Chinese immigrants

Migration behavior is a complex process. There is no agreement on the selection of possible factors. Based on previous research, migration factors generally fall into three categories: human capital, social capital, and location-specific amenities. Human capital refers to an individual's training, skills, and life cycle locations, such as age, education, occupation, and marital status (Kritz et al. 2011). Those personal characteristics intervene with the pull-and-push factors from destinations and origins, influencing the final decision to move or stay. People move when benefits outweigh costs (Lee 1966). Age, education, and labor force classifications are common proxies of human capital. While age decreases migration propensity, education increases migration likelihood. For labor force classifications, white-collar work (managerial, professional, and technical) encourages migration

because it opens up opportunities and networks in other places. On the opposite, immigrant populations are less likely to leave areas with high percentages of the labor force involved in manufacturing and services (Bagchi-Sen 2020; Gurak and Kritz 2000). Self-employment has a particular meaning in the immigrant economy. Self-employed immigrants may provide their co-ethnics job opportunities or housing information, which benefits from residing in ethnic concentrations (Kritz and Nogle 1994; Chen 2017; Wang 2010).

Social capital reflects individuals' capability to obtain scarce resources through their social networks (Sassen 1995). Such benefits may include easy access to housing and employment information and goodwill in grocery stores (Lee 2018; Fang and Brown 1999; Gurak and Kritz 2000). For example, Zhou (2010) observed that residing in the Chinatown enclave of New York City provided Chinese people a familiar working environment and a channel of employment and housing information. Two of the most common social capital indicators are population size and years residing in the US. Ethnic concentration is a deterrent to the immigrant population from moving out of their communities (Lobo and Mellander 2020; Pieterse 2003; Zhou and Lin 2005; Hall 2009; Hall 2013; Zhou 1998). With the growth of an ethnic community, churches, clubs, and microenterprises develop. People can meet most of their daily needs without going out of the ethnic community (Portes and Bach 1985). Immigrants who spend more time in the US have developed their social networks at a place. They would not migrate easily to another location, forfeiting what they have accumulated (Gurak and Kritz 2000).

The third set of factors, location-specific amenities, reflects the impacts on migration patterns from spatial variations in housing, employment, and other economic conditions (Gurak and Kritz 2000). Some scholars used amenity characteristics to measure the quality of life among US states. These characteristics were commuting time, crime rates, air quality, student-teacher ratios, and state and local taxes and expenditures (Gabriel et al. 2003). Previous studies connected individuals' migration behavior with economic developments (Kasarda 1988; Sassen 1995; Zelinsky 1971). For example, since the 1960s, the US saw a migration trend from the traditional northeast industrial regions to

places in the South and West, such as Texas and California. The economic change led to a large population loss within New York City during that period (Kasarda 1988; Sassen 1995). Zelinsky (1971) contended that people migrated from rural areas to urban areas for more employment opportunities during industrialization. After the economy matured, the dominant migration became urban-to-urban. However, there have been employment suburbanization and immigrant suburbanization trends during recent decades (Liu and Painter 2012; Mukherjee and Pattnaik 2020). Due to the significant role of economic conditions in migration decisions, economy-related measures are commonly used as a proxy for location-specific amenities, such as employment growth, per capita income, and unemployment (Gurak and Kritz 2000; Kadarik 2019). Both employment growth and per capita income negatively affect out-migration since people are less likely to leave areas where salaries are relatively high. The unemployment rate has a mixed effect on migration (Gurak and Kritz 2000). Kadarik (2019) suggests that income plays a more critical role than ethnicity in ethnic migration studies.

Ethnic enclaves keep attracting Chinese immigrants from elsewhere in the US. Two classic theories explain migration patterns and possible factors of ethnic populations: spatial assimilation theory and ethnic enclave theory. From the perspective of spatial assimilation theory, though ethnic minorities enter into a receiving country at a disadvantage, with the increase of their socioeconomic status, they eventually move to residential locations that match their social levels (Massey and Mullan 1984; Yu 2018). On top of improved socioeconomic status, residing in the suburbs promotes spatial assimilation (Kadarik 2019; Mukherjee and Pattnaik 2020). White-concentrated neighborhoods are closer to job growth (Liu and Painter 2012). The employment choices could be a factor for drawing immigrants into non-ethnic areas.

Ethnic enclave theory emphasizes the positive roles of ethnic resources in immigrant relocation, such as a familiar working environment and a channel for employment and housing information (Forbes 1984; Fang and Brown 1999; Lobo and Mellander 2020; Zhou 2010). Immigrants tend to live in their

ethnically distinct community in the host country, forming social and spatial segmentations (Portes and Bach 1985; Stillwell and Duke-Williams 2005; Wang 2007). Moreover, ethnic space is the place to fulfill and reproduce identity in various forms, such as languages, storefront signs, and arts (Bodenner 2014). However, with all the benefits from abundant ethnic resources and employment opportunities, there are different voices. Yu (2018) stated that ethnic resources do not benefit people in the long run if people rely heavily on local ethnic resources. Local ethnic resources could hinder people from broader physical, social and cultural mobility. Moreover, ethnic enclaves do not help Chinese immigrants, especially women, improve their status in the labor market (Wang 2010).

Besides ethnic enclave theory and spatial assimilation theory, some scholars use resurgent ethnicity to explain the new trends in ethnic communities. Resurgent ethnicity theory re-evaluates the role of ethnic enclaves. Ethnic concentrations are not necessarily related to limited opportunities or low socioeconomic status (Allen and Turner 2009; Lee 2018). People chose to stay in the ethnic concentrations even when they can move to areas matching their improved socioeconomic status because they may lose more than gain if moved (Walton 2012; Wen et al. 2009). These high-status ethnic areas provide an alternative to assimilation and mobility for Chinese immigrants (Lee 2018; Logan et al. 2002). Sometimes, spatial assimilation, ethnic stratification, and resurgent ethnicity all occur in the incorporation process of Chinese immigrants (Walton 2017). While there is an increasing Chinese population moving into non-ethnic neighborhoods, the Chinese concentrations have undergone growth simultaneously. Among these Chinese neighborhoods, stratification exists in social, economic, and cultural resources. Some are rich, and some are not (Walton 2017).

Besides the previously mentioned theories, some scholars proposed the concept of heterolocalism. Heterolocalism suggests that ethnic populations enter an area and chose a dispersed residential pattern with their co-ethnics. However, they stay in touch in various ways, such as through ethnic associations. In this way, they can keep their identity in a foreign country though not cluster in ethnic neighborhoods (Mukherjee and Pattnaik 2020). Though the heterolocalism phenomenon appears more

often in privileged neighborhoods, it also happens in low socioeconomic areas (Zelinsky and Lee 1998). It is hard to separate the influences of the above theories from each other in a migration study. Migration factors emphasized in them are intertwined (Fang and Brown 1999; Wang 2007).

2.4 Add More Geography into Migration Studies

The geographic dimension in migration studies can be seen in locality and scale. Places matter in providing specific local contexts (Schiller and Caglar 2011; Yu 2018). It is necessary to add a geographic perspective in the migration study (Johnston et al. 2009). Demographic elements of the ethnic population, interacting with the political and economic contexts at international, national, and local scales, help determine where the residents choose to live and the spatial patterns of their assimilation process (Zhou 1998). The assimilation process could be different for people with similar demographic profiles if their local contexts are different (Newbold 2010; Newbold and Foulkes 2004). Employment niche is a local phenomenon. Employment niches are uneven across a metro area. Scholars noticed that people are more likely to niche in the industry close to the ethnic concentration (Ellis et al. 2007). Local and infrastructure policies provide another example to use the lens of the locality. Economic diversity, minimum wages, residential mobility all affect the occupational choices of the immigrants in the US (Liu and van Holm 2019).

An effort to include scales into migration research has been seen in the following research. Chacko and Price (2020) stressed the role of scales in understanding the precariousness of temporary immigrants. Macro-level processes include economic policies at the national level, labor market, and cultural resources in the local area. Micro-level factors are demographic characteristics (Chacko and Price 2020). Almost all geographic problems relate to scales (Fotheringham et al. 2017). The sensitivity to processes of different scales helps explore ethnic migration problems. In studying the social integration of immigrants, Kadarik (2019) applies the method of scalable neighborhoods: 500

neighbors for adults, 400 for children (since they have a smaller active domain), and 50 neighbors for ethnic economic capital.

2.5 Conclusions

This literature review has covered the background of Chinese enclaves (their history and developmental trends), the Chinese immigrants' moving behavior, and its related factors. The literature review shows a need to value the role of hierarchies among the Chinese population and their settlements in understanding the dynamics of migration. During the formation of the Chinese enclaves, spatial variations in the Chinese settlements appear. These settlement variations have been slowly strengthened or altered by the continuous flow of Chinese immigrants to the US in the long run. The settlements provide contextual effects on the population. With these factors changing across space, their impacts on the migration behavior of the Chinese population vary from one place to another. Geographic methods which account for spatial variations will reveal patterns and relationships that are undetectable with other tools.

CHAPTER III

DATA AND METHODOLOGY

3.1 Introduction

To examine the spatial variations in the migration behavior of Chinese immigrants, I perform analyses using both microdata and areal data. First, using public use micro sample (PUMS) data, I performed a classification tree analysis at the individual level. The decision tree analysis helps select the most significant predictors for people with various migration statuses. Neighborhoods are the spatial units of contextual factors. Therefore, a clustering tool is used to study the characteristics of neighborhoods. With the chosen predictors, this study then runs regression analysis at the PUMA scale using three different models: ordinary least squares regression (OLS), regression on each neighborhood, geographically weighted regression (GWR), and multi-scale geographically weighted regression (MGWR). R is the programming language and statistical computing environment used in this study. R codes are available upon request. The three research questions that structure this study are restated below.

Question 1. What contextual factors may affect a Chinese immigrant's migration behavior in the New York-Newark-Jersey City MSA?

Question 2. What are the distribution patterns and characteristics of each neighborhood type in the New York-Newark-Jersey City MSA?

Question 3. How do local factors in the New York-Newark-Jersey City MSA impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially?

3.2 Evaluating Migration Factors with Microdata

3.2.1 Data: Public Use Micro Sample

Research Question 1 involves a detailed examination of the impacts on migration behavior from various factors using PUMS data. PUMS data contain rich demographic information about Chinese immigrants in the US. In PUMS, each record represents a single person, or a household in the American Community Survey (ACS) questionnaires, specifically from the five-percent PUMS of the 2015 ACS. The data file contains a full range of responses. PUMAs are the most detailed geographically contiguous areas in states attached to an individual in the Public Use Microdata Sample (PUMS) data.

3.2.2 Classification Tree Analysis

The decision tree approach is used to study migration factors. These factors include demographic information, socioeconomic status, and ethnic characteristics specific to the Chinese people based on my readings (Chen 2017; Gurak and Kritz 2000; Kritz et al. 2011; Kritz and Nogle 1994; Lee 2018; Ling 2005; Liu and Painter 2012; Zhou and Logan 1991). The classification analysis is to develop accurate classification rules and, more importantly, to gain understanding and insights into mechanisms that create the predictive structure of data. To derive this structure requires considering the influence on the dependent variable from each independent variable and interactions among them. A linear regression analysis builds a global model which assumes relationships stay the same across the study area. While there are always spatial variations, using a global linear regression model would erase specialties from different groups and is not predictive of any group's behavior.

A classification tree produces local models for each sub-group. The procedure is to split data into child nodes that contain similar values of the target variable. This splitting process iterates so that the nodes are progressively pure. The final product is a "tree" showing the data structure. The algorithm selects the most critical variable for each split, which has the strongest influence on the target variable.

Another reason to use the decision tree approach relates to the data in this study. The PUMS dataset in this study is large and complex. Its complexity is shown in the mixture of data types, high dimensionality, nonstandard data structure, and non-homogeneity. There are 21796 weighted records. Each record contains 30 dimensions ranging from geography to housing situations, demographic information, employment-related statistics, migration status, and ethnic-specific characteristics. These variables are a mixture of data types: numerical, ordinal, and categorical. Decision trees can analyze big complex data.

To address Research Question 1, I focus only on migrated population. The analysis results identify groups of people with different migration behavior and the most significant factors. The algorithm also identifies the main types of moving patterns for people from mainland China. The results reduce the number of useful factors for later regression analyses. Moreover, it lays the foundation for understanding the role of contextual factors in the migration behavior of Chinese immigrants. However, the classification tree analysis does not deal with unbalanced data structure well. Regression analysis in later chapters will provide different angles on the research.

3.3 PUMA Neighborhood Classifications

Previous studies highlight the role of neighborhood contexts in a migration decision (Newbold 2010; Newbold and Foulkes 2004; Schiller and Caglar 2011; Yu 2018; Zhou 1998). A neighborhood's definition significantly impacts a predictive model's performance. If areas with similar traits cluster together, analysis results are representative of the areas. However, if areas in a neighborhood vary

considerably in their characteristics, the analysis results would be an average of the places and could not accurately reflect any area's character. In general, the more "pure" a neighborhood is, the more accurate predictor results are. So it is necessary to classify neighborhoods based on their characteristics.

PUMS data offer rich information about the Chinese population on the individual level, which is necessary for performing regression analysis on their migration behavior. However, PUMA-level aggregated data of areal features are not available. So it is not feasible to directly define neighborhoods on the PUMA scale. The solution in this dissertation is a two-step approach. The first step is to define neighborhoods on the census tract scale. The US Census Bureau website offers areal feature data on various geographic scales, including census tracts. Census tracts are the building blocks of PUMA. So neighborhood classifications on the census tract scale can be transferred into the PUMA scale, which is the second step of the approach. However, there is a pre-process needed: variable reduction.

3.3.1 Too Many Variables?

Neighborhoods provide the spatial unit of migration contexts. I select 35 variables in this section to characterize neighborhoods. It is not always a good thing to have many variables. Variables do not exist independent of people or places. One variable provides one angle to look at people or places. One cluster of variables is a set of variables that are significantly associated. When the number of variables increases, angles do not necessarily increase at the same scale since information caught by different variables may overlap. When variables closely correlate, they outweigh some characteristics in variables (Linoff and Berry 2011). If included in a regression model, highly correlated variables contribute to the model multiple times similar information.

Furthermore, too many variables create a high-dimensional space, potentially leading to sparse data issues (Linoff and Berry 2011). It is difficult to capture data distribution patterns of sparse data. Also,

data overfitting may be a problem from too many variables (Linoff and Berry 2011). There are other problems specific to modeling techniques due to a large number of variables. For example, it is difficult to interpret a splitting rule with too many variables in decision trees (Linoff and Berry 2011, 497). The number of clusters, not the number of variables, determines the total information contained. A set of variable clusters assembles different angles and eventually shows a complete image of places or people.

Therefore, before performing neighborhood clustering analysis, it is necessary to reduce the number of variables. The purpose of variable reduction techniques is to capture as much information as possible with a few most significant independent variables. Some classical variable reduction methodologies include forward selection, backward selection, decision tree approach, and principal component analysis. This study applies a variable clustering approach, a powerful tool to study the structure of variables. Compared to traditional variable selection and reduction techniques, variable clustering goes one step further by revealing the data structure under which variables are assigned into clusters (Linoff and Berry 2011). The principal component or the most critical variable can represent each cluster. Using the variable clustering approach, I can reduce the number of variables, facilitating the neighborhood classification and explanation later.

3.3.2 Neighborhood Clustering Analysis on Census Tracts

To address Research Question 2, I applied a clustering tool to study neighborhoods at the census tract scale in this study. People's socioeconomic factors and behavior patterns constitute the overall context of their neighborhood. On the one hand, accurate capture of these contextual factors provides a precise definition of neighborhood. On the other hand, a practical approach to understanding people's behavior is studying observations in their neighborhood contexts. This chapter is intended to classify neighborhoods based on information about residents and their living environment. Therefore, clustering analysis is crucial in understanding the migration behavior of the Chinese people. The

algorithm assigns "adjacent" objects in the same group, whereas "distant" ones are in different groups. Thus, it paves the road for regression analysis on individual neighborhoods later.

3.3.3 Transferring Neighborhood Classification to PUMA level

After classifying census tracts into different neighborhoods, the next step is to transfer neighborhood classifications to the PUMA level. A PUMA's neighborhood definition relies on each census tract's type and frequency in the PUMA. A PUMA's neighborhood type is the neighborhood type with the highest frequency in its census tracts. It is not easy to evaluate the loss of accuracy in transferring neighborhood types from the census tract level to the PUMA level. However, the regression models built on PUMA neighborhoods are expected to be more predictive than the global model for the whole study area. Therefore, model performance statistics, such as R-squared, could help evaluate the effect of neighborhood definitions. Moreover, in meaningful neighborhood types, diverse migration patterns and their underlying processes would be more clear, which can provide another evaluation for a neighborhood partition.

3.4 Regression

3.4.1 Data Aggregation and OLS Regression

To address Research Question 3, I build predictive models with three types of regression methods: OLS, regression on individual neighborhoods, and GWR and MGWR. OLS regression models are built as the reference point for the other two techniques. PUMA is the spatial unit for all models. Aggregated data about Chinese immigrants are not available. US Census Bureau data tables contain only general statistics, such as the total Chinese immigrant population. Other detailed characteristics on the individual level are ignored and dropped because of the Chinese immigrants' small sample size. Therefore, I aggregate microdata into areal data at the PUMA level using the R statistical programming environment. For each PUMA, weighted means are derived for numerical variables and weighted ratios for categorical dummy variables. One dummy variable corresponds to one value of a

categorical variable. One categorical variable of k values can be transferred into k-1 dummy variables. For example, the marital status variable has six values (married, spouse absent, separated, divorced, widowed, and single). Each value corresponds to one dummy variable. However, any dummy variable can be derived if the rest five values are known. Therefore, only five dummy variables should be used to replace the original categorical variable. This transferring process considerably increases the number of independent variables.

A large number of independent variables adds computational steps and time. Moreover, it complicates the model building and interpretation. Applying a collection of model selection and strategies, I can choose the best model with the most significant variables for each migration behavior. The OLS model selection process decreases the possible choices of independent variables for regressions on individual neighborhoods and MGWR. Model fit measurements in the OLS models, such as adjusted R-squared, Akaike information criterion (AIC), and AICc (corrected AIC for small data sets), allow comparison across different types of regression models.

3.4.2 Regression Analysis on Individual Neighborhood

This section evaluates and explains which neighborhood types have the most considerable measurable contextual effect on migration behavior with previous efforts in classifying neighborhoods. To ease the comparison, I use the same regression modeling method as the OLS model for the whole study area: stepwise regression with cross-validation resampling for each neighborhood type. The performance of predictive models is compared across the models for the entire study area and each neighborhood type. These statistics include R2, residuals, and predictors in each regression model. The derived models will reveal specific migration patterns in each neighborhood type.

3.4.3 GWR and MGWR

Geographically weighted regression (GWR) and multi-scale geographically weighted regression (MGWR) are applied to explore the spatial non-stationarity of migration behaviors and the underlying

processes. In an OLS model, relationships are constant across space. Unlike the OLS model, GWR considers the spatial dependence in data through the weighting vector of the regression point. It is consistent with Tobler's first law of geography. "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970, 236). While GWR offers a measure to explore the spatial scales of relations, it assumes all relationships in a model operate at the same spatial scale, which is a critical limitation. MGWR improves GWR by allowing processes to work at various scales (Fotheringham et al. 2017; Oshan et al. 2019).

GWR and MGWR offer the potential to explore the roles of scale in various geographical processes. GWR produces an optimized bandwidth, whereas MGWR produces an optimized bandwidth for each process. As pointed out in the literature review, the Chinese ethnic community has formed as the product of processes at multi-scales: economic restructuring at the international scale, the changes in US immigration policies at the national scale, and the stratifications of Chinese immigrants at the local scale. GWR and MGWR can help reveal the spatial non-stationarity of relationships in migration behavior, facilitating an understanding of the nature of different processes.

3.5 Conclusions

The methodologies presented in this study apply different weighting schemes. Various weighting schemes indicate different data associations: either data associations in the geographical dimension or associations in non-geographic dimensions. GWR and MGWR are examples of the former type. Near places have stronger associations than farther places. Regressions applied to individual neighborhoods are examples of the latter situation. Non-geographic characteristic correlations between PUMAs are the determining factor of their neighborhood type. By addressing the research questions with different methods, I can see my research problem from different angles and better understand the role of geography in relational studies.

CHAPTER IV

CLASSIFICATION TREE ANALYSIS ON MICRODATA

4.1 Introduction

This chapter addresses Research Question 1: What contextual factors may affect a Chinese immigrant's migration behavior in the New York-Newark-Jersey City MSA? This question provides a foundation for regression analysis in later chapters. Moreover, by explaining this question, I can gain insight into subgroups of Chinese immigrants, which offers ways to explain their diverse migration behavior.

This chapter applies classification tree analysis to examine micro-data level factors. A classification tree is a decision tree used for data with categorical variables. In the methodology chapter, I gave the reasons why I choose decision trees in this dissertation. Before discussing results, I will review the decision tree method, algorithm, and parameters because applying a decision tree is not straightforward. The application involves choosing the most appropriate parameter for each calculation step based on the study purpose and data attributes. Next is the data cleaning and manipulation process. This process removes errors and improves data quality. In the effort of performing a systematic evaluation of as many as 30 variables, the data set in this

study is large and complex. Therefore, data cleaning and manipulation are crucial to enable reliable analysis results.

4.2 Decision Trees

The Decision tree analysis is to develop accurate classification rules, and more importantly, gain understanding and insights into the mechanism that creates the predictive structure of data. Decision trees divide data into child nodes that contain similar values of the target variable. This splitting process iterates so that the nodes are progressively pure. The final product is a "tree" showing the data structure. Generally speaking, the decision tree algorithm addresses two questions: how to grow a tree and when to stop the tree's growth (Breiman et al. 1984).

The first question relates to the choice of splitting measures. The default splitting parameter in R is a Gini index (Therneau et al. 2019), which is named after the Italian economist and statistician Corrado Gini (Linoff and Berry 2011). The index measures the purity of observations at a node, varying between 0.5 and 1. The greater the value is, the purer a node is. A Gini value of one implies that all observations belong to the same class, whereas a Gini value of 0.5 shows that two classes are equally represented in a binary split (Linoff and Berry 2011).

The R package "rpart" stands for recursive partitioning and regression trees. Functions in the package follow closely to those in the book "Classification and Regression Trees" (Therneau et al. 2019). It has several parameters to control a decision tree's growth, such as the minimum bucket, the minimum split, and the max tree depth. The minimum bucket value refers to a child node's size, whereas the minimum split value is the threshold for a parent node to originate a split. All tree nodes are equal to or greater than the minimum bucket. When one of the two parameters is determined, the other is set automatically by the program using a formula: minimum bucket value multiplied by three equals the minimum split value (Therneau et al. 2019). The max tree depth criterion specifies a tree's maximum

depth, setting the root node as depth zero. When the max depth is greater than 30, the results are no longer reliable for a 32-bit machine (Therneau et al. 2019).

Although these parameters provide stopping criteria of a tree's growth, they are too subjective and do not guide the best subtree selection. Therefore, there was a methodological transition in the algorithm development history (Breiman et al. 1984). The idea of pruning trees gradually replaced the stopping criterion methods. The concept of pruning a tree is a two-step method. First, it allows a tree to grow fully. The second step is to prune the tree in a bottom-up manner, starting from the bottom tree branches to the root node. After getting a fully grown tree (or close to), the program works backward to cut branches that contribute the least to the pure measurement. A deeper tree is not always better than a smaller tree because when the nodes get too small, the tree runs the risk of overfitting training data. Therefore, the adjusted error rate is introduced to punish tree overgrowth. It serves as a barrier. When the contribution to the pure measurement from a split cannot overcome the penalty from the parameter, the branch would be cut in a pruning process. The adjusted error rate increases in a gradual manner to create a set of subtrees. Then the program chooses the best subtree whose misclassification rate is the smallest for validation data. When the R package "rpart" grows a tree, it performs 10-fold cross-validation on the data.

4.3 Data Manipulation

To address Research Question 1, I applied decision trees on individual-level Chinese population data in the New York-Newark-Jersey City metropolitan statistical area (MSA 35620). Public use microdata sample (PUMS) data offer detailed information about Chinese immigrants. One record represents one person or one group of persons who share the same characteristics, depending on the record weight. A record weight indicates the number of persons represented by it (Ruggles et al. 2021).

There is an inexact correspondence between the MSA and public use microdata area (PUMA) geographical scales. The reason is that for PUMS data, PUMA is the only sub-state-level geographical scale. PUMAs are primarily aggregations of census tracts, cities, and counties. For PUMAs wholly nested in an MSA, there is little or no error. For PUMAs crossing the boundaries of MSAs, errors arise. There is one PUMA area (PUMA 4200500) crossing the MSA 35620 boundary. This area consists of Pike, Wayne, and Susquehanna. About one-third of the land lies inside the study MSA area, and the rest is outside. None of the three counties in Pennsylvania have a large Asian population (IndexMundi 2021). In PUMS data, there are seven records in the PUMA 4200500 (whose boundary is shaded blue in the map below). To exclude these counties would not cause a substantial loss of observations in the data. Therefore, the following analysis will remove data of the PUMA area 4200500. The final data contains 21796 pieces of records from 150 PUMAs in the New York-Newark-Jersey City MSA.

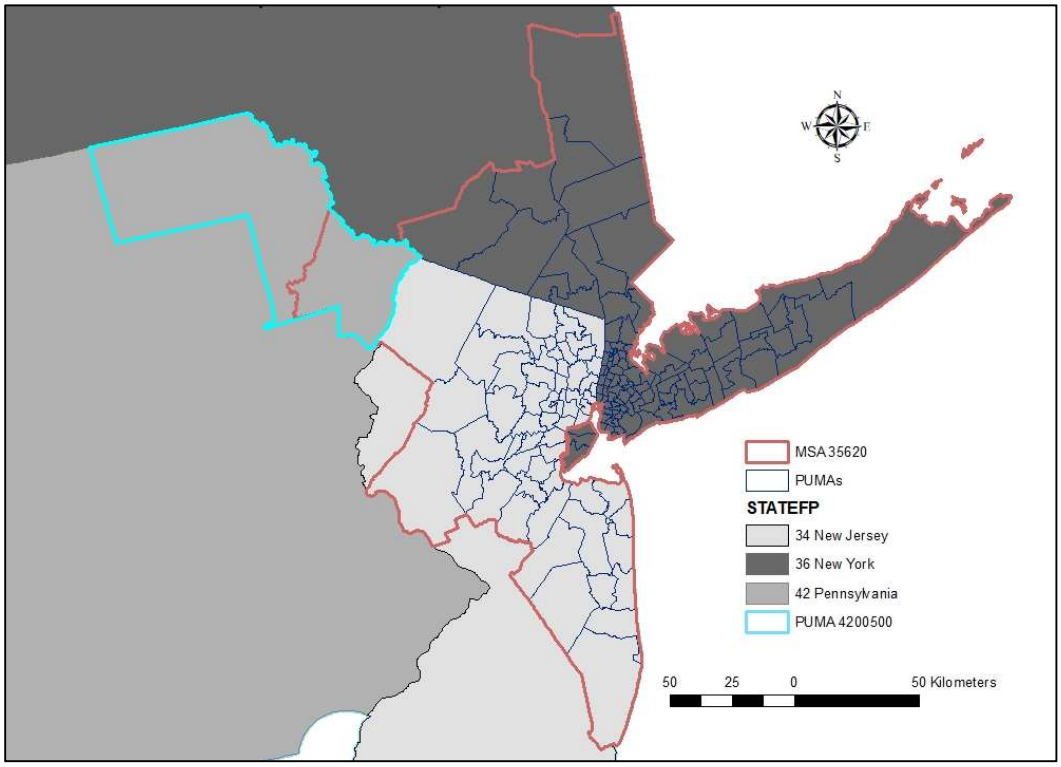


Figure IV-1 Correspondence between PUMAs and NY-N-JC MSA

In this study, the individual-level data are from IPUMS USA (Ruggles et al. 2021). The microdata consists of 30 variables in geography, housing characteristics, demographic information, employment-related statistics, migration status, and ethnic-specific variables. Variables describing housing information include group quarter status, ownership, house acres, values of house units, numbers of bedrooms, and rent. The variable rent is the contract rent plus expenses on utilities and fuels. It is more comparable to contract rent. Other variables describing facilities in a housing unit, such as sinks, stoves, plumbing, bathtubs, or showers, do not vary much in a PUMA area and are therefore omitted.

Demographic variables include sex, age, marital status, educational attainment, income, and poverty status. Based on previous literature, salary from work is one main reason to move to a new place. Therefore, income from other possible resources, such as business and farm income, welfare income, or retirement income, are not included in this dataset. Employment-related statistics are employment status, worker class, and occupational income score. Two more variables are also in this group: transportation means and travel time to work. Occupational income score is a constructed variable to assess occupations based on their financial reward to people. Its values are the median total income of workers in that particular occupation. Based on previous research, jobs are one primary reason to move. The occupational income score was included to represent the drawing force from a better-paid job. The variables of occupation and industry may be helpful but are not included in the study. They are categorical variables whose levels surpass the decision tree algorithm limits of 32. Migration status reports if people moved in the past year. For movers, the variable indicates their residing places one year ago in the same state, in a different state, or abroad.

The study group is the Chinese population in the New York-Newark-Jersey City metropolitan area. The US Census Bureau considers race a sociopolitical construct, not an anthropological or scientific concept (Ruggles et al. 2021). The idea of race has been continuously changing in ACS survey history. Since the 2000 survey, people are allowed to choose more than one race to define themselves.

Using birthplace together with race helps pinpoint the specific Chinese immigrant population. The primary purpose is to exclude the Chinese population born out of China. Chinese children born in the US are one example. They may define themselves as Chinese. However, under the influence of the US culture, they show distinct characteristics and social behaviors compared to their peers who have grown up in China. Therefore, it is best not to include those population groups in the study. Other race-related variables include citizenship status, numbers of years living in this country, the number of years became naturalized, and languages spoken at home.

Recoding variables

Besides regular data cleaning procedures, it is necessary to recode variables for this high-dimension data set. Variables with too many levels could lead to the issue of sparse data, "the curse of dimensionality" (Breiman et al. 1984). During decision tree analysis, high dimensionality increases processing time and causes difficulties in interpreting classification rules. On the contrary, recoding variables could let significant levels stand out. Three variables are recoded: education levels, languages spoken at home, and poverty status. Based on previous research, having or not having a college degree is most influential among all education levels. The original 43 education levels are recoded into two: Bachelor's degree (or higher) and less than a Bachelor's degree. Since the study's primary concern is whether people speak Chinese at home, the variable of spoken languages at home is recoded into three levels: Chinese, English, and others. Poverty was a numerical variable as the percentage of the poverty threshold. Since the study's interest is to determine whether a poverty status of yes would make a difference on travel behavior, the original poverty variable is recoded into two levels: at or under the poverty threshold or above the poverty threshold. In the survey, the poverty values are percentages of a family’s income to the poverty threshold. For example, a value of 1 for the original poverty variable means one percent of the poverty threshold. Therefore, the original poverty variable ranging between zero and one corresponds to poverty status. A poverty variable of greater than one in the survey means above the poverty threshold.

4.4 Migration indicators of the Chinese immigrants

4.4.1 Migrated Population

My first attempt was to conceal the differences between movers and non-movers using a decision tree algorithm. The algorithm fails to classify the data. Because 87.4% of the original data fall into the category of staying at the same house, it is hard to improve its purity with such a dominant class. Therefore, the following analysis focuses only on migrated population. After removing non-movers, there are 2601 records left.

Five splitting variables appear in the pruned classification tree in Figure IV-2. They are employment status, citizenship status, class of workers, number of bedrooms, and rent, among which the indicator rent appears twice. The first three variables are characteristics of people, whereas the last two are housing characteristics. In previous research, wages and rent are the two main moving-related factors of white people; workers' class was a critical moving indicator of Chinese immigrants.

The program R also produces surrogate variables at each split. Surrogate variables are essentially a replacement of the primary splitting variable in case the latter is missing. Each surrogate variable has an agreement ratio: the percentage of data assigned in the right direction using the surrogate variable. It indicates its influence on the primary splitting variable (Therneau et al. 2019). Surrogate variables may offer extra information at a split.

Node 1. The splitting variable at node 1 is employment status. Among 2601 records, the employment rate is 51% (as shown at node 2), and the unemployment rate is 49% (as shown at node 3). Within-state migration is the primary type for employees. The within-state migration ratio increases from 0.62 at node 1 to 0.74 at node 2 after the split. The first surrogate variable is transportation means, which sends 98% of observations in the correct direction. Though the agreement is high, the surrogate splitting variable did not provide extra information. Un-employed people have very different

transportation patterns, including transportation means and transportation time. Therefore, the surrogate splitting variable indicates the same impacts of employment status on migration behavior.

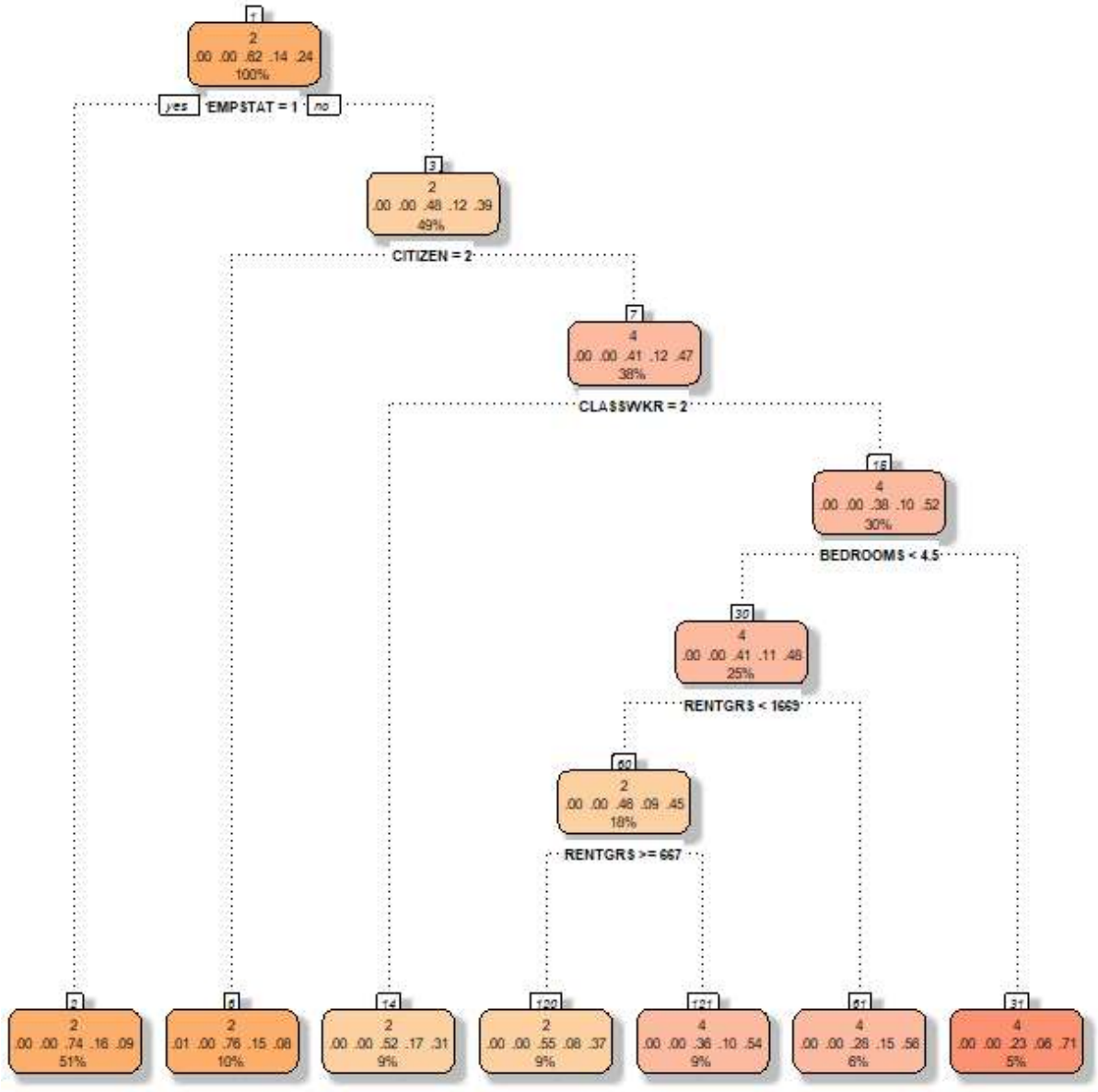


Figure IV-2 Migration indicators of Chinese immigrants

Node 3. For unemployed people, the first indicator of migration behavior is their citizenship status. If they are naturalized citizens, they are more likely to do a within-state migration. The likelihood of within-state migration increased from 0.48 for the Chinese immigrants (shown in node 3) to 0.76 for naturalized citizens (shown in node 6). Age is the first surrogate variable of citizenship status. It is not

hard to understand. It takes a long time to go through the naturalization process. About 10% of the surveyed population are naturalized citizens, and the remainder is noncitizens. A high proportion of noncitizens had moved into the nation from abroad in the survey year (shown in node 7).

Node 7. Among the noncitizen Chinese population, the first moving indicator is the worker's class (whether self-employed). It seems contradictory to classify an unemployed person as self-employed. However, employment status is a constructed variable based on what respondents did in the previous week. In contrast, the class of workers is a long-term self-identification variable. A worker class coded 1 indicates self-employed people, while 2 indicates working for wages. Compared to self-employed people, people working for wages are more likely to move within the state or move between states (as shown in node 14).

Node 16. Self-employed people constitute 30% of the whole survey population. About 53% of the self-employed population have moved from abroad in the survey year. Among the Chinese immigrants who identified themselves as self-employed, there is a general migration trend. They gradually move to smaller housing units with a lower rent as the years passed. This trend can be seen from node 16 to node 30, then to node 60, compared to node 31. Also, the population identified as unemployed is more diverse than its counterpart. So they are classified into sub-groups in the decision tree. A sub-group of movers from abroad differentiate them by large houses and high rent. This sub-group constitutes about 11% of the total surveyed population (nodes 31 and 61). Another sub-group of movers from abroad is identified by low rent and small houses (node 121).

4.4.2 Immigrants from mainland China

When the focus is on people from mainland China, decision tree analysis generates different results from the overall moved population (Figure IV-3). There are 2547 records for people who came from mainland China and have moved in the survey year. The most significant splitting variables in the tree are employment status, citizenship status, college degrees, and class of workers. The variable college

degree did not appear in the classification tree for the overall migrated population, but the other three variables did.

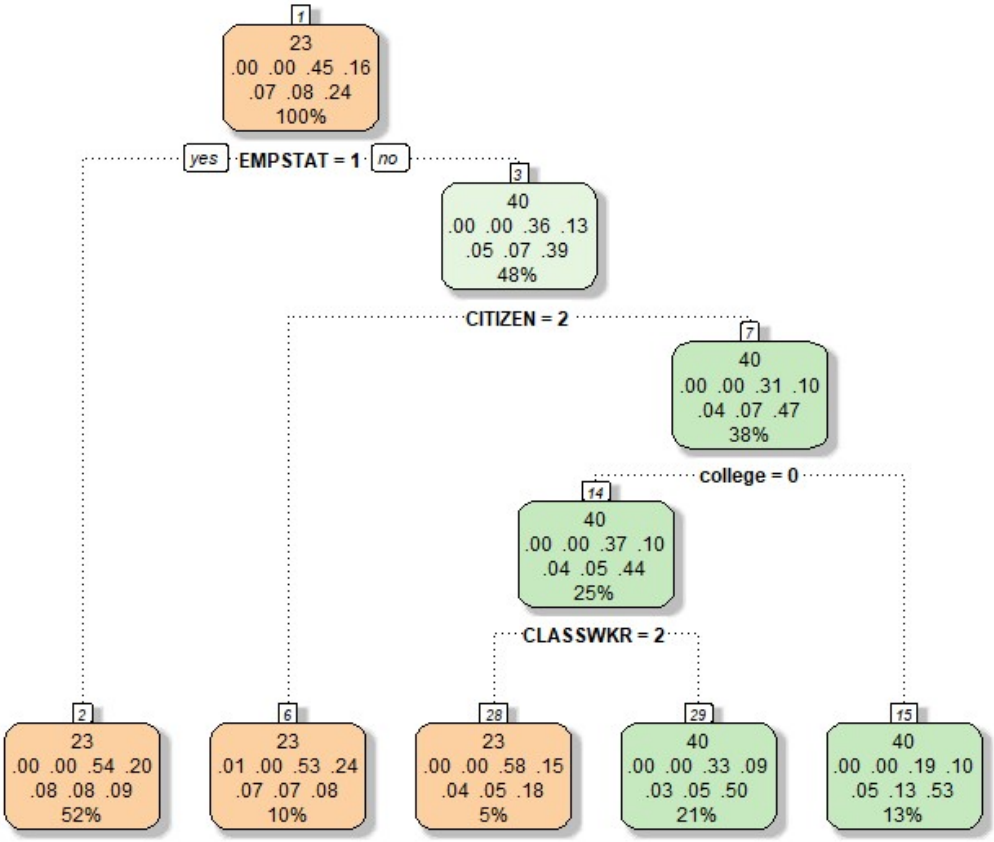


Figure IV-3 Migration indicators of immigrants from mainland China

Compared to unemployed people, employed people are more likely to make a move (node 1). Within-state migration is still the primary moving type for employed people (node 2). Citizenship is the second important splitting variable (node 3). People who have naturalized citizenship are more likely to make within-state migration than people who do not (node 6).

For people who do not have citizenship, having or not having a college degree is the splitting criterion to separate them into different subgroups. One subgroup has a college degree which constitutes one-third of the population without US citizenship. Many of them are new immigrants within the survey year (node 15). The remaining are people without a college degree (node 14), which has two subgroups itself. They differentiate each other by class of workers. For people who work for wages,

within-state is the primary moving type (node 28). People in the other subgroup identify themselves as self-employed. Among the self-employed subgroup, while some people have moved from other PUMAs of the same state, more are new immigrants who moved into the US in the survey year (node 29).

4.5 Conclusions

There are two general types of moving patterns: for work purposes or housing characteristics. The first moving behavior is strongly related to employment status, whether or not self-employed, and income. On average, the population whose migration behavior falls into the first category has a high citizenship ratio and is a long-time resident in the US. For work-purpose migration, the primary type is moving within the state.

On the contrary, housing characteristics (size and rent) are their main concerns for the second type of population. Most of them moved into the nation from abroad within a year. Among them, a sub-group of movers from abroad differentiate them by big houses and high rent. The other sub-group is identified by low rent and small homes. However, the splitting variables concerning housing characteristics are at lower ranks on the classification tree. They are not as significant as employment status, citizenship, and the class of workers in splitting the data. It suggests that housing conditions are not the first concern to move.

When only focusing on the Chinese immigrants from the mainland, one sub-group appears: migrating for educational purposes. This result agrees with the background information about the Chinese immigrants. Many people, especially those from mainland China, are willing to put effort and energy into education. The results also show that the education-purpose movers generally have a college degree. It indicates that people applying to study in a graduate college constitute the primary portion of the movers for educational purposes, compared to other categories (such as to enroll in an undergraduate program or a high school).

I performed decision tree analysis on other sub-dataset as well, but the algorithm failed to run. The first attempt was trying to tell the differences between movers and non-movers. As discussed earlier, the un-balanced data structure could be the reason. Similarly, decision trees did not work to differentiate within-state movers from between-state movers. The within-state migration has a much higher ratio than the between-state migration. Due to the data size, the overfitting issue happened when the focus was on people from Taiwan. Therefore, the decision tree analysis on microdata did not capture all significant indicators in the Chinese people's migration behavior. Regression analysis in later chapters will provide different angles on the research.

In summary, while the decision tree algorithm identified several significant factors related to migration behavior, it missed the necessary spatial component to deny or justify a migration theory. Employment and educational attainment have positive influences on social mobility. Ethnic-specific factors, naturalized citizenship and self-employment, improve social mobility as well. However, within-state migration concealed other spatial patterns in migration since it is the dominant type. The subsequent chapters will offer different spatial perspectives to the research.

CHAPTER V

NEIGHBORHOOD CLASSIFICATION

5.1 Introduction

The goal of this chapter is to reveal the spatial variations across PUMAs. In the process of reaching this goal, I address Research Question 2: What are the distribution patterns and characteristics of each neighborhood type in the New York-Newark-Jersey City MSA?

The answer to this question can situate migration behavior in the right environment, enabling more accurate results in modeling the relationship between migration behavior and its contextual factors. This chapter includes three sections, which build upon each other, leading to the final neighborhood classification results. The first section is to study the variable structure and then reduce the number of variables. This variable manipulation process is crucial for gaining insight into variable mechanics, which is the basis for assigning PUMAs into different neighborhood types. The following two sections relate to the actual classification steps.

It is not straightforward to identify neighborhood types of PUMAs since PUMA-level aggregated data of areal features are not available. To achieve the neighborhood classification goal, I apply a

two-step approach corresponding to the two sections in this chapter. The first step is to define neighborhoods on the census tract scale. Census tracts are the building blocks of PUMAs. Therefore, neighborhood classifications on the census tract scale can be transferred into the PUMA scale, which is the second step of the approach. Before performing any neighborhood classification, I will first examine the variable structure of areal features on the census tract level.

5.2 Variable Reduction

I include in this study measurements for residents and their living environment on the census tract level to characterize neighborhoods. The measures range from socioeconomic indicators to people's behavior patterns and place amenities (such as unemployment rate and percentages of immigrant populations). One measurement provides an angle to look at people or places. The complete information we can gain from a data set is not equal to the number of variables. Because there are almost always correlations between variables, the information contained in variables overlaps. Therefore, it is necessary to reduce the number of variables before performing any data analysis.

The first subsection evaluates some variable reduction methodologies and suggests a procedure to analyze areal features on the census tract level. Next, I will summarize the data and data preparation process. The last two sections are variable clustering tendency analysis and partitioning results.

5.2.1 Variable Reduction Techniques

Variable independence is the fundamental idea of variable reduction (Linoff and Berry 2011). In reality, one-hundred-percent independence of variables seldom exists, although variable independence is an assumption in many modeling techniques such as linear regression and logistic regression. A weakened assumption is that there are no highly correlated variables. The purpose of variable reduction techniques is to capture as much information as possible with a few independent variables. One way to achieve this goal is to include only relatively independent variables, such as forward selection. Forward selection can be used with various modeling techniques, though the most

common one is linear regression. In linear regression, R^2 is the traditional measurement to calculate the total variation captured by independent variables. The statistic of adjusted R^2 improves R^2 by considering the influence of the number of variables in a model. In a forward selection process, the variable with the most significant R^2 is the first independent variable. Then it repeats the process among the rest variables. Variables strongly correlated with the chosen variables are not likely to be included in the model since their information has been vastly captured by the chosen ones (Linoff and Berry 2011).

Backward selection starts with dropping variables that are highly dependent on others. These variables contribute little to the overall information contained in the set of independent variables. Initially, the model includes all variables. Then variables with a small contribution to the model are removed, one at a time. Backward selection works better than forward selection when two or more variables combined are very predictive, but single individuals are not. However, the backward selection technique is inefficient when there are many independent variables due to the amount of work needed in the modeling process (Linoff and Berry 2011).

Regression-based variable reduction techniques find significant variables in a global model (Linoff and Berry 2011). When the study focus is locally optimal properties, other approaches are needed. At times, some variables are only meaningful in certain regions, not across the whole study area. These variables are good at capturing data variation representing a sub-region, while in other areas, a different set of variables may contribute more to the model. The decision tree technique is common for locally optimal analysis (Linoff and Berry 2011). A powerful alternative practice of decision trees is variable selection. Its first step is similar to the forward selection, in the sense of selecting a variable that shows the most data variation within the entire data set based on specific statistics (purity measurements in this case). After splitting data into subsets, the process repeats on each child tree, respectively. Unlike the forward selection technique, variables selected in those child trees may be highly correlated (Linoff and Berry 2011).

In all the above variable reduction methods, selected variables are a subset of original variables. An alternative variable reduction method, the principal component analysis, generates components representing the input variables. The number of principal components is the same as the input variables, but we only use the first several components based on their eigenvalues in most situations. Matrix algebra first introduced principal components, which are linear combinations of input variables. For example, the first principal component is a projection line, which maximizes linear combinations of inputs (Linoff and Berry 2011). The technique creates a scree plot illustrating the amount of cumulative variance captured by principal components. This chart helps determine the number of crucial components. Since the output variables are linear combinations of the original input variables, it is often hard to interpret these principal components.

To reveal the mystery, we want to explore the system's interior structure. Its structure is the set of elements and relationships between them (Newman et al. 2006). Variable clustering goes one step beyond variable selection and reduction by revealing the structure of variables (Linoff and Berry 2011). In reducing and selecting variables, the number of themes determines the number of selected variables, though there is often more than one correct answer. There are some questions we would like to answer. For example, why could we choose certain variables over others? Shall we replace a selected variable with another one because of data quality? What are the influences on the rest variables after we keep or drop a particular variable? Variable clustering is a powerful tool to study the structure of variables. The variable clustering approach is a dynamic process, displaying the changing relationships of variables during each variable selection and reduction process.

The variable clustering approach differentiates main variables from secondary variables, in which variable clusters start to appear. The main variables form the structure of data. Secondary variables reinforce or weaken the structure formed by main variables. The relationship between main variables and secondary variables is analogous to a tree trunk and its branches. While it may lose some information, picking out secondary variables removes noises in the data set and enables a more

explicit variable structure. With principal components, the variable clustering approach removes two types of secondary variables: variables heavily cross-loaded to several principal components and variables with minor contributions to principal components (Weeks et al. 2010).

R Packages

This study used the R package ClustOfVar for variable clustering analysis. Before the development of ClustOfVar, there were R functions for observation clustering: hclust (R Core Team 2021), agnes (Maechler et al. 2005), diana, and pam (Kaufman and Rousseeuw 2009). To apply these approaches for the variable clustering purpose, practitioners need to calculate dissimilarity matrices first. They cannot cluster variables directly. Besides, the observation-clustering functions do not produce a synthetic variable for each cluster (Chavent et al. 2011). For direct variable clustering, SAS software has a widely-used process called VARCLUS (Dhillon et al. 2003; Vigneau and Qannari 2003). It is not available in R software. One great advantage of ClustOfVar is that it can process quantitative and qualitative variables (Chavent et al. 2011).

Building variable clusters requires knowledge in two aspects. One is a distance measure, and the other is a technique to derive variables from clusters (Linoff and Berry 2011). Among all approaches, the most common distance statistic is Pearson correlation. Variables with strong associations are assigned into one cluster. Each cluster derives one synthetic variable to represent the cluster. The derived variables could be different in various techniques. The two most common ways are to use its first principal component or manually pick the most critical variable of a cluster based on professional knowledge. The former approach captures more cluster variance while the latter approach simplifies the interpretation of clusters' meanings. There are also more complicated ways to combine the use of these two methods.

In the hierarchical clustering algorithm of ClustOfVar, correlation squares are used to measure variable distances, which ignores the signs of correlation coefficients. A cluster contains variables

with relatively small distances. The synthetic variable of each cluster is its first principal component. It is often hard to interpret the first component except for some commonly recognized associations, such as the correlations between education and socioeconomic indicators. Therefore, it is an alternative to pick the input variable with the greatest association with its first component in each cluster.

5.2.2 Data Preparation

As stated earlier, this dissertation's primary data are on the PUMA scale, whose boundaries are built based on census tracts and counties. Neighborhood characteristics in this chapter are on the census tract scale, which can merge into PUMAs for later analysis. Based on previous research on migration behavior, I selected variables of the following categories: demographic information, employment, and economic situations, mobility status, characteristics specific to foreigners, and physical housing characteristics.

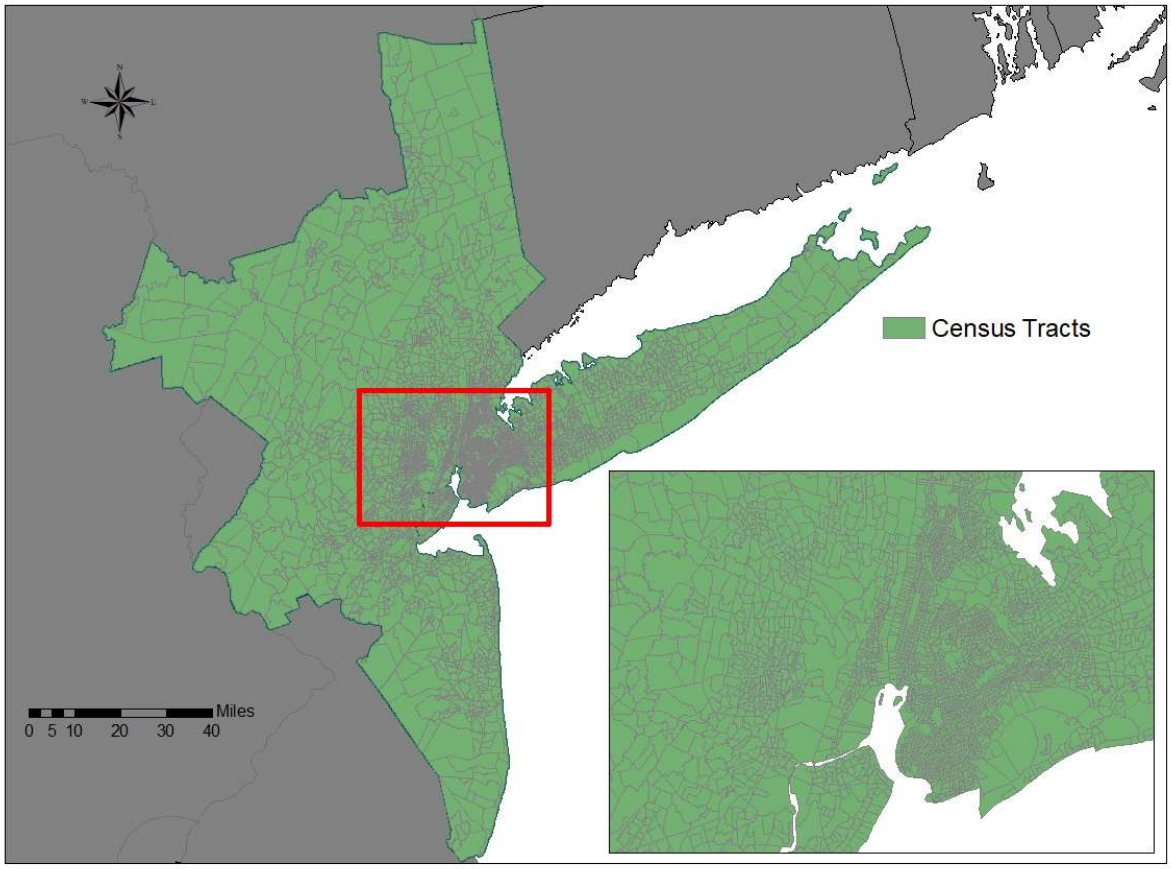


Figure V-1 Census Tracts in the NY-N-JC MSA

Considering the data quality issue, I excluded census tracts with populations of 100 or less. This procedure removed most missing values in the census tract data. Income-related variables, such as house values and median rents, are sensitive and have more missing values than other variables. I replaced them with their means. Based on previous research, having or not having a bachelor's degree is the most critical factor among all education elements. Therefore, the percentages of the population with a bachelor's degree or higher are derived by combing related groups, diminishing the number of variables. The data dictionary table below lists elementary data analyses of the variables used for neighborhood clustering. There are 35 variables, including one variable of place names and two geographic code variables.

Census Tract Data Dictionary				
Variables	**Mean**	**St. Deviation**	**Min.**	**Max.**
Demographics				
Total population	4316	1843	120	28926
Percent male	48.37	4.79	11.20	100.00
Median age	38.68	7.18	11.30	82.00
Percent White non-Hispanic population	47.12	32.87	0.00	100.00
Percent married	45.49	14.10	0.00	83.90
Percent Bachelor (equivalent and above)	35.94	18.89	0.29	100.00
Employment				
Median earnings	56214	19949	5000	221125
Unemployment rate	8.89	5.44	0.00	66.70
Percent poverty	14.07	12.11	0.00	75.71
Percent renter	43.71	25.76	0.00	100.00
Percent self-employed in own incorporated business workers	4.24	3.53	0.00	36.61
Percent self-employed in own not incorporated business workers	4.21	3.20	0.00	45.16
Percent occupations in MBSA	38.90	15.26	0.00	93.33
Percent service occupations	20.40	9.29	0.00	59.84
Percent sales and office occupations	24.17	5.63	0.00	80.00
Percent occupations in NRCM	7.08	4.56	0.00	43.18
Mobility				
Percent moved within same county	5.48	4.98	0.00	49.40
Percent moved from a different county, same state	2.02	3.43	0.00	54.80
Percent moved from a different state	1.36	2.38	0.00	47.30
Percent abroad one year earlier	0.84	1.20	0.00	18.80
Chinese-Specific Variables				
Percent foreign borns	29.33	17.04	0.00	81.55
Percent naturalized	58.12	21.47	0.00	100.00
Percent Asian	10.44	10.77	0.00	89.80
Percent Chinese population	3.77	6.18	0.00	83.10
Chinese population	159.50	293	0	7473
Housing Characteristics				
Total housing	1698	817	0	12840
Percent vacant housing units	8.71	10.06	0.00	97.00
Median number of rooms per housing unit	5.13	1.18	1.30	8.50
Median house value	461705	225236	24500	1979200
Median gross rent	1402	449	235	3471
Percent housing units built since 2000	7.96	9.77	0.00	96.30
Percent housing units built between 1980 and 1999	12.78	14.17	0.00	100.00

Table V-1 Census Tract data dictionary

5.2.3 Variable Clustering Dendrogram

Variable selection and reduction are dynamic processes involving background knowledge, correlation analysis, and principal component analysis before, in the middle, and after applying variable clustering techniques. As pre-processes, they clean redundant variables and variables of poor quality. For example, one survey question generates four occupation variables of MBSA (management, business, science, and arts occupations), service, sales, and NRCM (natural resources, construction, and maintenance). We can derive any occupation percentage by subtracting the remaining three categories from value one, leading to collinearity. Based on previous research, the percentage of NRCM has the lowest correlation with Chinese immigrants and was therefore eliminated. Some base variables were used to derive other variables. It is necessary to drop the total population, the total number of housing units, and the Chinese population. The percentages of Asian and foreign-born people significantly correlate with Chinese percentages, with a Pearson value of 0.76 and -0.69, respectively. Since the focus is on Chinese immigrants, it strengthens the Chinese population's impact on the neighborhood study by removing the influence of other Asian populations.

Standardizing variables

A critical concept in data cleaning is scale. Variable measurements are on different units or over different ranges. Ignoring variable units would lead to several variables of large numbers dominating the set of variables. For example, income is generally measured by thousands of dollars, whereas ages are between 1 and 100. When we put them together in a mathematical model, the variable income will diminish ages' contribution to the model. Scaling provides a way to standardize all variables with the same mean value, in most situations, 0 and over the same ranges. This way, each variable contributes its fair share to the overall data.

Clustering Dendrogram

R generates a cluster dendrogram showing the result of variable clusters with variables connected in a tree (Figure V-2). Near variables combine into one node, forming a new synthetic variable. This

process repeats to develop a higher-level tree node until all variables connect. A synthetic variable of each cluster could be derived from principal component analysis. On the chart, the height index is the loss of homogeneity when two clusters merge. A cluster homogeneity measures the links between cluster variables and their synthetic variable, based on squared Pearson correlations for quantitative variables (Chavent et al. 2011).

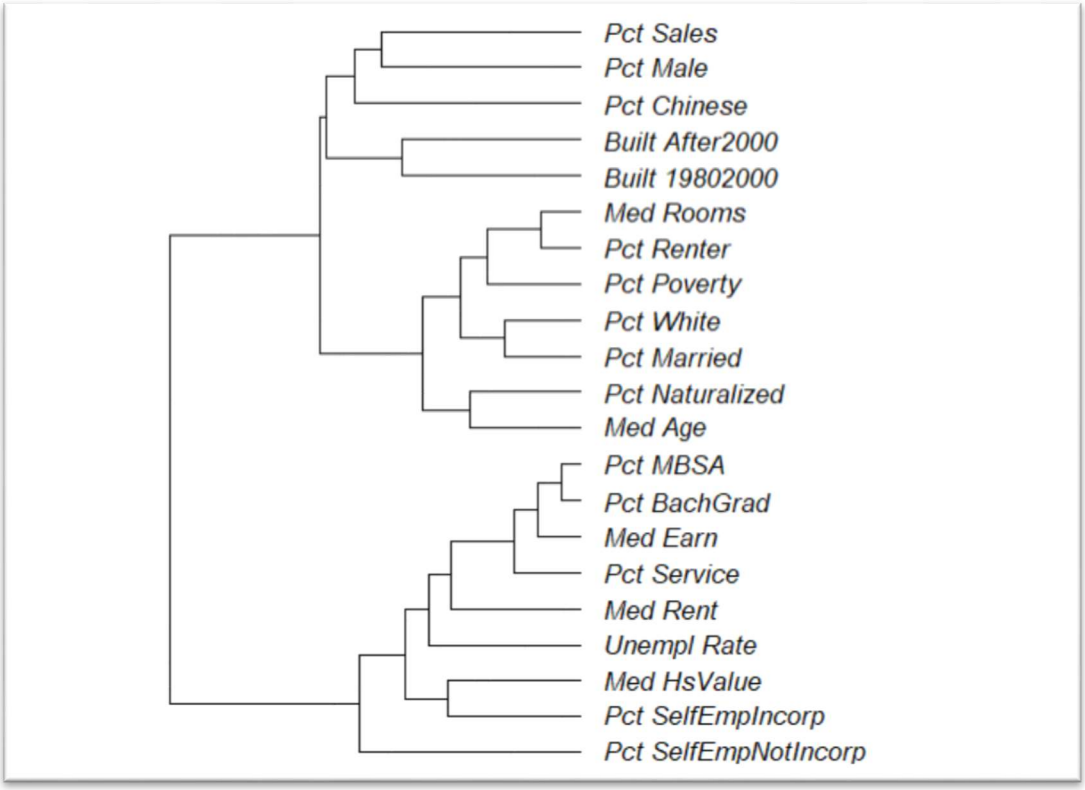


Figure V-2 Clustering dendrogram of 21 variables

Some variables are separate from others: either they are vital or not important to the research problem. Dropping unimportant variables leads to a different clustering dendrogram with a cleaner structure. Linoff and Berry suggest dropping variables with small loadings on main components or variables cross-loaded evenly across several principal components (Linoff and Berry 2011). The male percentage variable has no loadings on the first two principal components, whose variances account for almost half of the total variances. The median age variable is evenly cross-loaded on the first three

principal components. After dropping the two variables of median age and male percentage, the proportion of cumulative variance for the first two principal components increased from 48% to 52%.

Similarly, two variables of houses built from 1980 to 2000 and built after 2000 are reasonably far from all the rest variables. The 1980s was a period when a significant Chinese population came to the US. They provide a perspective of city development in the physical environment. However, after dropping male percentages and median age variables, the percentage of houses built from 1980 to 2000 has the smallest loadings on the first principal component. Furthermore, the variable of houses built after 2000 has the smallest loading on the second principal component (-0.129). Thus, these two house-related variables are not included in the following analysis.

Another isolating cluster relates to self-employment: incorporated self-employment and unincorporated self-employment. These two variables represent distinctive demographics of Chinese immigrants. Based on previous literature, unincorporated self-employed workers are more likely to be in the early stages of their businesses and have un-paid family workers. They represent an early form of businesses run by Chinese immigrants in old Chinatowns. The percentage of unincorporated self-employment has no strong correlation with the rest variables in my data set.

On the contrary, incorporated self-employed workers tend to have a high marriage rate, have received more education, earn much more, and have a higher percentage of citizenship (Hipple and Hammond 2016). The correlation analysis results in this study are consistent with the above statement. These two self-employment variables do not have heavy loadings on the first two principal components. One possibility is due to the small Chinese population compared to white people. However, they closely relate to the Chinese percentage variable and will stay in the analysis.

5.2.4 Variable Partitions

The final variable clustering dendrogram is shown in Figure V-3. It has a high clustering tendency denoted by a Hopkins statistic of 0.87. Hopkins statistic evaluates the null hypothesis that all data

points are from the same distribution. The bigger the statistic value, the more confident we are that meaningful clusters exist. A value of 0.75 or higher is preferred (Lawson and Jurs 1990). The next step is to assign variables into clusters. The ClustOfVar package also provides a bootstrap approach to evaluate the stability of variable partitions applying Rand indices. Rand index measures partition similarities between two data clusterings (Hubert and Arabie 1985; Rand 1971). It ranges between 0 and 1. A stable partition has a high value of mean and a small variation for the Rand index.

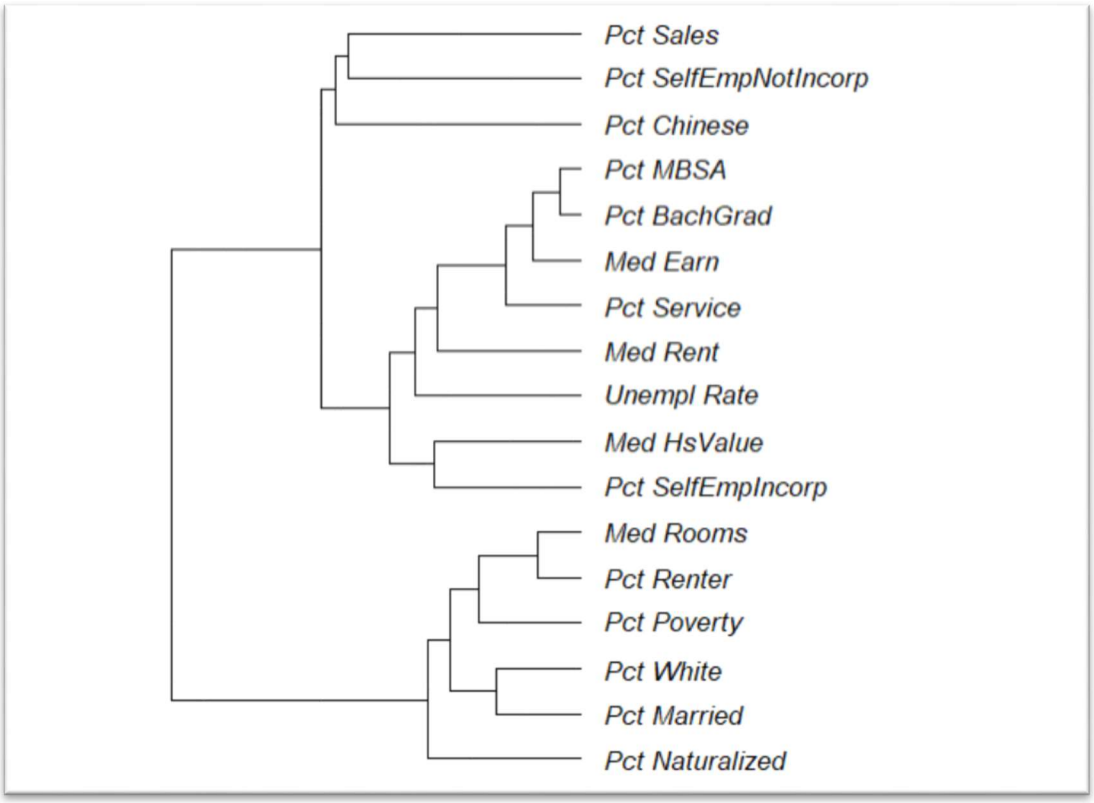


Figure V-3 Clustering dendrogram of 17 variables

The stability plots of partitions suggest three or five variable clusters (Figure V-4). The first graph illustrates changes in the mean of the adjusted Rand index. The greater the mean is, the more stable a partition is. The horizontal axis is the number of clusters. When the number of clusters arrives at three or five, the stability line reaches or closes to a local maximum. Then the mean stays relatively stable and rises to another local maximum at eleven clusters. The dispersion chart of the adjusted Rand index generally agrees with the index's means. Partitions of three, four, or five have short bars

without outliers, indicating small dispersions. The dispersion magnitudes for seven or eight partitions are acceptable as well. However, the increase of variables could quickly increase the difficulty in analyzing and interpreting neighborhood types discussed in the next chapter. Therefore, the goal is to obtain a stable partition with fewer variables.

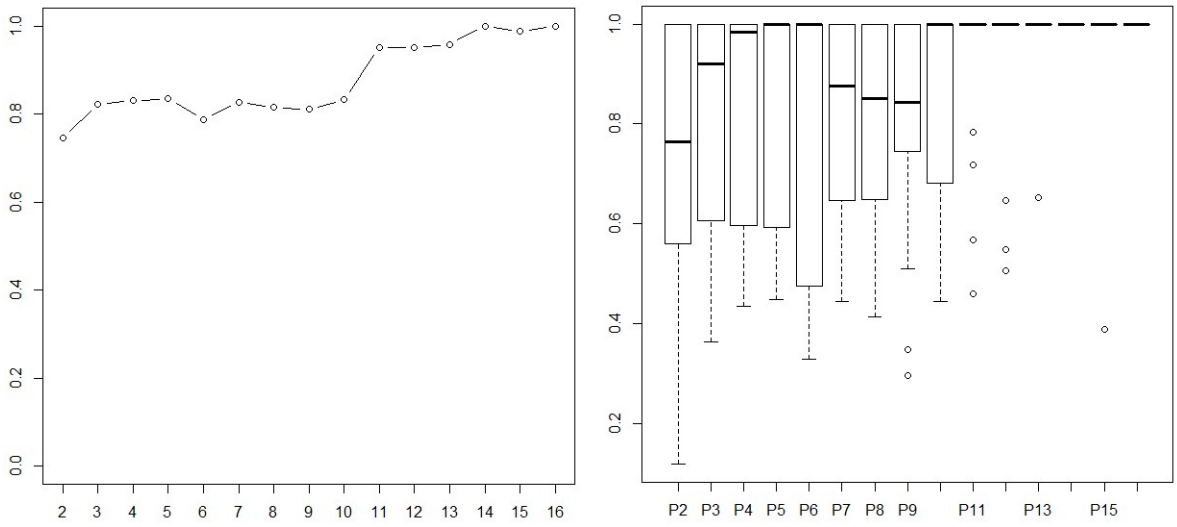


Figure V-4 Mean and dispersion of Rand index

The principal component analysis offers another approach to evaluating partition stability. A homogeneous cluster should only have one dominant component whose eigenvalue is greater than one. I choose three partitions over five after studying the principal component analysis results (see Table V-2). These two partitions are relatively similar except for one issue. Variables of Chinese people, unincorporated self-employment, and sales occupation fall into the same cluster when assigning variables into three partitions. In contrast, they form their own cluster when using five partitions. Grouping the above three variables is more consistent with previous research.

Cluster 1 measures the stability status of the neighborhood. Stable neighborhoods have high percentages in the white population, married population, the naturalized population, and low rates in renters and people under poverty. These areas typically have large houses as represented by the number of bedrooms. Variables in cluster 1 have strong correlations with their first principal

component. A neighborhood with a positive value of the first principal component is an unstable neighborhood: high in renters and poverty percentages and low in white, married, or naturalized percentages. To ease the interpretation of neighborhood features in the following analysis, the negative values of the first principal component will be used to represent cluster 1.

Cluster 2 has three variables: Chinese percentage, unincorporated self-employment percentage, and percentage of sales occupation. Compared to cluster 1, cluster 2 is more dispersed. Its three principal components are close in their eigenvalues. Instead of the first principal component of the cluster, I use Chinese percentage to represent cluster 2. The variable of Chinese percentage is more important for the purpose of this study. Similarity results show that the other two variables have little correlation with the Chinese percentage, though they are assigned in one cluster. One possible reason for the small correlations is that data are at the census tract level, on which ethnic-related influences are canceled out. Therefore, the clustering process represents the entire population's average characteristics, and not just the Chinese population.

Cluster 3 depicts the socioeconomic status (SES) of an area. In cluster 3, the first three variables with the most significant correlation (absolute values above 0.9) with the central synthetic variable are the percentage of people with a bachelor or higher degree, percentage of people working in MBSA occupations, and median earnings. Cluster 3 describes areas with people of high educational levels, high-paying jobs, and incomes. These areas usually have houses of high value or high rental costs and low unemployment rates. Some people own an incorporated company. Variables in cluster 3 are strongly correlated. For the same reason as cluster 1, the negative values of the first principal component will be used to represent cluster 3.

In summary, three statistics are used to characterize census tracts in the study area, replacing the original data set. These three statistics are stability status, Chinese percentages, and the SES index. With the above three statistics, I can continue on the neighborhood clustering at the census tract level.

Summary of Variable Clusters				
Variable	**Squared Loading**	**Correlation**	**Component Eigenvalue**	
Cluster 1				
Percent renter	0.85	0.92	1	3.97
Median number of rooms	0.69	-0.83	2	0.63
Percent married	0.69	-0.83	3	0.54
Percent poverty	0.66	0.81	4	0.45
Percent white	0.62	-0.79	5	0.30
Percent naturalized	0.47	-0.68	6	0.11
Cluster 2				
Percent sales	0.47	-0.69	1	1.10
Percent unincorporated self-employment	0.41	0.64	2	0.98
Percent Chinese	0.22	0.46	3	0.92
Cluster 3				
Percent Bachelor (equivalent or above)	0.86	-0.93	1	4.88
Percent occupations in MBSA	0.85	-0.92	2	0.83
Median earnings	0.81	-0.90	3	0.66
Percent service	0.70	0.84	4	0.62
Median rent	0.49	-0.70	5	0.53
Percent incorporated self-employment	0.43	-0.65	6	0.25
Unemployment rate	0.38	0.61	7	0.17
Median house value	0.36	-0.60	8	0.07

Table V-2 Summary of variable clusters

5.3 Neighborhood Clustering on Census Tracts

Clustering is crucial in understanding people's behavior. Clarita's PRIZM Premier segmentation system has classified every US household into one of 68 segments based on their geodemographic information and consumption behaviors (Claritas 2021). Similarly, neighborhood clustering analysis in this study will help gain insights into neighborhood contexts for the Chinese immigrants and associate diverse migration patterns with contextual factors more precisely.

In the following subsections, I will introduce some key concepts related to distance. Distance measurements are fundamental in assigning areas into groups. Subsequently, it is a brief review of some classical data clustering methods, which allow me to choose the appropriate method for the areal data in my study. Primary analysis steps include: calculating clustering tendency, determining

the number of clusters, and splitting data into clusters. Data clustering is both an objective and subjective process, which requires much deliberation in every step. For example, researchers have developed more than 30 indices to help determine the number of clusters. Though these indices provide some objective standards, the final decision relies heavily on personal opinions and background knowledge. An optimal number of clusters should provide insight into the study problems, even if this number is not the best derived from statistical indices. Moreover, different technique choices may lead to different clustering results, requiring an active engagement from the practitioner in every analysis step.

5.3.1 Distance Measures

Clustering analysis assigns "adjacent" objects in the same group, whereas "distant" ones are in different groups. Various distance definitions may lead to clusters of different shapes and numbers. Euclidean distance is the default setting in many clustering analysis software packages. Based on Euclidean distances, observations with similar values are closer (Greenacre and Primicerio 2014; Shirkhorshidi et al. 2015). Another two classical measurements are Pearson correlation coefficients and Spearman correlation coefficients. Spearman correlation coefficient measures the similarity of two points based on their ranks, which is preferable for categorical data. Pearson correlation coefficient is more common for numerical data. Pearson correlation coefficients measure linear correlations of data, focusing more on overall data distribution patterns (Greenacre and Primicerio 2014). Two strongly related observations are considered fairly close even if their Euclidean distance is big. Here data similarity is based on Euclidean distances since data's magnitudes are essential in studying their relations.

Distance statistics measure similarities between points, and the statistic of linkage is for clusters. Complete linkages, single linkages, average linkages, and Ward's linkages are standard measurements (Shalizi 2009). A complete linkage is the maximum pairwise distance between a point in cluster one and cluster two. On the opposite, a single linkage is the minimum pairwise distance of two clusters.

An average linkage is an average distance between two clusters. The Ward's linkage minimizes within-cluster variance. Ward's linkage tends to produce more compact clusters (Flynt and Dean 2016). I use Ward's linkages to derive small intra-neighborhood variance and big inter-neighborhood variance.

5.3.2 Review of classical clustering algorithms

A partitioning clustering algorithm splits data into k clusters, in which k is a pre-determined number. It requires background knowledge from the practitioner to choose an appropriate value of k. Although there are techniques to help decide the best k value, the final decision, on a large part, is still subjective and cannot rely purely on statistical measures. Among all partitioning clustering algorithms, k-means is the most commonly used one. After indicating the k value, the algorithm randomly chooses k points as the seeds of each cluster. The next step is to assign the remaining points to a cluster based on their distances to the seed points. After completing all data point assignments, the center of each cluster is re-calculated. The data assignments and center calculation processes repeat until there are no more changes in the partitions (Hartigan and Wong 1979; MacQueen 1967).

K-means clustering can deal with large data sets since it only calculates distances between a point and its cluster center. However, it is quite sensitive to outliers. Furthermore, the final cluster assignment is a local best solution, varying each time with different seeding points (Linoff and Berry 2011). Scholars have developed various approaches to improve the k-means clustering technique. For example, instead of choosing k ransom seeding points, a k-medoids clustering algorithm calculates and uses the most centrally located points (medoids) as seeds in the algorithm (Kaufman and Rousseeuw 1990). The CLARA algorithm (clustering large applications) can handle large data applying sampling techniques (Kaufman and Rousseeuw 2009).

Hierarchical clustering is another clustering technique category frequently used in research. There are two general types: agglomerative clustering or agglomerative nesting (AGNES) and divisive

clustering (DIANA). In an agglomerative clustering algorithm, each point at the beginning is a cluster itself. Then adjacent clusters merge into a new cluster. This process iterates until all data points merge into one big cluster (Flynt and Dean 2016; Linoff and Berry 2011).

In contrast to agglomerative clustering, divisive clustering is a top-down algorithm. It starts from a single cluster that contains all data points. The data point with the maximum dissimilarity breaks away, and the original cluster splits into two sub-clusters. This process continues until each cluster contains only one data point (Flynt and Dean 2016). Agglomerative clustering and divisive clustering often produce similar results. However, agglomerative clustering can capture small clusters, whereas divisive clustering is preferable for generating large clusters (Kassambara 2017). In this study, clustering analysis is performed on the area of thousands of census tracts. The ideal number of clusters is no more than ten. Thus divisive clustering is preferred.

5.3.3 Analysis and Results

Clustering Tendency

Clustering techniques produce clusters, whether they are meaningful or not. Therefore, clustering tendency analysis is necessary. It serves as an early-stage validation technique for detecting valid clusters. This chapter applies two R functions to estimate the clustering tendency of the study area. The first method is the Hopkins statistic. The result of a Hopkins statistic of 0.95 suggests that it is safe to reject the null hypothesis and conclude that the data contain meaningful clusters.

The second method is called visual assessment of the cluster tendency approach or VAT. The VAT approach starts by calculating the dissimilarity matrix of a given data set based on Euclidean distances. Elements in the dissimilarity matrix are re-ordered in a way that similar objects are adjacent to each other. Clusters formed by these similar objects are displayed as squares in the ordered dissimilarity image. Therefore, blocks along the diagonal are visual evidence of meaningful clusters (Bezdek and Hathaway 2002; Kassambara 2017).

The clustering tendency result from the VAT method agrees with Hopkins statistic. Color levels visualize dissimilarity values. The image is visibly different from a random image. Pure black shows a low dissimilarity level, whereas pure white represents a high dissimilarity. It detects several dark squares along the diagonal, indicating the presence of meaningful clusters.

However, the number of clusters in the image is vague. One small square around the center of the image has a very dark color tone, popping up from its surroundings. This small square indicates a meaningful cluster. However, other parts along the diagonal do not have a clear cluster structure. The bottom left corner is a medium-sized square with a darker center and lighter-shaded edges. We can view it as either one loose cluster or two or three small clusters. The cluster structure in the top righter top corner is similar to the bottom left corner yet is even more ambiguous. It needs further investigation to detect the number of clusters.

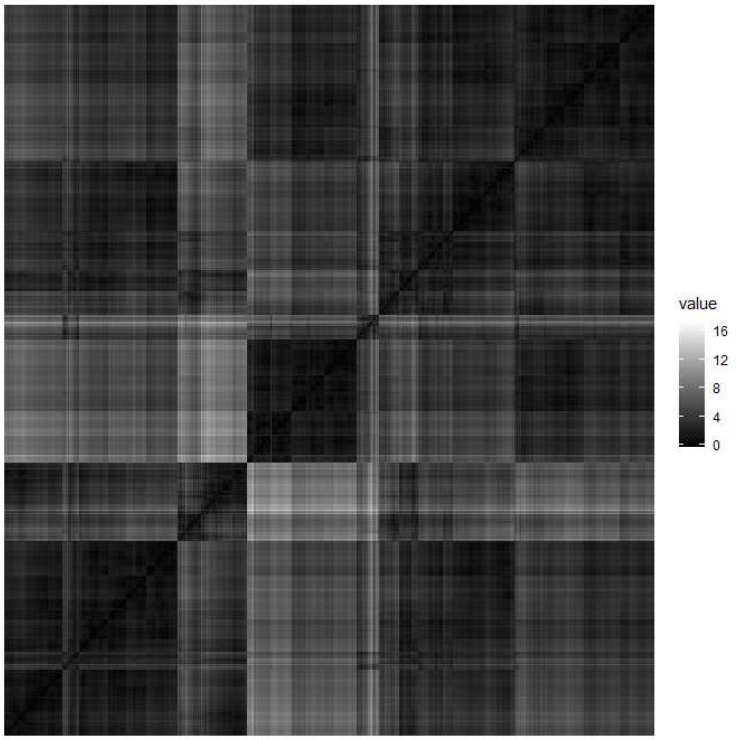


Figure V-5 Ordered dissimilarity image

Determining the optimal number of clusters

I use three methods to detect the optimal number of clusters: elbow, average silhouette, and gap statistic. The elbow method's basic idea is to find a value of k so that the contribution of k to the total variance is no longer significant. For each k, it calculates the total within-cluster variance. Then it plots the total variances against the corresponding number of clusters. With k increasing, the variance decreases. The improvement will not be as substantial at a certain point (elbow point), indicating an optimal number of k (Kassambara 2017).

Similarly, the average silhouette method optimizes a criterion: the average silhouette of all observations. For each data point, the silhouette coefficient measures the quality of a clustering method. A compact cluster indicates points with a small within-cluster average distance and a sizeable between-cluster distance, called a high average silhouette width (Kassambara 2017; Kaufman and Rousseeuw 2009).

The gap statistic method involves null hypothesis testing. It compares the within-cluster variance with its expected value under the null reference distribution for a particular number of clusters k. How to choose an appropriate null distribution is out of this study's discussion scope, but uniform distributions and random distributions are two common types. A gap statistic can be viewed as the deviation of observed within-cluster variance from its expected value. The purpose of the approach is to pick up the k that maximizes the gap statistic. It means that it is far from a random distribution or a uniform distribution (Kassambara 2017; Tibshirani et al. 2001).

The elbow method, silhouette method, and gap statistic offer objective standards in detecting the optimal number of clusters. However, they are not making the task easier cause their suggestions of an optimal number are not entirely the same (Figure V-6). In the images, each horizontal axis indicates numbers of clusters (k), and each vertical axis marks total within-cluster variance, average silhouette width, and gap statistic, respectively. The average silhouette method suggests two clusters

as the dividing result. This result partially agrees with the elbow method. There is no clear turning point or elbow on the image, but the turning at the point of two is more substantial than other points. The gap statistic suggests dividing the data into nine clusters.

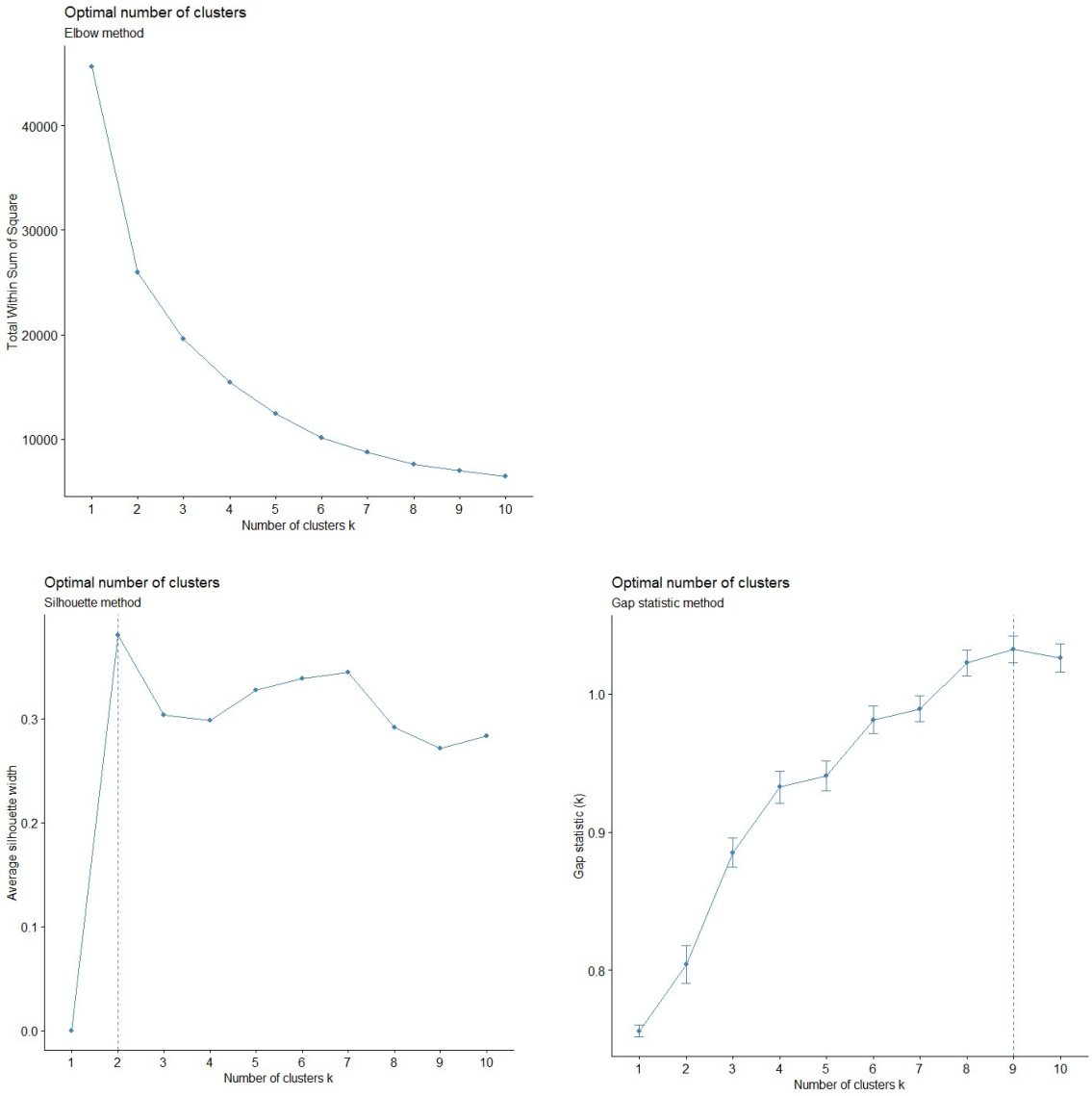


Figure V-6 Determining the optimal number of clusters

However, the differences are not as significant when we look at the vertical axes in the image. Despite the curve's up-and-down pattern on the silhouette method figure, the average silhouette widths on the vertical axis are around 0.3 for the most part. When the optimal number is equal to or greater than six on the gap statistic image, the gap statistic is around 1.0.

Another more intuitive approach is to divide the cluster dendrogram visually and see if it makes sense in terms of the research purpose. A dendrogram is a tree-like structure derived at the early stage of hierarchical clustering analysis. With a small value of the number of clusters, clustering results are more steady. However, it runs the risk of concealing characteristics of particular groups. On the other hand, there are drawbacks to a large number of clusters. It would not be easy to describe every cluster accurately since the dataset in this study is large. After experimenting with several possible values, I split the dendrogram into two and six clusters (Figure V-7).

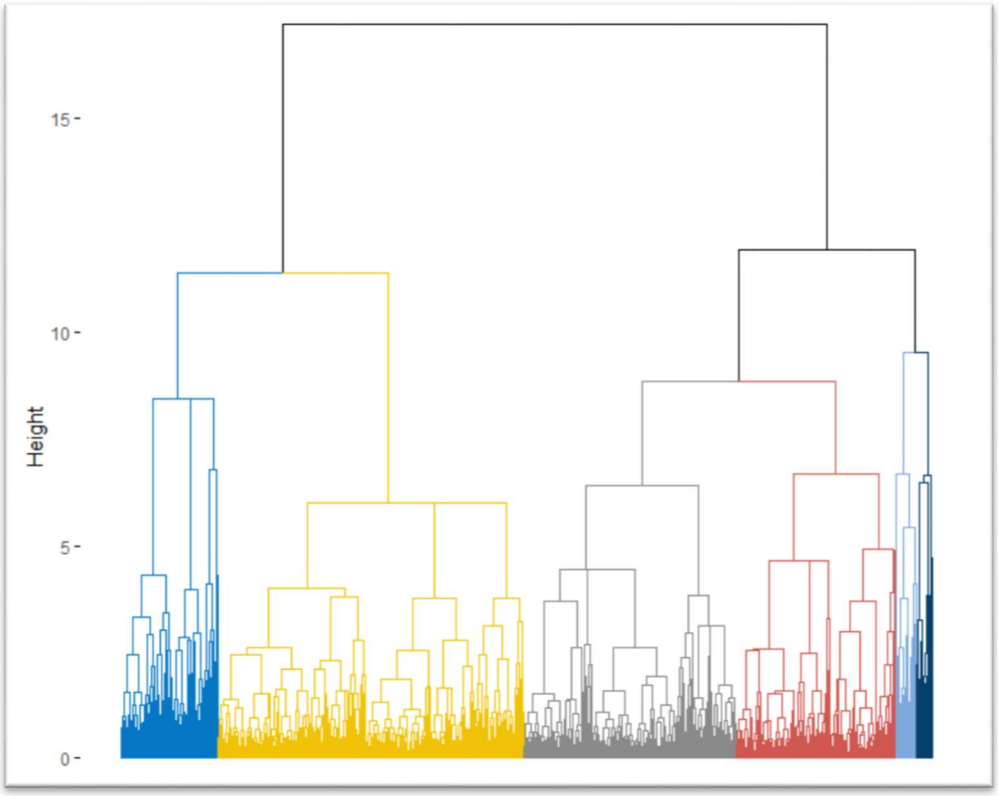


Figure V-7 Cluster dendrogram

Cluster Dendrogram

Table V-3 shows the results of partitioning the study area into two or six neighborhood types. Being stable and holding high SES characterize neighborhood type 1. On the contrary, neighborhood type 2 areas have low Stability and low SES. Neighborhood type 2 has a greater Chinese concentration than

type 1, with the former (0.13) slightly above the study area's mean percentage and the latter (-0.13) marginally lower than the mean.

When the study area further divides into six clusters, neighborhood type 1 is further split into two sub-types (s1 and s2), whereas neighborhood type 2 is further divided into four sub-types (s3, s4, s5, and s6). Generally, a neighborhood with a positive SES has a positive Stability index. Based on the SES and Stability magnitudes, the census tracts fall into neighborhoods of different SES and Stability levels. Neighborhood s1 is marked as the highest SES and a decent Stability index, followed by neighborhood s2, whose SES is 0.79 yet a higher Stability of 1.68. The SES and Stability of neighborhood s5 are barely above the average. The other three neighborhood types (s3, s4, and s6) all have negative SES and Stability values. Neighborhood s4 has the lowest SES and Stability. Neighborhood s6 is the second to the last. The SES and Stability values of neighborhood s3 are just below the average.

Cluster06	Frequency	SES	Stability	PctChinese Std.	PctChinese	Cluster02
s1	550	4.17	0.96	**0.04**	**4.12**	1
s2	1750	0.79	1.68	-0.18	2.23	1
s3	1210	-0.95	-0.70	-0.20	2.06	2
s4	912	-2.65	-2.76	-0.30	1.20	2
s5	119	0.26	0.08	**2.82**	**27.97**	2
s6	92	-1.51	-1.37	**5.21**	**48.48**	2

Table V-3 Summary statistics of 6 neighborhood types

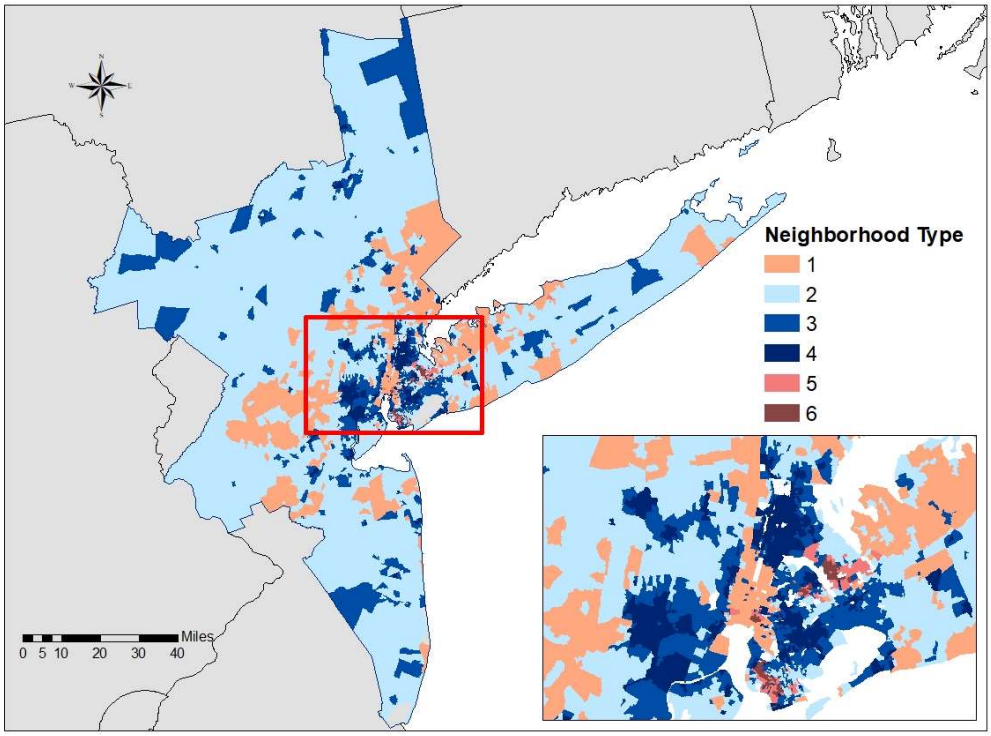
As mentioned earlier, data in the clustering process are standardized, with a mean of zero and a standard deviation of one. To better understand the size of the Chinese population in these neighborhoods, I transform the standard values back to their original percentages (as listed in Table V-3). Two neighborhood types have the highest Chinese percentages (48.48% in s6 and 27.97% in

s5). Neighborhoods s6 and s5 correspond to the two types of Chinese enclaves as described in previous research. The first type (with 92 census tracts) started to form in the early Chinese immigration history, indicated by people with low SES (-1.51) and low Stability (-1.37). Later, with the accumulation and rise of their SES, some people move into suburban areas. At the same time, new immigrants with relatively high SES (0.26) and Stability (0.08) come into these areas. During this process, a new form of Chinese enclaves (with 119 census tracts) appears.

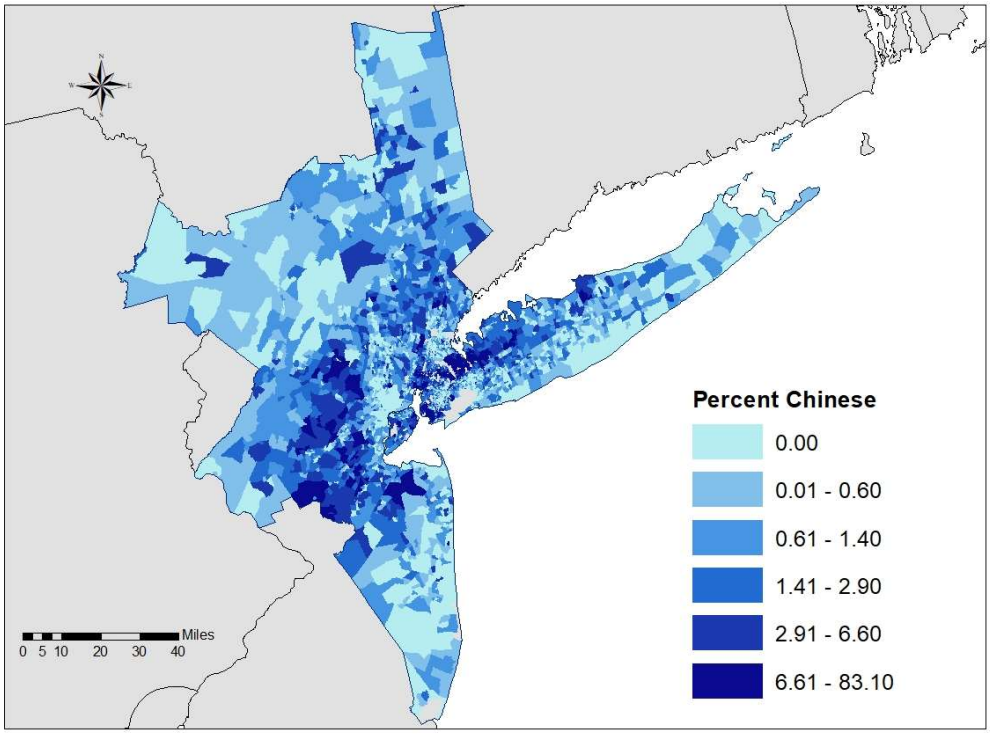
Although low SES and Stability characterize neighborhood types s6 and s5, there is a third neighborhood type whose Chinese percentage is above the average. Neighborhood type s1 has the highest SES and a decent Stability index. This neighborhood type has 550 census tracts. The Chinese percentage (4.12) in this neighborhood is just above the average of 3.77.

Figure V-8a shows the six neighborhood types, and Figure V-8b is the Chinese percentages in each census tract. In neighborhoods s1, s5, and s6, the standardized Chinese percentage is above the average value, while in neighborhoods s2, s3, and s4, the standardized Chinese percentage is below the average value. Neighborhoods s5 and s6 are in the darkest tones on the map indicating the highest Chinese percentages. There are just a few census tracts in neighborhood s5 or s6. These two neighborhood types are around the center of the study area. Moreover, areas of neighborhood s5 look like satellite ethnic concentrations expanded from areas of neighborhood s6.

Therefore, while many Chinese immigrants live in low SES and Stability neighborhoods, a decent portion of the population lives in middle and upper levels using the six-cluster strategy. However, dividing data into two clusters would generate different results: most Chinese immigrants live in neighborhood type 2 (low SES and Stability). The following analysis will use the two neighborhood dividing strategies so that each neighborhood's data size would not be too small. After classifying census tracts into different neighborhoods, the next step is to transfer neighborhood classifications to the PUMA level.



V-8a Neighborhood types



V-8b Chinese percentages

Figure V-8 Chinese percentages and neighborhood types by Census Tract

5.4 Transferring Neighborhood Classification to PUMAs

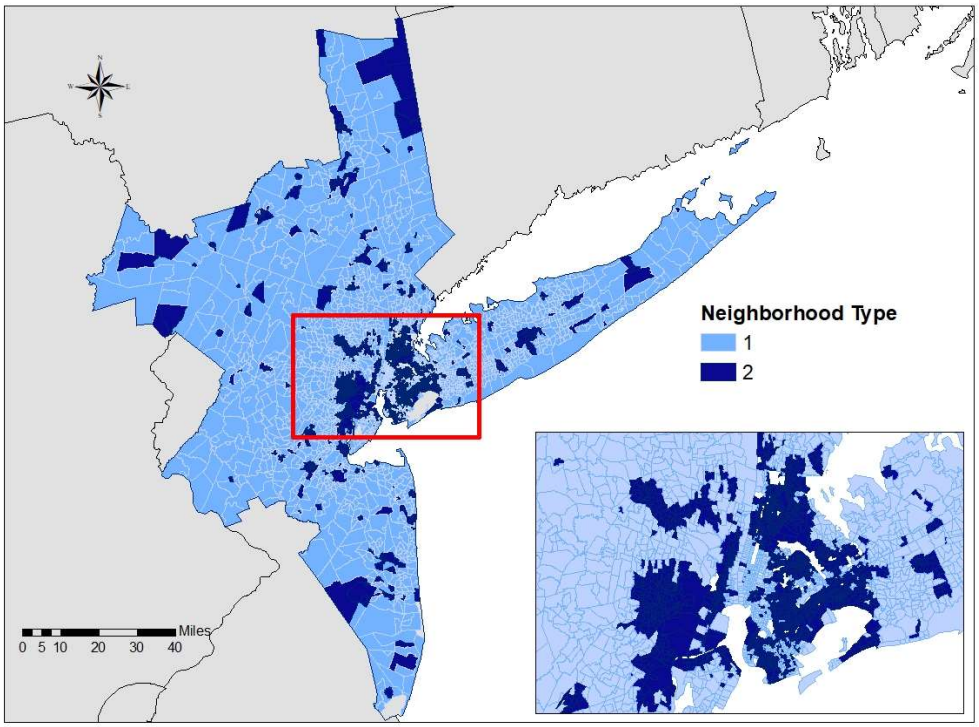
Since census tracts are building blocks of PUMAs, one PUMA contains several census tracts. Identifying a PUMA's neighborhood type includes checking each census tract's type and frequency in the PUMA. A PUMA's neighborhood type is the neighborhood type with the highest frequency in its census tracts. The optimal population of a census tract is around 4000, ranging between 1200 and 8000 (US Census Bureau 2019). Because the population's size in each census tract varies in a reasonable range, identifying a neighborhood's type does not consider the impacts of its population size.

The calculation steps are as follows. The first step is to build the relationship between two feature classes applying the ArcGIS tool: census tracts and PUMAs. A geographic feature class is essentially a data table with locational information. There are generally two ways to study two feature classes' relationships: either through geographic information or non-geographic data tables. A spatial join is an example of the former approach. However, it needs to pay particular attention to topological issues such as slivers and overlaps in performing a spatial join. This research uses a census tract relationship file as the second approach to associate the two feature classes. In the relationship table, the geographic information of a census tract is coded as a specific ID number which uniquely identifies the census tract with the county ID number and the state ID number where the census tract belongs. Another attribute column of the table attached to every census tract is the ID number of PUMAs. It is a many-to-one relationship between census tracts and PUMAs since more than one census tract is within each PUMA.

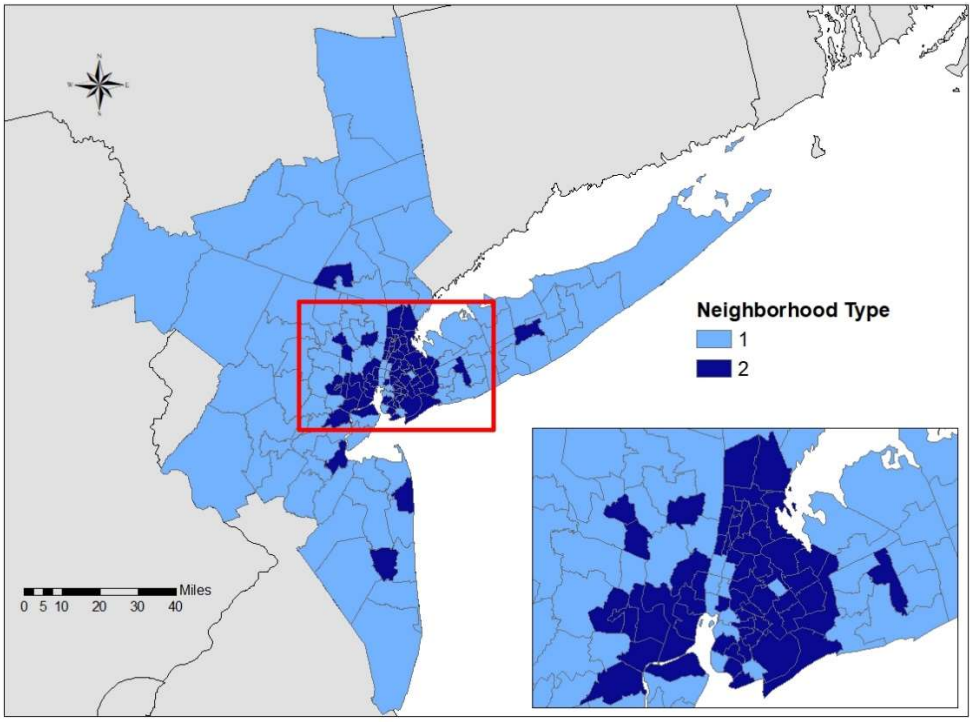
After locating census tracts within their PUMAs, the next step focuses on the constitution of census tracts in each PUMA. Since there are only two neighborhood types, a PUMA's neighborhood type is the type whose appearance is more than 50% in its census tracts. For example, if a PUMA consists of nine census tracts, six are type 1 neighborhoods, and three are type 2 neighborhoods, this PUMA is a

type 1 neighborhood. It is not easy to evaluate the loss of accuracy in transferring neighborhood types from census tract level to PUMA level. However, it is expected to have a better model performance after defining PUMA neighborhood types than viewing the study area as a whole. Model performance statistics, such as R^2, could help evaluate the effect of neighborhood type identifications.

Figure V-9 shows the two neighborhood types on the census tract scale (V-9a) and the PUMA scale (V-9b). There are 87 neighborhoods in type 1 (colored in light grey) and 64 in type 2 (dark grey) on the PUMA scale. The distribution patterns of neighborhoods on the two maps are not entirely the same. "Hot spots" are more evident on the PUMA level. Here hot spots refer to a cluster of census tracts with the same color, mainly neighborhood type 2 in dark grey. They cluster in the center of the study area. There are also some mismatches on the two maps. In Figure V-8a, there are some census tracts of neighborhood type 2 dispersed around the peripheral areas. After transferring the neighborhood defining rules to the PUMA level, the neighborhood types of those isolated areas become the same as their surrounding areas. On the map, there is a spatial pattern that similar values cluster. A PUMA of neighborhood type 1 has a bigger chance to be adjacent to an area of the same neighborhood type. This clustering rule applies to PUMAs of neighborhood type 2 as well. A global Moran index of 0.29 suggests that the spatial pattern of neighborhoods at the PUMA level is not random, which agrees with the spatial pattern on the map.



V-9a Census tract neighborhoods



V-9b PUMA neighborhoods

Figure V-9 Neighborhood types

5.5 Conclusions

In this chapter, three indicators were derived or chosen from the original data set to catch the spatial variations of the study area. Two indicators are synthetic variables: a stability index and an SES index, with each representing a cluster of strongly correlated variables. With these indexes, the NY-N-JC PUMAs can be divided into two neighborhood types. They display a core-and-peripheral pattern in the NY-N-JC MSA. Most type 1 neighborhoods are in the peripheral regions, and most type 2 neighborhoods cluster in the center of the study area (except Manhattan). Neighborhood type 1 is high in both SES and Stability; neighborhood type 2 is low in both indexes.

The third indicator is the Chinese percentage in each PUMA. Due to the small population size, I needed to split the data into six clusters further to reveal the distribution patterns of Chinese immigrants and their relations with the other two indexes (Stability and SES). These relations are covered and unnoticeable when the research focus is on the total population of an area. The type of neighborhood with the highest Chinese percentage includes 92 census tracts. Almost half of the people are Chinese in these areas. Low SES and Stability characterize these areas. Following that are 119 census tracts whose SES and Stability are slightly above the average of the study area. The Chinese percentage in this neighborhood type is close to one-third. The last neighborhood type with a Chinese percentage greater than the average has the highest SES and decent Stability. This neighborhood type has 550 census tracts.

With the increase in the number of clusters, there is more information I can gain in the clustering analysis. However, I will choose to divide the 150 PUMAs into two clusters to guarantee that each subgroup has a big enough sample size for regression analysis. Moreover, since the PUMS data are only about Chinese immigrants (no White people or other ethnic groups), their migration behavior would not be covered. In closing, neighborhood clustering analysis laid the foundation for regression

analysis in chapter VII. Regression analysis built on each neighborhood is expected to reveal spatial patterns within different migration behavior.

CHAPTER VI

OLS REGRESSION

6.1 Introduction

To analyze the Chinese immigrants' migration status in the NY-N-JC MSA, I apply three different regression methods: OLS in this chapter, regression on individual neighborhoods in Chapter VII, and MGWR in Chapter VIII. Each chapter offers a different angle to Research Question 3: How do local factors in the New York-Newark-Jersey City MSA impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially?

The purposes of this chapter are three folds. The first purpose is to evaluate the predictive power of migration-related factors at the PUMA level. The results are average estimations across the study area since OLS is a global model. OLS regression cannot show the spatial variations of relationships. It mainly serves as a reference point for the other two regression methods, which is the second purpose of this chapter. The final purpose of applying OLS models is to select the most significant predictors at the PUMA level. These independent variables are the basis for deriving independent variables for individual neighborhoods in Chapter VII and MGWR in Chapter VIII.

MGWR models could have different independent variables from the OLS model, but the MGWR software does not have the function yet. Therefore, the selection of independent variables in the MGWR models will be based on OLS models.

The rest of this chapter is organized as follows. The first section aims to aggregate microdata into areal data for regression analysis at the PUMA level. The methodology section describes and compares some selection criteria and selection procedures for regression models. I will use the between-states migration to go through the selection process since it is the most studied category in previous research. I will build twenty-seven OLS models on the migration pattern of those who moved between states. One best model will be selected. The model's parameters will be passed into the OLS models for other migration categories, which will be explained in detail in the model fit section. The last section of this chapter presents some discussions and conclusions.

6.2 Data Aggregation

The PUMS data about Chinese immigrants are at the individual level. In Chapter IV, the decision tree algorithm has selected the most significant variables related to the migration study. Among the variables, some are numerical, and some are categorical. The difference in their structure leads to different aggregation processes. For each PUMA, weighted means are derived for numerical variables and weighted ratios for categorical dummy variables. Numerical variables are age, number of bedrooms, rentals, and income. For numerical variables, mean and median are two standard statistics for calculating average values. Median is the value of the middle point in a vector of data. It does not consider the impact of data magnitudes other than sorting data. On the contrary, the calculation of means is based on data values rather than their orders. The measure of mean is more representative of the data as long as there is no outlier problem. In this work, a weighted mean is calculated to summarize numerical variables after cleaning outliers.

Outliers in this data set include extremely large values and values of zero. Almost half of the records for the variable rent are zero. A code of zero represents not applicable situations, such as owning a house. But other possibilities exist as well. For example, people may not want to report their rent. In the calculation, rents of zero were removed from the data set. After data aggregation, eleven PUMAs are missing rent values.

Income is always a sensitive question in surveys and thus having many missing values. The inclusion of occupational income scores was to compensate for missing income values. However, there are about 32% missing values in the variable occupation score, which did not improve the data quality much. Therefore, the following analysis will keep the income variable and remove the occupation score. In calculating the weighted mean of income, missing values are dropped. It is also reasonable to drop records of zero income and top-coded income. An income of zero may indicate self-employment or unemployment. It has no contribution to the mean income of PUMA areas.

Compared to numerical variables, it is more complicated to process categorical variables. The first step is to derive dummy variables. A dummy variable corresponds to one level of a categorical variable (as shown in Table VI-1). For example, the citizenship status variable has three levels: born abroad of American parents, naturalized citizens, and not citizens. Each level turns into one dummy variable. If we know the values of two dummy variables, we can easily calculate the third one. Therefore, it is necessary to drop one dummy variable in the same group to avoid multi-collinearity. After transferring categorical variables into dummy variables, it is not hard to calculate each dummy variable's weighted percentage as the summary statistic. There are twenty-seven variables after the aggregation process, with four weighted mean of numerical variables and twenty-three weighted percentages from categorical variables.

Categorical Variables			
Categorical Variable	Level		If drop (×)
migration status (dependent variable)	1	Same house	
	2	Moved within state	
	3	Moved between states	
	4	Abroad one year ago	
race	400	Chinese	
	410	Taiwanese	×
birthplace	50000	China	
	50010	Hong Kong	
	50040	Taiwan	×
marital status	1	Married, spouse present	
	2	Married, spouse absent	
	3	Separated	
	4	Divorced	
	5	Widowed	×
	6	Never married/single	
college degree	1	With college degree	
	0	Without college degree	×
poverty status	1	Under poverty line	
	0	Above poverty line	
	999	missing	×
citizenship	1	Born abroad of American parents	×
	2	Naturalized citizen	
	3	Not a citizen	
speak only English	0	N/A or blank	×
	1	Does not speak English	
	3	Yes, speaks only English	
	4	Yes, speaks very well	
	5	Yes, speaks well	
	6	Yes, but not well	
employment status	0	N/A	×
	1	Employed	
	2	Unemployed	
	3	Not in labor force	
class of worker	0	N/A	×
	1	Self-employed	
	2	Works for wages	

Table VI-1 Levels of categorical variables

6.3 Methodology

6.3.1 Model selection criteria and strategies

A large number of independent variables adds computational steps and time and complicates model building and interpretation. Applying a collection of model selection and strategies, I can choose the best model with the most significant variables for each migration behavior. The issue of multi-collinearity, rising from interactions among variables, must be taken into consideration as well. To choose the best model is to select variables that contribute most to the model fit by evaluating specific criteria. This research applies three statistics (RMSE, MAE, and adjusted R^2) to select the best regression models. Models derived from different selection criteria may vary.

Standard selection criteria for a regression model are (adjusted) R^2, root mean squared error (RMSE), and mean absolute error (MAE). R^2, the coefficient of multiple determination, is the ratio of variation in the dependent variable explained by independent variables. It measures the explanatory strength of regression models (McGrew Jr and Monroe 2009). The R^2 value increases with the addition of variables, even if a newly added variable is not helpful for the model fit. Adjusted R^2 takes into account the number of independent variables. It has a punishing mechanism for the number of independent variables. If adding a variable does not increase the model fit larger than chance, the adjusted R^2 stays the same or decreases (Bhandari 2020). RMSE and MAE both provide an average model error (Burt et al. 2009). MAE is less sensitive to outliers. Like R^2, these two metrics tend to increase with the addition of variables even if they do not improve the model fit.

Selection Strategies

There are many selection strategies. Forward selection, backward elimination, stepwise regression (or sequential replacement), and all possible models have been widely used. The all possible models method calculates and compares all possible models, preferable for models with a small number of independent variables. The model calculation increases rapidly with the increase of variables. If the

number of independent variables is k, then a total number of $2^k - 1$ models need to be built in general (PennState 2018). With over thirty possible variables in this dissertation, the required number of model fits would be over 1 billion. So the method of all possible models is very impractical to build models with many variables. However, it has a low risk of omitting the best model.

In contrast to the method of all possible models, forward selection and backward elimination are not computationally expensive, even in situations of a large number of independent variables. A traditional forward selection starts with zero independent variables and sequentially adds one variable at a time. Each time the independent variable with the most significant contribution to the model is added. This process continues until no statistically significant improvement happens with the addition of any independent variable. In contrast, a traditional backward elimination model starts with all independent variables and drops one variable with the least contribution to the model fit. The dropping process ends when no variables could be removed based on their significance tests (Kassambara 2017).

A stepwise regression method is a combination of forward selection and backward elimination. It starts with no independent variables and sequentially adds one variable into the model (like a forward selection procedure). Each time after including a new variable, the model performs tests of significance and removes a variable if it is no longer statistically significant (like a backward elimination procedure) (Kassambara 2017). By considering the interactions between independent variables, the stepwise regression method could considerably improve the model performance.

R caret package

The R "caret" library can run 233 different models (Kuhn 2013). The beauty of the library is that we can use the same function to run all models with slight changes in their arguments. The "caret" library incorporates various selection criteria and selection procedures. It also conducts automatic resampling and parameter tuning, seriously reducing the workload.

Data resampling offers a more realistic prediction on model fit than using the original data set. Resampling sets aside testing data and uses the rest of the data for modeling. After repeating this process many times, the model fit's final evaluation is the average of all the repetitions. On the contrary, without resampling, the data set for building a model is the same data set for evaluating model fit metrics, leading to an overfitting issue. The purpose of a predictive model is to determine how well it works on new data sets. Model fit metrics derived on resampled data are not as good as on the original data set. However, the model performance from the resampled data is more reliable (Ross 2017).

In the R "caret" library, the default resampling technique is bootstrap. A bootstrap resampling randomly selects data with replacement. Thus some data may be picked more than once, and some never get chosen. Those chosen data are used to develop a model, and those "out-of-pocket" data are for testing purposes. After repeating this process many times (the default number of repetitions is 25), the program derives a mean evaluation of the model performance (Ross 2017). Bootstrap resampling has a relatively large bias. Compared to the bootstrap method, k-fold cross-validation is a more robust method. The first step of k-fold cross-validation is to split data into k subsets randomly. While one subset is reserved as the test data, the other k-1 subsets are training data. Next is to repeat this process until each subset has a chance to be the test data with a prediction error. The average value of the k prediction errors is the final value used to measure a model's fit (James et al. 2014). Choosing the number of data subsets k is not simple. A small k could introduce more bias, while a big k may generate a wide variance.

The bias-variance tradeoff in data resampling techniques has complicated the model selection. Parameter tuning multiplies the complexity of the study. The tuning method aims to choose the best values of one or more parameters to optimize a model's performance. It is impossible to determine the best value of a tuning parameter using an equation or other analytical algorisms. The parameter tuning approach passes different values to the model. By evaluating the performance of different

models, the practitioner can determine the best parameter value. In a regression model, the number of independent variables is a tuning parameter when the data have a high dimension. Different modeling approaches, such as forward, backward or stepwise regression, affect the selection order and the number of independent variables in a regression model.

6.3.2 Regression Models

Table VI-2 represents twenty-seven regression models for the migration pattern of moved between states. The table shows model details, including adjusted R^2 and predictors in the regression models. These models vary in modeling approaches: forward, backward or stepwise regression. They are also different in the value of k for the k-fold cross-validation, the number of data groups produced in the resampling process. Since the k-fold cross-validation resampling randomly assigns data into subgroups, these subgroups contain different data points each time, even for the same k. Therefore, three models are built for each k in the effort of covering diverse situations.

In terms of the adjusted R^2, 5-fold cross-validation resampling is more likely to produce relatively high R^2 values than 4-fold and 3-fold. Models built with stepwise regression have a higher value of adjusted R^2, followed by the forward regression and the backward elimination models. In the table, the highest value for adjusted R^2 is 0.29. There are ten models whose adjusted R^2 values are above 0.2. Among those ten good models, only two are built with backward elimination regression. Therefore, in the following analysis, other migration categories will apply 5-fold cross-validation resampling combined with stepwise regression. One note here is that this combination may not always produce the best model. There is no agreement on the selection of resampling and modeling techniques. The performance of a specific method varies with data.

Model selection Migration pattern: moved between-states							
Model	adj. R^2	Independent Variables					
		college	naturalized	single	separated	employed	Not self-employed
k-fold cross validation (k = 5)							
1-forward	0.26	√	√				
1-backward	0.13	√	√	√	√	√	
1-stepwise	0.29	√	√	√	√		√
2-forward	0.19	√	√				
2-backward	0.25	√	√	√	√	√	
2-stepwise	0.15	√	√				
3-forward	0.24	√	√	√	√		√
3-backward	0.13	√	√	√	√		
3-stepwise	0.27	√	√	√	√		√
k-fold cross validation (k = 4)							
4-forward	0.17	√	√	√	√		√
4-backward	0.21	√	√	√	√	√	
4-stepwise	0.22	√	√	√	√		
5-forward	0.23	√	√	√	√		
5-backward	0.09	√	√	√	√	√	
5-stepwise	0.14	√	√	√	√		√
6-forward	0.17	√	√		√		
6-backward	0.12	√	√	√	√		
6-stepwise	0.06	√	√		√		
k-fold cross validation (k = 3)							
7-forward	0.23	√	√	√	√		√
7-backward	0.09	√	√				
7-stepwise	0.16	√	√		√		
8-forward	0.11	√	√	√	√		√
8-backward	0.18	√	√	√	√	√	
8-stepwise	0.12	√	√		√		
9-forward	0.22	√	√				
9-backward	0.12	√	√		√		
9-stepwise	0.09	√	√	√	√		√

Table VI-2 Model selection for between-states migration

In this study, the sample size of 150 PUMAs is relatively small. If k is too big, the size of its subset data would be too small to be representative of the population. Model selection is not the major research topic of this study. If so, 270, instead of 27 models, would be more persuasive. A locally high value of adjusted R^2 could appear with any number of predictor variables. An increase in the number of predictor variables in a model does not necessarily increase the adjusted R^2.

6.4 Model fit

As discussed earlier, stepwise regression models built with 5-fold cross-validation sampling offer the highest R^2 on average. Therefore, the same modeling approach is applied to analyze other migration patterns, including the form and strength of models and assumption validations. The table below summarizes the values of adjusted R^2 and predictors for each migration pattern. The regression model for being abroad (one year ago) has the most predictive strength, followed by moved between states and moved within a state. The regression model for the within-state migration performs poorly with an adjusted R^2 of 0.1. The maximum number of predictor variables in a regression model is set as five. The regression results show that the migration patterns vary in their most significant predictors.

Regression results for different migration patterns (5-fold cross-validation, stepwise)						
Migration pattern	adj. R^2	Predictors				
		V1	V2	V3	V4	V5
Moved between-states	0.38	naturalized	college	separated	single	not self-employed
Moved within state	0.10	naturalized	self-employed			
Abroad (one year ago)	0.46	age	not in labor	speak English well	single	wage
Same house	0.37	naturalized	age	college	married	speak only English

Table VI-3 Regression results for different migration patterns

6.4.1 Migration pattern: moved between states

There are five variables in the final model for predicting a between-states migration. Based on their orders included in the stepwise regression, they are naturalized citizenship, college degrees and above, separated, single, and not self-employed. They account for about 38 percent of the variation in a between-states migration. All five predictor variables have a variance inflation factor (VIF) around

one. Therefore, multi-collinearity is not a problem in this regression model. A rule of thumb is that a desirable value of VIF should be less than five (O'brien 2007).

The variable of naturalized citizenship has a negative relation with the dependent variable. Areas with a high percent of naturalized citizens tie to a small percentage of between-states migrations. The other four variables all have a positive relationship with the dependent variable. College is a crucial predictor of a between-states migration. People with a college degree or above have a far greater likelihood of moving to a different state than those who do not possess a college degree. Marital status variables play a vital role in a between-states migration as well. An individual whose marital status is single or separated has fewer family-wise concerns or obstacles to make a between-states move than a married individual. Areas with better-paid wages are likely to attract immigrants from other states. In contrast, a relatively high percent of naturalized citizens indicates a more stable place with fewer between-states migrations.

The magnitude and direction of residuals provide information to evaluate the fit of a regression model. In Figure VI-1, the first graph shows the distribution of residuals against fitted values. Except for several outliers (numbered 11, 29, and 82), the residual values range between -10 and 10, with most fitted values ranging between -5 and 5. The second graph (top right) is a normal Q-Q plot to check if the residuals obey a normal distribution. The majority of data points are along the expected line except for some outliers in the tail. The pattern of the scatter points on the third graph agrees with the first graph. Compared to the first graph, its y-axis is the square root of standardized residuals instead of the original residual values. The fourth graph is a scatter plot of standardized residuals against leverage with Cook's distance contours lines. The closer to Cook's distance contours, the more influential a data point is to the parameter estimates (Crawley 2012). Three data points in the graph are marked as influential: points 49, 82, and 91. In summary, from the above four plots, we can obtain a basic idea that the residuals are generally constant and obey a normal distribution. However, some outliers exist, which may have affected the final model estimates.

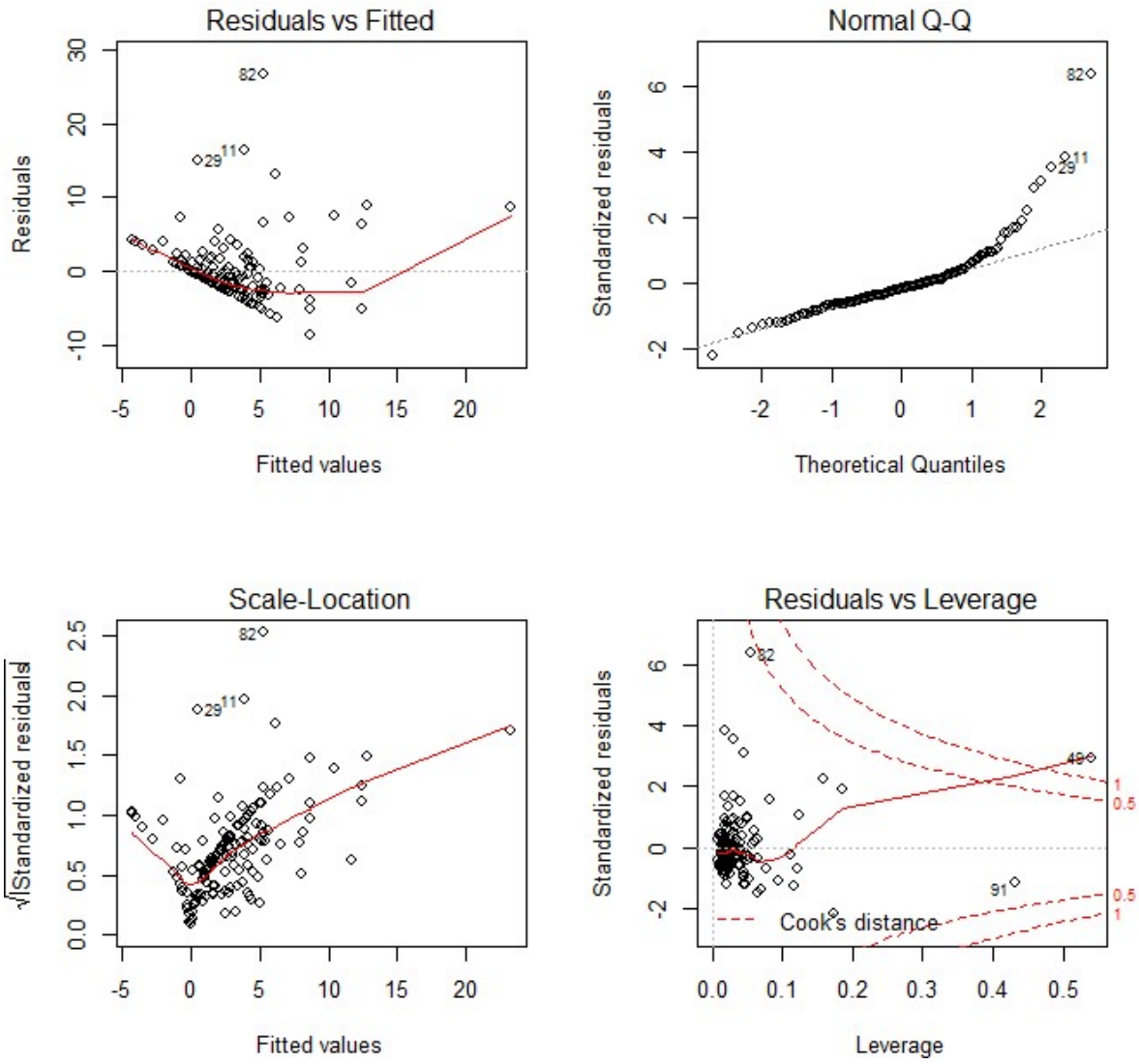


Figure VI-1 Residual assessment (migration status: moved between states)

Figure VI-2 displays the regression residuals of the between-states migration for the Chinese immigrants. On the map, warm colors (shades of brown) indicate positive residuals, whereas cold colors (shades of blue) indicate negative residuals. This rule applies to the residual maps of other migration patterns in this chapter as well. Since the data have been normalized before building the regression model, they have a mean of zero and a standard deviation of one. The dominant color on the map is light blue, indicating that the regression model overestimates the percentage of between-states migrations in these PUMAs. The errors in these overestimated areas range between zero and 4.2 standard deviations. On the contrary, in those areas colored in light brown, the regression model

underestimates their percentages of between-states migrations. Other areas shaded in dark colors (brown or blue) scatter around on the map, with no apparent pattern.

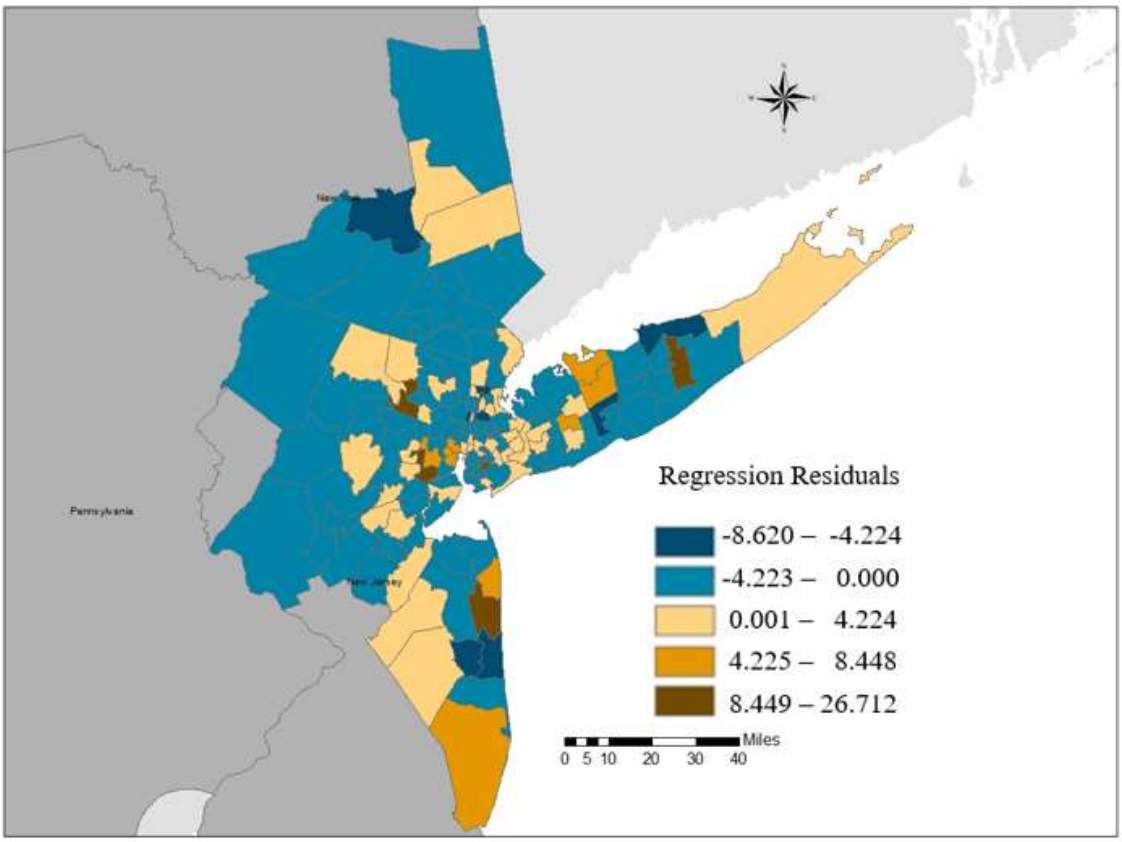


Figure VI-2 Map of residuals in modeling the between-states migration

Dubin Watson Test is for checking the independence of residuals. The result is not significant, so there is no reason to reject the hypothesis that the residuals are independent of each other. Global Moran's I derived from ArcMap agrees with the Dubin Watson test. A Moran index of about zero suggests that the residual pattern is not significantly different from random.

6.4.2 Migration pattern: moved within the state

Regression for the in-state migration performs poorly. The independent variables explain about a ten percent variation in the dependent variable. There are two predictors in the model: naturalized citizenships and self-employed. Based on the VIF results (around one), the regression model does not have a multi-collinearity problem. The predictors contribute little to an in-state migration behavior. A

high percentage of naturalized citizens serves as a hindrance to in-state migration. On the contrary, self-employed workers tend to move within the state. Too many reasons could lead to within-state migrations, so the regression result is poor for within-state migrations.

In Figure VI-3, the first graph (top left) and the third graph (bottom left) plot residuals against fitted values in the regression model of in-state migration in different scales. There is no obvious scatter pattern in the first plot, such as an S-shape or a banana shape. Also, the width of points does not change with the increase of fitted values, which is good. However, there are two minor issues. The first issue is the existence of outliers numbered 29, 74, and 94. Also, there is a downward trend in the scatter of the first graph (shown in the red line compared to the dotted line). As fitted values grow, there are more negative residuals which means that the dependent variable is overestimated for bigger values. The third graph illustrates the same issue of outliers as the first graph. If those outliers were removed, the scatter's width along the y-axis stays almost the same except for the first several data points whose fitted values are close to zero. While the overall trend shown in the red line has a slight upward trend, it is not a pronounced one. On the second graph of the normal Q-Q plot, most data points are very close to the dotted line representing a normal distribution of residuals. The exceptions are several outliers numbered 29, 74, and 94. The fourth graph highlights several influential points: 10, 94, and 117, which are very close to Cook's distance contours. Datapoint 10 has dragged the overall data trend toward the end. In summary, the regression model did not capture many variations in the variable of in-state migration. Its regression residuals are generally constant and are normally distributed, except for some outliers.

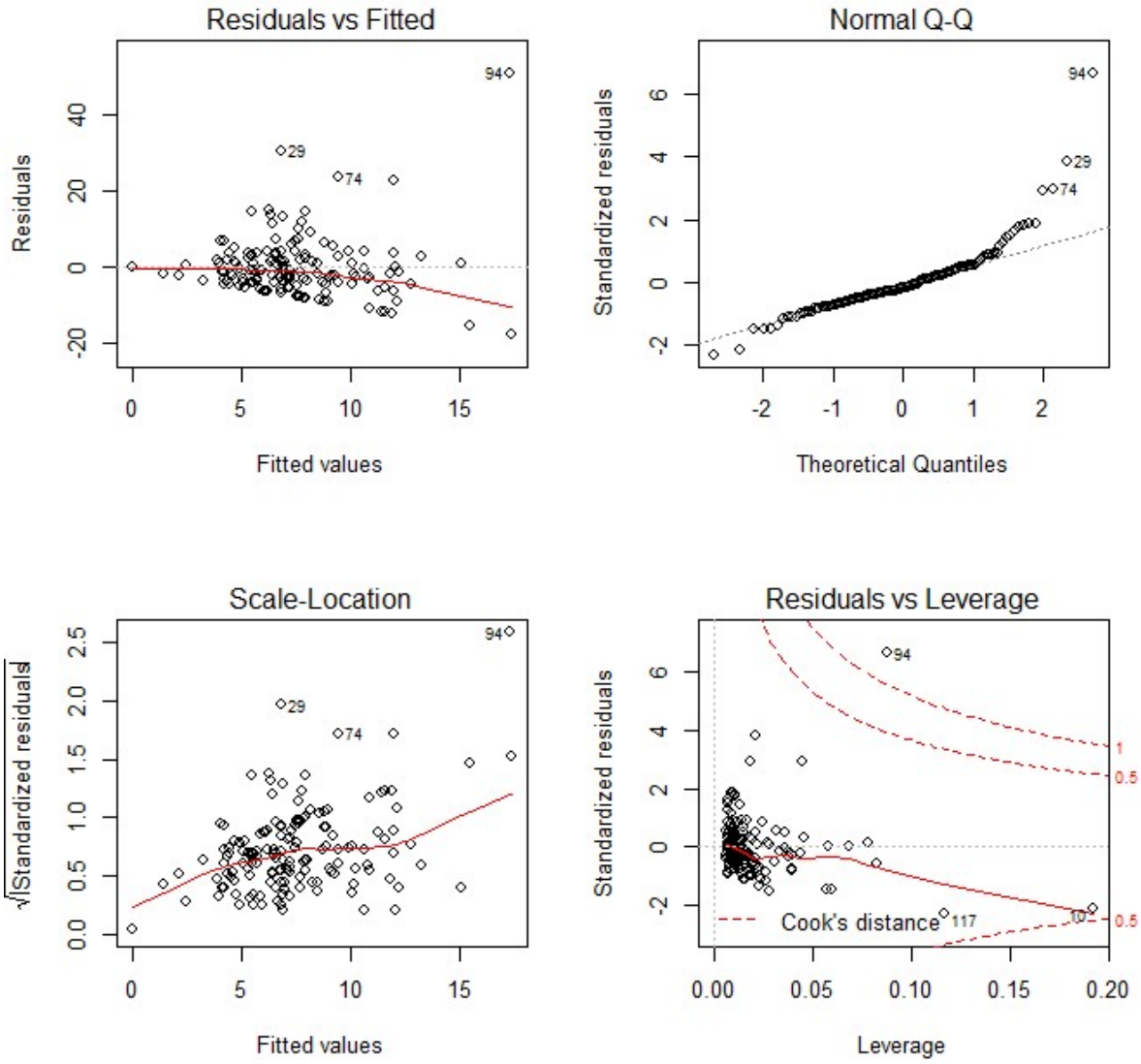


Figure VI-3 Residual assessment (migration status: moved within state)

Figure VI-4 shows the regression residuals for the within-state migration status. The residuals have a wide range between -17.4 and 51, indicating that the regression model does not fit the data well. There is no clear color pattern on the map. A significant number of PUMAs (shaded in light blue) are overestimated in their within-state migration percentages. Underestimated PUMAs cluster around New York City extending along the Long Island. The result of the Durbin Watson test is not significant, so there is no reason to reject the hypothesis that the residuals are independent. However, Moran's I statistic suggests a cluster distribution, and the likelihood is less than 10% that the clustered pattern is a random result.

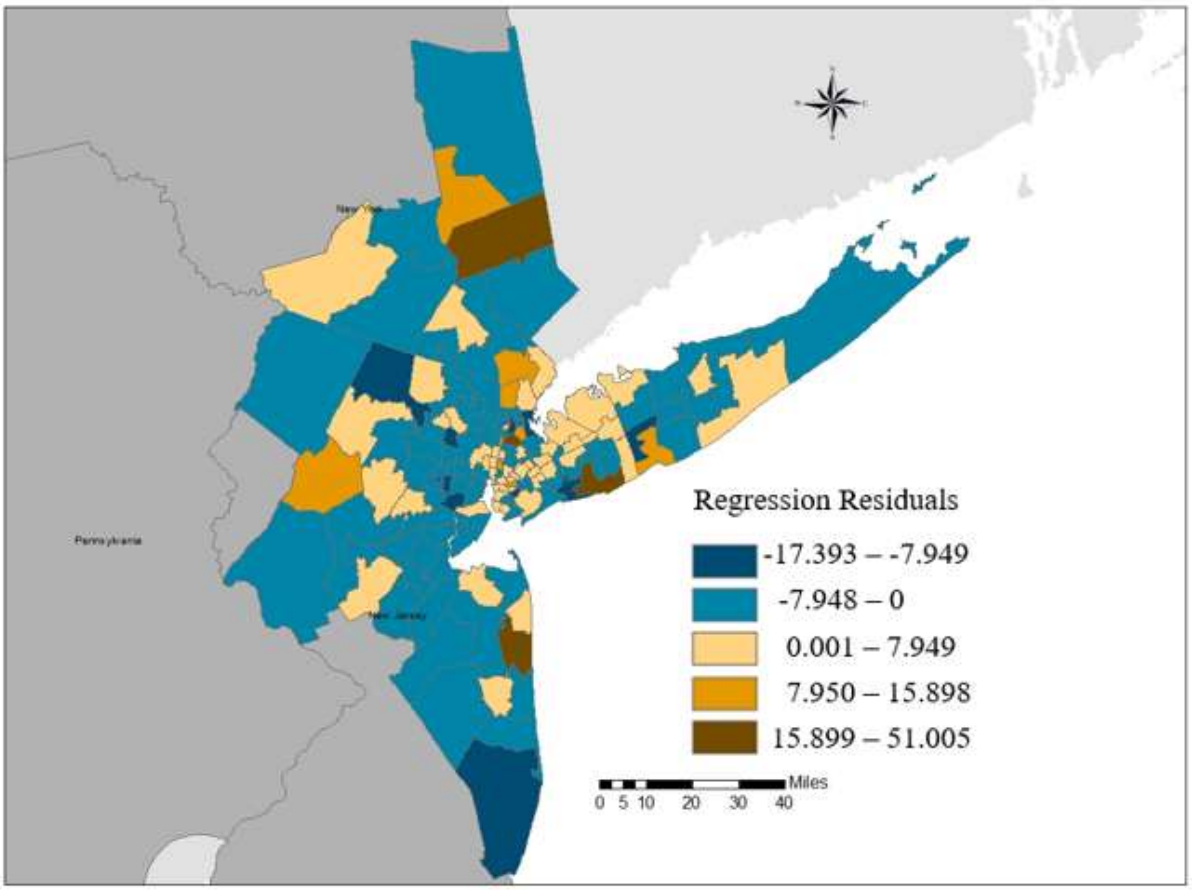


Figure VI-4 Map of residuals in building the within-state migration

6.4.3 Migration pattern: abroad one year ago

Regression on the population who was abroad one year ago has the highest adjusted R^2 (0.46) among all migration patterns. There are five independent variables in the final model: age, not in the labor force, speak English well, single, and wage. There is no collinearity issue in the regression model, with VIF values ranging between one and two. Age is the first added predictor in the regression model, negatively related to the dependent variable. Younger respondents are more likely to move to the US compared to older people. This migration group of the Chinese population has some shared characteristics: not in the labor force, speaking English well, and being single. An increase in wages is a cause of moving to the US. This group of people has been identified in an earlier chapter of decision tree analysis. They are education-purpose movers from mainland China in the data set.

In the regression model for the migration behavior of people who were abroad before, the influence from outliers is significant. Those influential data points are highlighted in the graph of residuals against leverage. They are data points 39, 55, and 149. These outlier points also deviated far away from the remaining points in the normal Q-Q plot. In both the first and third graphs, those outliers force the red lines to deviate from the original directions. If those outliers were removed, the scatters would not broaden with the increase of fitted values in the first and third graphs. In the first graph, the scatter trend (shown as a red line) is a curved line. The first half of the line has a downward trend. After passing the fitted value around 10, the red line turns up. This upward trend of the second section of the line results from outliers.

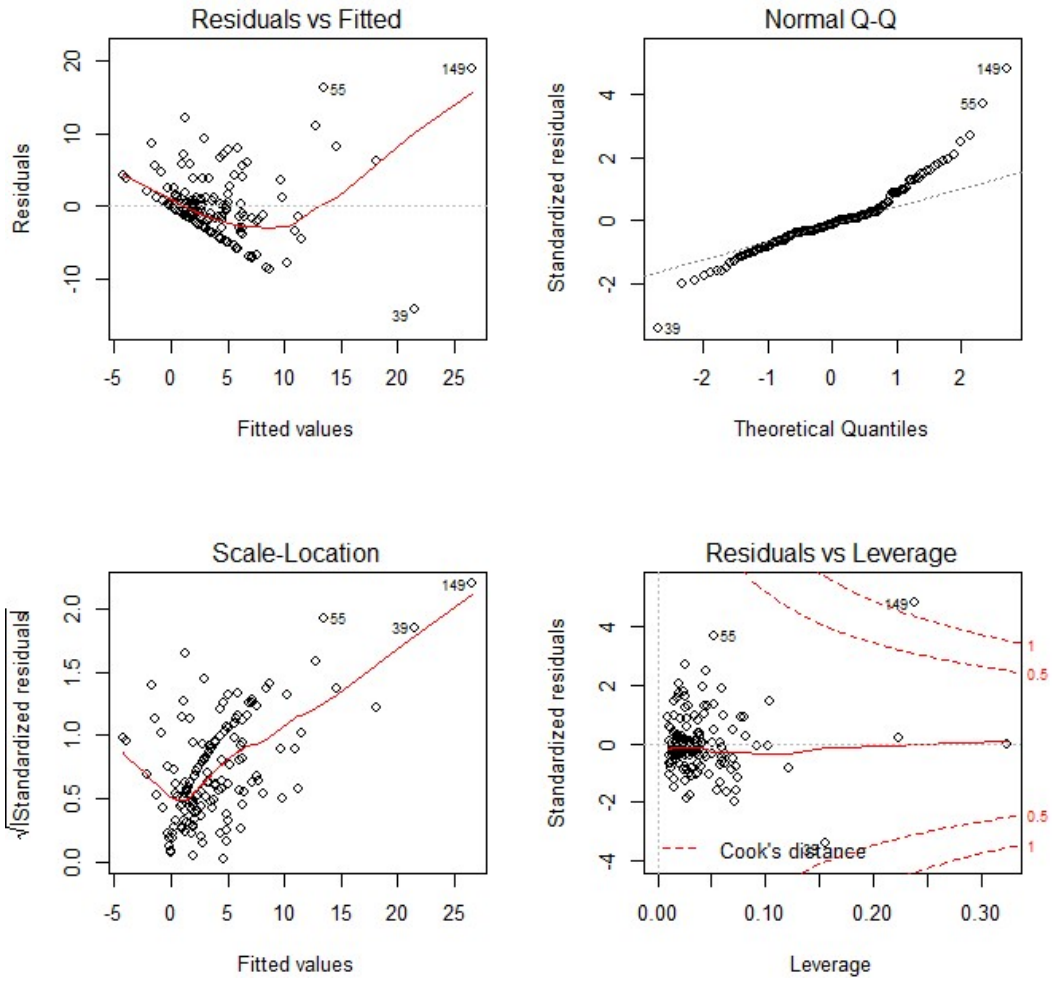


Figure VI-5 Residual assessment (migration status: abroad one year ago)

Figure VI-6 shows the regression residuals in modeling the migration status of a particular group of people. These individuals were abroad before and then moved into the US. There is no observable pattern on the residual map. Most PUMAs fall in the two legend groups with the smallest errors, with light blue indicating negative errors and light brown positive errors. The Durbin Watson test suggests that it is better not to reject the hypothesis of independent residuals. In agreement with the Durbin Watson test, Moran's I index indicates a random distribution of residuals.

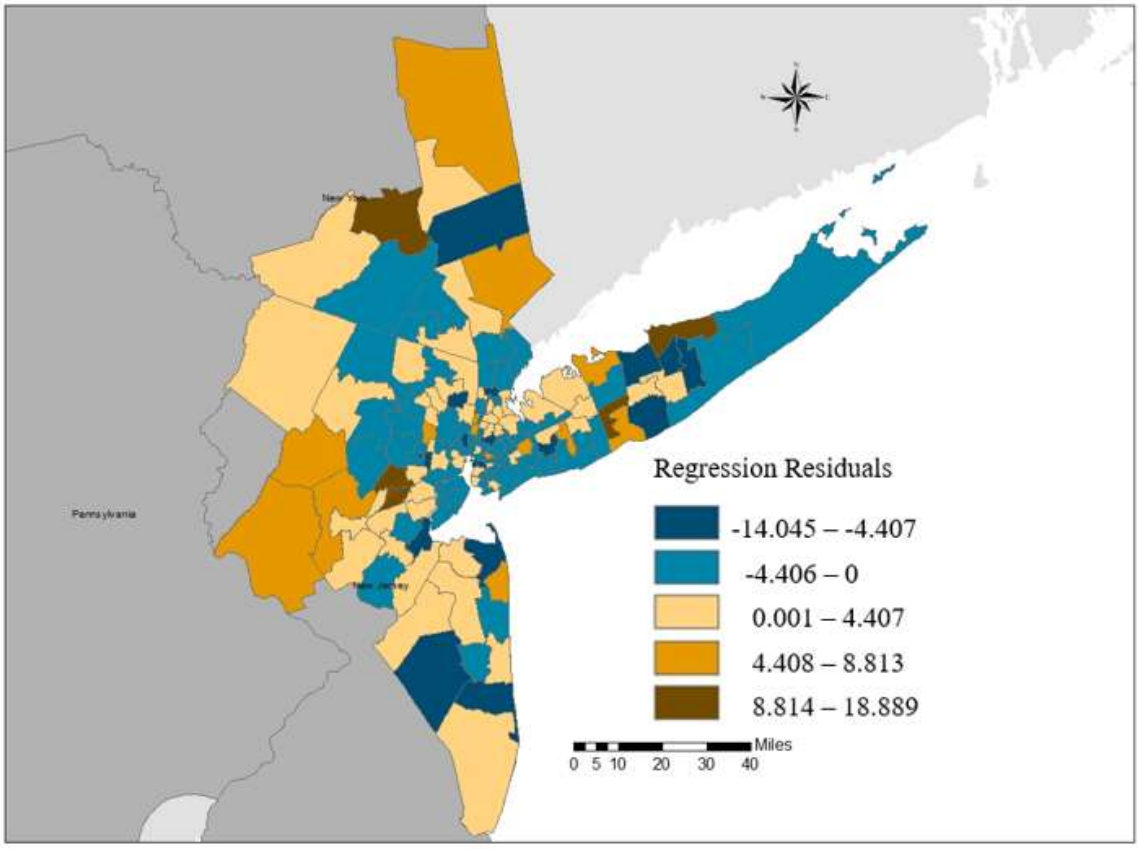


Figure VI-6 Map of residuals in building the migration status of abroad one year ago

6.4.4 Migration pattern: same house

Five predictors entered the regression model for people who stay at the same house: naturalized citizenship, age, college, married, and speak only English. There is no multi-collinearity in the regression model, in which the VIF values of independent variables are less than two. The variable naturalized citizenship explains almost 22 percent variation in the dependent variable. A high percentage of naturalized citizens ties to a high percent of non-movers. People are less likely to move

when they get older or married. The variable speaking only English has a positive relationship with the dependent variable as well. This group of Chinese people could be those who have been residing in the nation for a long time. They have passed the moving stage and have settled down in one place. On the contrary, high educational degrees encourage migration behaviors.

Similar to regression models for the other three migration response variables, outliers in Figure VI-7 are an issue. They are points numbered 29, 31, 82, and 94. Those outliers are close to Cook's distance contours and far away from the rest data points. Regardless of outliers, on the first plot, data points gathered into a ball shape around the fitted value of 85. The scatter ball is the widest around 85, indicating the greatest variance of residuals. The third graph shows the same pattern. On the second graph of the normal Q-Q plot, the middle part of the line formed by the data points is relatively straight. But two ends of the line start to deviate from the ideal line.

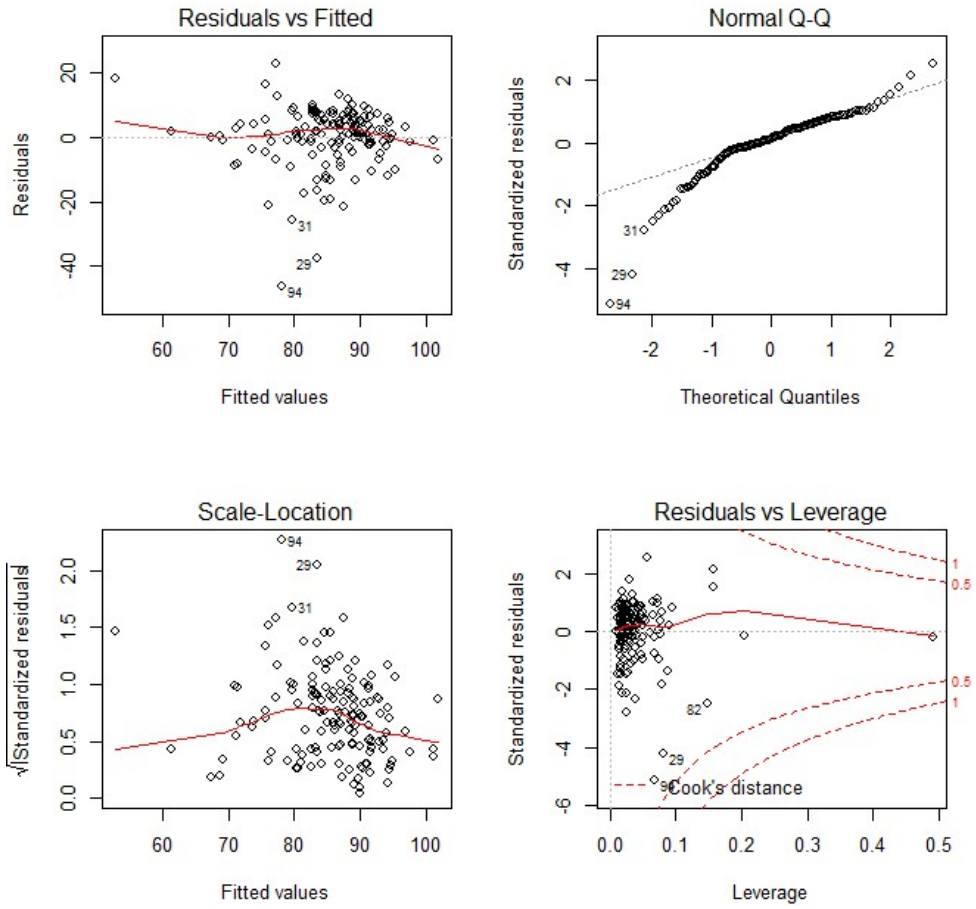


Figure VI-7 Residual assessment (migration status: same house)

Figure VI-8 is the residual map showing the migration status of staying in the same house. There is no obvious pattern on the map. Areas with light brown occur the most often. The regression model underestimates the percentages of people staying at the same house in these areas. Both Moran's I index and the Durbin Watson test suggest no autocorrelation in the model, and residuals are independent.

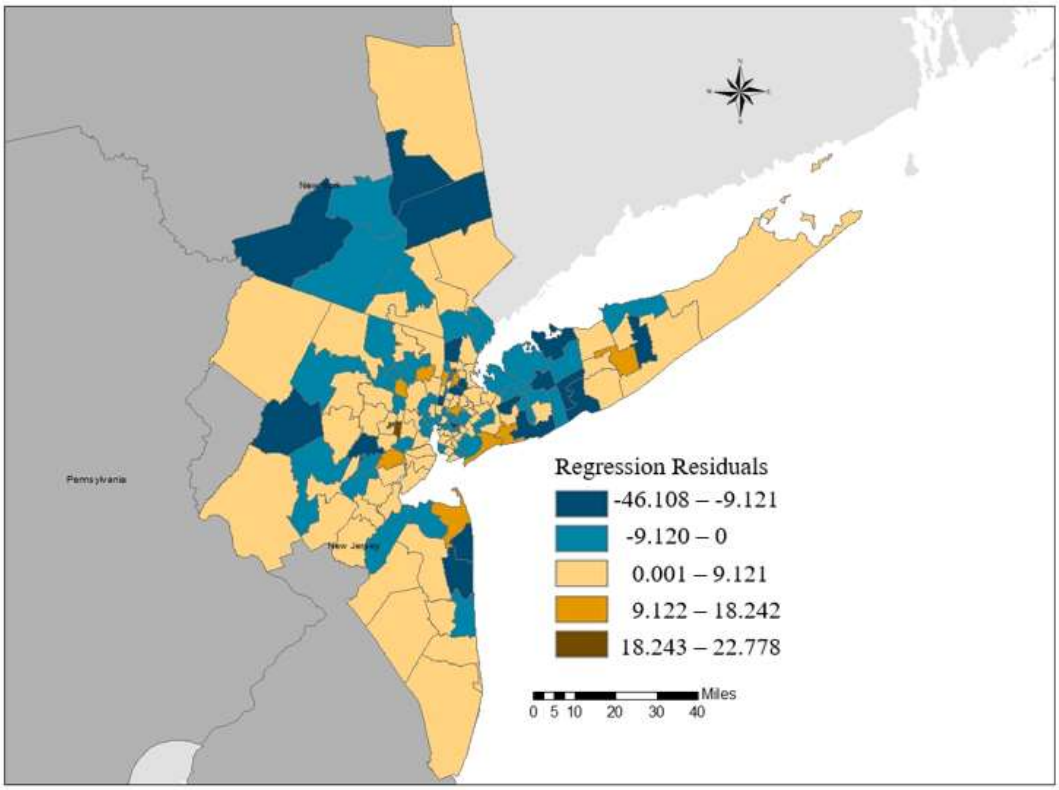


Figure VI-8 Map of residuals in building the migration status of same house

6.5 Conclusions

In summary, the above migration models differ from each other while sharing some common traits. Predictors in the above models fall into one of these categories: personal capability, family information, and job-related factors. Personal capability includes English fluency and educational attainment. As a foreigner, speaking English well is crucial in the migration process for migration from different states or abroad. Places with a high percentage of college degrees are associated with a high between-states moves. Citizenship status is a reflection of both people's capability and in-country time. Citizenship status is a significant indicator of being stable. Since it takes time for a foreigner to obtain citizenship, citizenship status often indicates an older individual. Family information mainly refers to marital status. Marital status has a significant influence on migration. There are more obstacles and considerations for a married person to move than a single or separated

person. Employment is the most crucial job-related factor, particularly for migration from different states or abroad.

The OLS regression results offer an examination of migration-related factors at the PUMA scale. Some results are consistent with decision tree analysis, while some are not. Employment and educational attainment have positive influences on migration behavior, as indicated in the chapter of decision tree analysis. However, naturalized citizenship act as a stable factor for people who are less likely to move in this chapter, whereas it promotes social mobility in decision tree analysis. Moreover, self-employment is not a significant factor, whereas English proficiency is included in the regression model at the PUMA scale.

The amount of variation captured in the above models ranges between 0.1 for within-state migration and 0.46 for people who migrated from abroad. In terms of the regression residuals, while outliers cause some deviations, there is no visible geographical pattern in the distribution of residual values. However, it does not mean that the regression models have captured all spatial variations of the migration patterns and underlying processes. A further investigation is still needed. One reason is that the impacts of factors could be canceled out by each other in regression residuals. Another possibility is related to scales. It is meaningless to discuss a spatial pattern without mentioning the scale in which the pattern appears. Therefore, an accurate choice of scale is crucial to studying spatial patterns since whereas a different scale could disguise the expected patterns. In closing, while OLS regression demonstrates some common trends and indicators for diverse migration behavior, it does not reveal any spatial variations of underlying processes.

CHAPTER VII

REGRESSION ON NEIGHBORHOODS

7.1 Introduction

This study applies three regression methods to address Research Question 3: How do local factors impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially? The first method is OLS regression, as demonstrated in the previous chapter. While OLS models gave an estimate of the relationships between migration behavior and factors, they did not help with the second half of my third research question: how the relationships vary spatially. To study the spatial variation of relationships in migration is a main focus of this study. Therefore, I put more weight on this chapter and MGWR in the following chapter, with each offering a different angle. I will run regression analysis on each neighborhood type separately in this chapter as an approach to reveal the spatial variation of relationships in migration behavior.

In Chapter V, The NY-N-JC study area is classified into two neighborhood types based on their demographic, socio-economic, and migration characteristics. Based on the classification results, I will first select the most significant migration indicators for each neighborhood type in the following section. Next is the analysis of diverse migration behavior in each neighborhood. I will

further compare the regression results from the two PUMA neighborhoods with those from the study area as a whole.

7.2 Significant predictor for each neighborhood

Neighborhood type 1

For both neighborhood type 1 and type 2, stepwise regression with a 3-fold cross-validation resampling is chosen as the modeling method. This regression modeling method is the same as the regression model on the whole study area, except that this chapter uses a 3-fold instead of a 5-fold cross-validation resampling. The decrease in the number of data subsets results from the small data size in each neighborhood. Table VII-1 summarizes typical regression results (adjusted R^2 and predictors) for different migration patterns in neighborhood type 1. One benefit of running cross-validation in R is that we can repeat the analysis with the same block of codes. Each time the software assigns data into subgroups serving different purposes in the modeling process. Several models repeatedly appear for each migration behavior (see Table VII-1).

In the between-states migration of neighborhood type 1, two models are listed. They are similar in the values of adjusted R^2 (between 0.1 and 0.2) but differ significantly in the set of their predictors. The first model suggests three powerful predictors: college degrees, naturalized citizenship, and employment status. The second model has an entirely different set of predictors: the number of bedrooms, age, married, and spouse absence.

For the in-state migration behavior, three models are listed. The spouse absence variable is very predictive of the in-state migration since it is included in all three models. It is the only predictor in the third model with the highest adjusted R^2 compared to the others. Other predictor variables for in-state migration are self-employment, single, born in Hong Kong, and speaking only English. The adjusted R^2 ranges between 0.16 and 0.33 in the listed models.

Two models are listed for people who moved from abroad. The magnitudes of adjusted R^2 range between 0.1 and 0.2. Selected predictor variables are age, single, speaking English well, speaking English not well, and not in the labor force.

The last group of people, staying at the same house, has two models listed in the table. These two models are close in their adjusted R^2 and predictor variables. The values of the adjusted R^2 range from 0.18 to 0.27. The predictor variables are married, naturalized citizenship, and college degrees.

Regression results of migration behaviors for neighborhood type 1 (3 fold Cross-Validation, Stepwise)						
Migration status	adj. R^2	Predictors				
		V1	V2	V3	V4	V5
Between-states	0.19	college degrees	naturalized citizenship	wage		
	0.14	No. of bedrooms	age	married	spouse absence	
In-state	0.16	spouse absence	self-employment			
	0.29	spouse absence	singe	born in Hong Kong	speaking only English	self-employment
	0.33	spouse absence				
Abroad	0.21	age	speaking English well	not in labor		
	0.13	age	single	speaking English well	speaking English not well	not in labor
Same house	0.18	married	naturalized citizenship			
	0.27	married	college degrees	naturalized citizenship		

Table VII-1 Regression results for migration behavior of neighborhood type 1

Neighborhood type 2

Table VII-2 lists regression results for neighborhood type 2. Among all the migration patterns, the between-states migration model has the greatest adjusted R^2 value (0.38). Three predictors are in the model: married, separated, and educational attainment. The values of adjusted R^2 in in-state migration

models are the lowest among all migration patterns. They are all less than 0.1. With such a low $R^{2,}$ it is hard to say that any predictor is strong in predicting an in-state migration.

Regression results of migration behaviors for neighborhood type 2 (3 fold Cross-Validation, stepwise)						
Migration status	adj. R^2	Predictors				
		V1	V2	V3	V4	V5
Between-states	0.38	married	separated	college degrees		
In-state	0.03	naturalized citizenship				
	0.09	spouse absence	college degrees	naturalized citizenship	self-employment	
	0.09	no. of bedrooms	age	married	spouse absence	separated
Abroad	0.28	married	not in labor	rent		
	0.29	married	not in labor	rent	self-employment	
	0.29	college degrees				
Same House	0.21	naturalized citizenship				
	0.20	married	naturalized citizenship			

Table VII-2 Regression results for migration behavior of neighborhood type 2

The next group is people who moved into the nation from other countries. The three models are very close in their adjusted R^2 (0.28, 0.29, and 0.29). However, they differ significantly in predictor variables. Two models have three predictors overlapped: married, not in labor, and rent. The third regression model for in-state migration has only one predictor variable of educational attainment. The last group of regression models is for people who stay at the same house. These two models are similar in adjusted R^2 and their predictor variables. The magnitudes of the adjusted R^2 are around 0.2. The selected predictor variables are naturalized citizenship and married.

There is a challenge in the process of applying the cross-validation resampling approach. The analysis results vary each time. Sometimes, the variation is vast, which could be an obstacle to choosing the best model representing the whole data set's performance. It depends on the resampling process: how

data points are assigned into data subgroups. There are two requirements for the resampling results to be accurate. First, the training data set must be consistent with the testing data set in the model performance. Second, both training data and testing data are representative of the whole data set. Only when both needs are fulfilled would we be confident that the predictive model is reliable. However, this resampling process is random and in a "black box." Repeating the modeling process on different data subgroups can minimize the chance of missing any essential variable. The next step is to build models on the whole data set with selected variables.

The varying nature of the cross-validation resampling results is a strength of the approach at the same time. It provides valuable information in evaluating the "purity" degree of a data set. Take the regression analysis of neighborhood type 1 as an example. The two listed models for the between-states migration behavior have two completely different sets of predictor variables. The differences signify that the training data sets used to derive the two regression models are very different, which leads to low R^2 values. Therefore, for between-states migration, neighborhood type 1 has a low degree of purity. A similar impurity phenomenon happens in neighborhood type 2 for the population moved from abroad. One of the listed models has only one predictor variable of college degrees, which is absent in the other two regression models. A high percentage of college degrees is significant in some PUMAs of neighborhood type 2, but not for other areas.

7.3 Model fit

7.3.1 Migration status: moved between states

Table VII-3 summarizes regression model statistics for the whole study area, neighborhood type 1 and neighborhood type 2 for between-states migration. The first three predictor variables are naturalized citizenship, college degrees, and separated in modeling the whole study area. These three predictors account for more than 80% (0.31 out of 0.38) of the explained variance of the between-states migration. The regression model on neighborhood type 2 increased adjusted R^2 from 0.38 to

0.73. Correspondingly, the regression residuals have a noticeable drop compared to the global model. The residuals range between -8.6 and 26.7 in the global model and between -4.8 and 23.9 in the neighborhood type 2 model. Most of the model residuals in neighborhood type 2 fall between -2 and 2 instead of -4 and 4 in the global model (Figure VII-1). Since the data are normalized, model residuals have no unit, which allows for comparing different models.

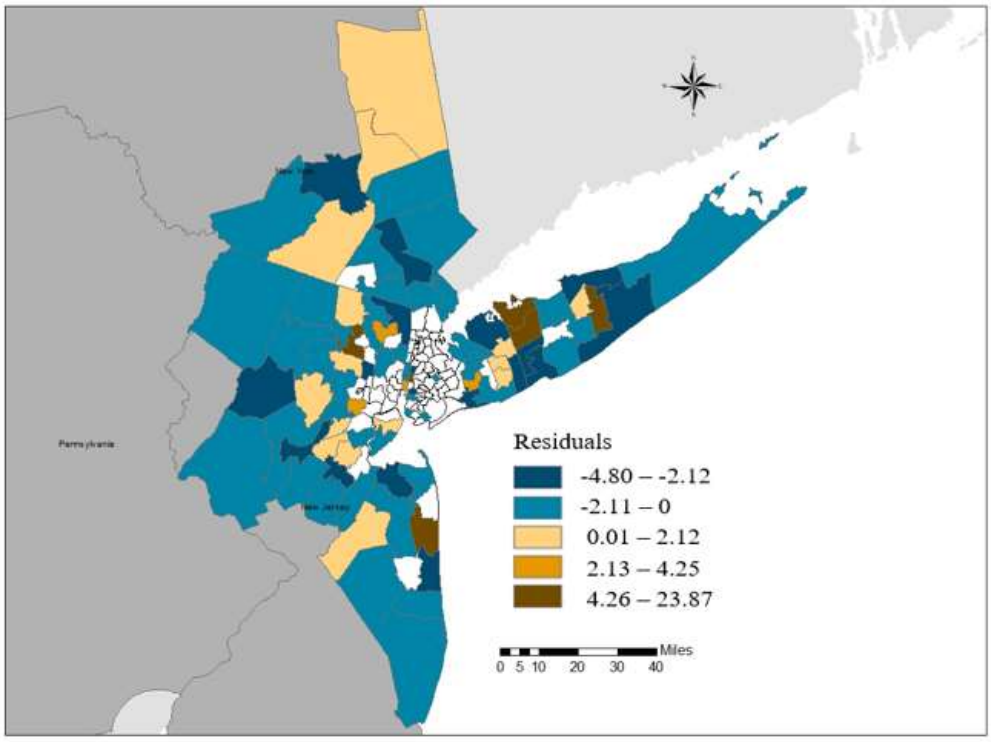
Migration Status: Between-States			
Neighborhood	Whole	1	2
adj. R^2	0.38	0.11	0.73
Pct citizenship	Yes	Yes	
Pct speak only English			Yes
Pct married			Yes
Pct single	Yes		
Pct separated	Yes		Yes
Pct college degrees	Yes	Yes	Yes
Pct self-Employed	Yes		Yes
Pct employed			Yes
Wage		Yes	

Table VII-3 Regression statistics for between-states migration

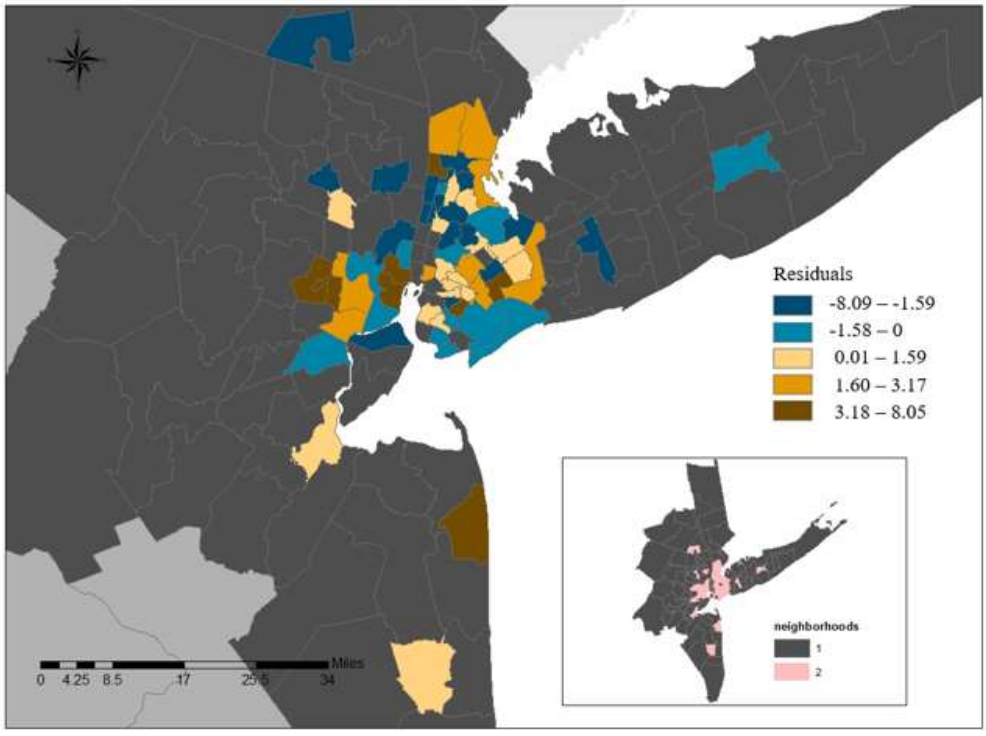
Predictor variables in the model of neighborhood type 2 are very similar to those for the entire study area, with two additional variables of speaking only English and employed. This set of variables depicts the critical characteristics of immigrants in the PUMAs of neighborhood type 2. They have a job, a high level of educational attainment, and not many obstacles for moving to a new place since most of them are single or separated. Most of them do not have naturalized citizenship, nor do they own a business. One push of a migration decision is a better job.

Compared to neighborhood type 2, neighborhood type 1 has a very low adjusted R^2 value of 0.11, which decreases the R^2 values for the entire study area. The model residuals did not improve compared to the global model (Figure VII-1). Therefore, we can see that the whole study area comprises data points from two different neighborhoods. The regression model performance is an

average of the models for the two neighborhoods separately. The average model is not representative of any neighborhood due to the big difference between the two. PUMAs of neighborhood type 1 create noises in the original data set. By removing noise from the data set, data become purer, and thus the relationships between predictor variables and the response variable get stronger and more evident. This explains why the model on neighborhood type 2 has a better performance than the whole study area.



VII-1a Neighborhood type 1



VII-1b Neighborhood type 2

Figure VII-1 Regression residuals for between-states migration

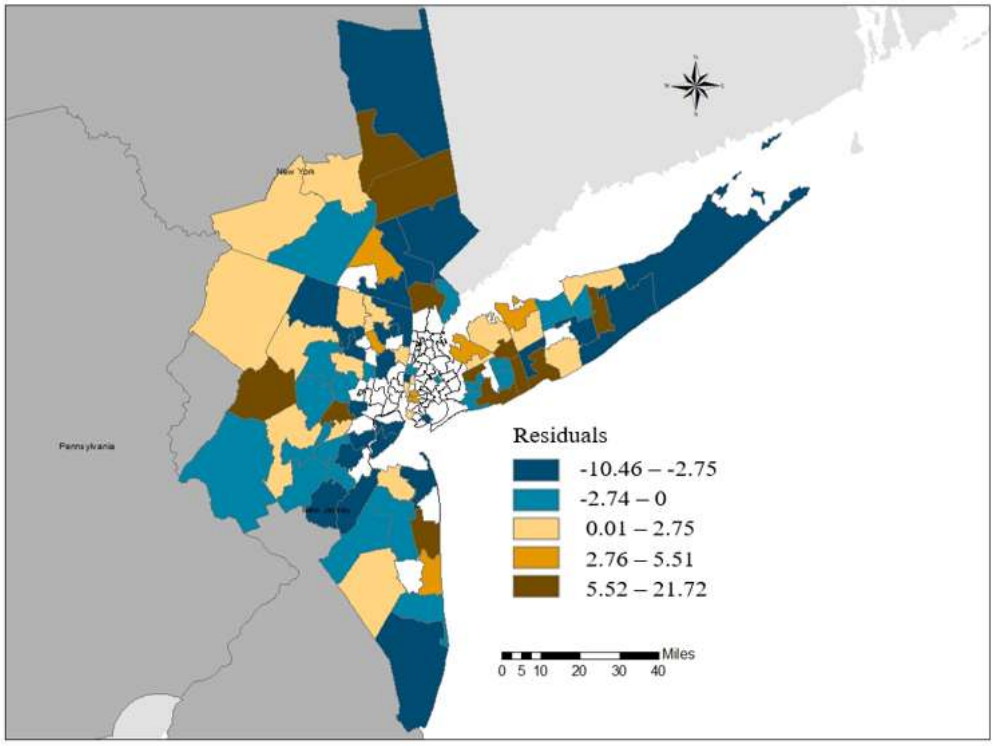
7.3.2 Migration status: moved within state

Built on the whole study area, the regression model in Chapter VI did not produce meaningful results since only 10% of the variance in the response variable in-state migration has been captured. The adjusted R^2 in the regression model for neighborhood type 2 is even lower (0.04). Therefore, it is meaningless to discuss the model fit for neighborhood type 2. The value of R^2 for neighborhood type 1, however, has a considerable increase (0.44). The maximal model residuals (both positive and negative) for neighborhood type 1 have decreased compared to the global model (Figure VII-2).

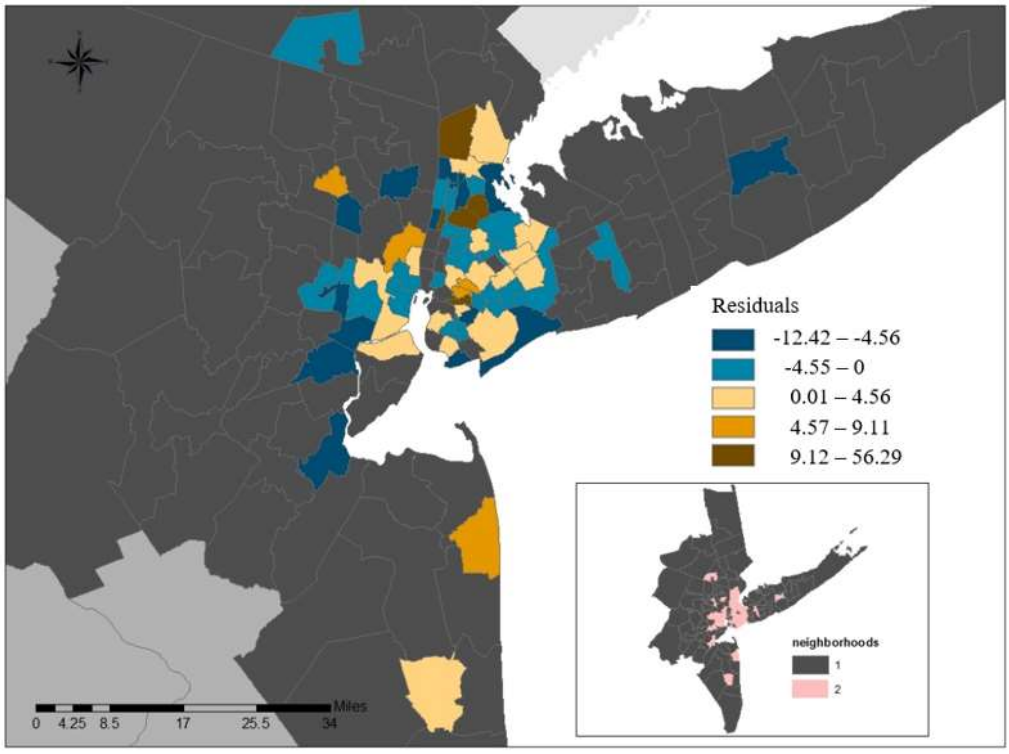
Migration Status: Moved Within State			
Neighborhood	whole	1	2
Adj. R2	0.10	0.44	0.04
Pct citizenship	Yes		Yes
Pct speak only English		Yes	
Pct born in Hong Kong		Yes	
Pct single		Yes	
Pct spouse Absent		Yes	
Pct self-Employed	Yes	Yes	
Number of bedrooms		Yes	

Table VII-4 Regression statistics for within-state migration

For neighborhood type 1, the first two most important predictors are spouse absence and self-employment. They both have a positive relationship with the response variable of in-state migration. The two independent variables together account for about 65% (0.29 out of 0.44) of the response variable's explained variance. The first predictor, spouse absence, is not a reason or push for a move. People whose spouse is absent or single do not have as many family-related obstacles to move as married people living with the spouse. Self-employment is a significant indicator for an in-state move. There are some common traits in this group of people: some of them own businesses, most do not live with their spouse or are single, English is not the only language they speak, and a modest portion of this group of the population was born in Hong Kong. Within this group of people, with the increase in the number of their bedrooms, the possibility of moving to a new place decreases.



VII-2a Neighborhood type 1



VII-2b Neighborhood type 2

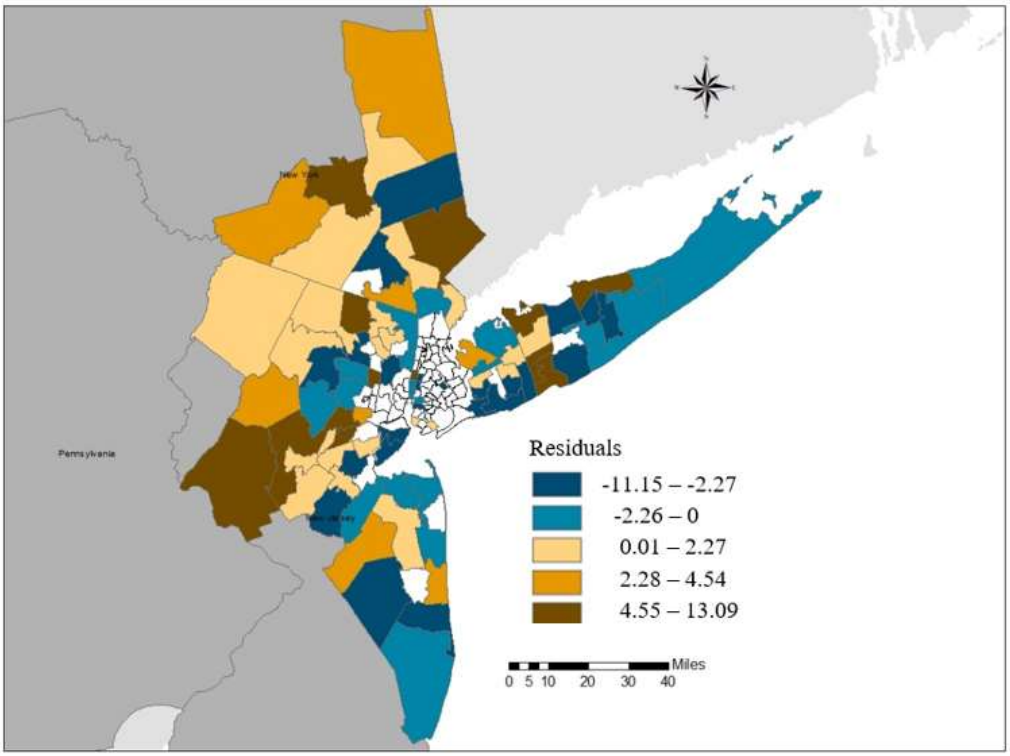
Figure VII-2 Regression residuals for within-state migration

7.3.3 Migration status: abroad

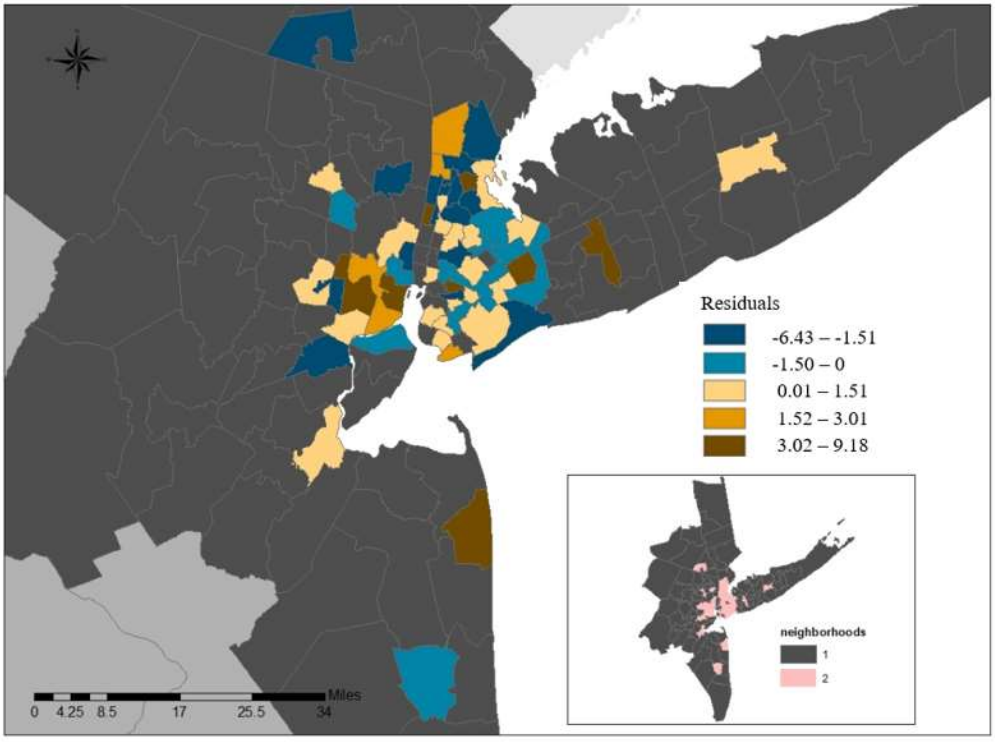
For migrants from abroad, the predictive strength of regression models on either neighborhood type 1 (0.53) or neighborhood type 2 (0.55) is slightly higher than the whole study area (0.46). The regression residuals suggest the same improvement (Figure VII-3). The regression models on neighborhood type 2 depict such a group of people who are self-employed. This group has a high ratio of college degrees and a status of married. Many live in big houses. This image differs from its counterpart in neighborhood type 1. Most people in neighborhood type 1 are single, not in the labor force, and able to speak English well. A decent wage is a primary moving reason. For both neighborhood types, age has a counter effect on a moving decision. The OLS model is closer to neighborhood type 1 in terms of the independent variables. One possibility is that the total population in neighborhood type 1 is substantially greater than neighborhood type 2. After mixing data points from the two neighborhoods, the migration pattern of neighborhood type 2 has been concealed.

Migration Status: Moved from Abroad			
Neighborhood	whole	1	2
Adj. R2	0.46	0.53	0.55
Pct naturalized citizenship		Yes	
Pct speak English well	Yes	Yes	
Pct speak English not well		Yes	
Age	Yes	Yes	Yes
Pct married			Yes
Pct single	Yes	Yes	
Pct of college degrees			Yes
Pct self-Employed			Yes
Pct not in labor force	Yes	Yes	Yes
Wage	Yes		
Rent			Yes

Table VII-5 Regression statistics for migration from abroad



VII-3a Neighborhood type 1



VII-3b Neighborhood type 2

Figure VII-3 Regression residuals for migration status: abroad one year ago

The model residuals for neighborhood type 1 did not offer visible improvements over the OLS model (Figure VI-6; Figure VII-3). The range of the model residuals for neighborhood type 2 has decreased compared to the OLS model. Some PUMAs' residuals dropped, while some PUMAs' residuals enlarged. Most PUMAs of neighborhood type 2 in the OLS model have negative residuals. Among them, some areas' residuals turn positive when just modeling neighborhood type 2. There are no visible patterns in the residual maps of neighborhood type 1 or neighborhood type 2.

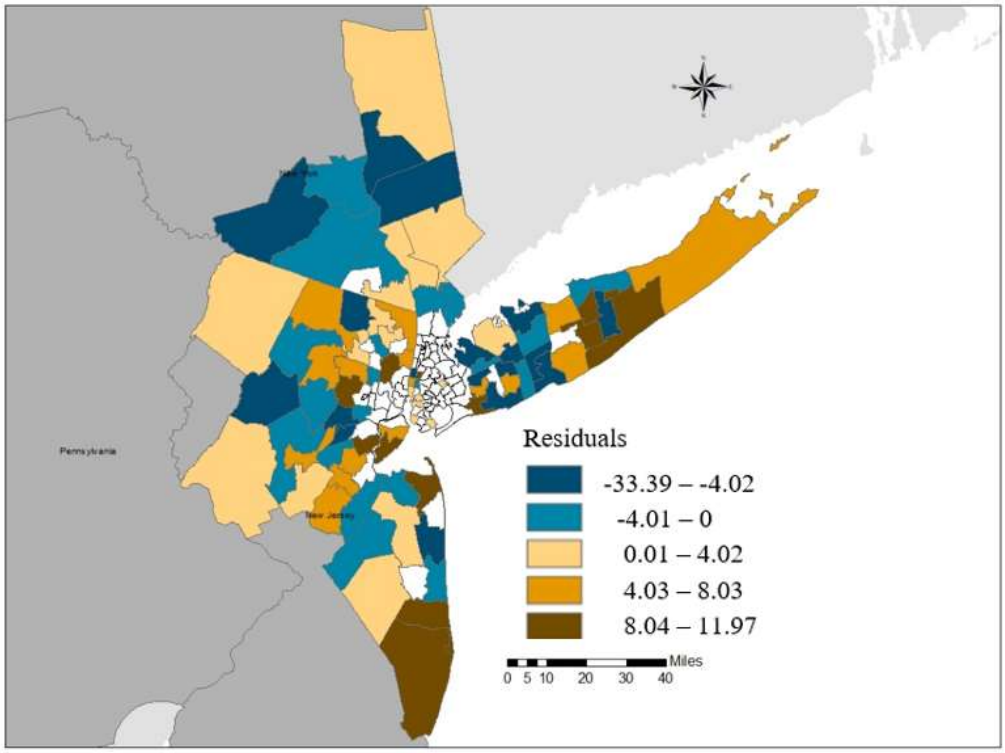
7.3.4 Migration status: same house

While it is complicated to untangle various migration patterns and unveil their traits, there are fewer variations in the characteristics of people who have stayed at the same house in the survey year. Neighborhood type 1 and type 2 are relatively consistent with the regression model built on the whole study area in terms of predictor variables in the models (see Table VII-6). One key difference is that neighborhood type 2 is associated with a high ratio of self-employed people. No matter which neighborhood they live in, people in this group have some common traits. Most of them have naturalized citizenship and do not have a college degree. Some speak only English. The average age of this group of people is higher than people with other migration patterns.

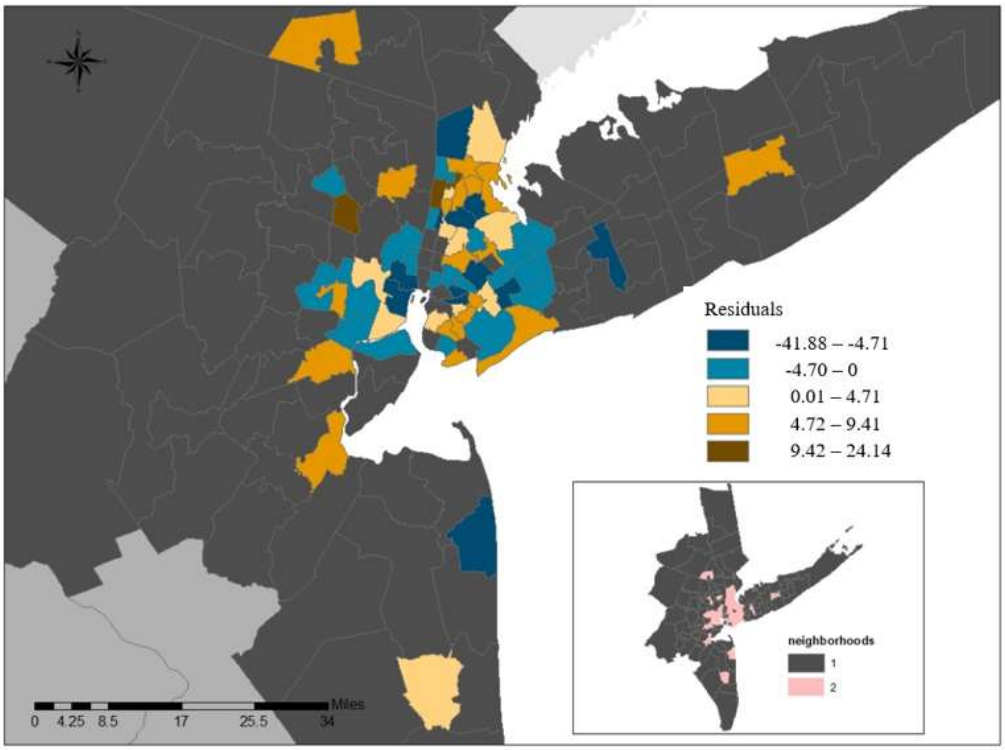
Migration Status: Same House			
Neighborhood	Whole	1	2
Adj. R2	0.37	0.45	0.38
Pct naturalized citizenship	Yes	Yes	Yes
Pct speak only English	Yes	Yes	
Age	Yes	Yes	
Pct married	Yes	Yes	Yes
Pct of college degrees	Yes	Yes	Yes
Pct self-Employed			Yes
Pct not in labor force		Yes	

Table VII-6 Regression statistics for staying at the same house

The predictive strength of regression models on either neighborhood type 1 (0.45) or neighborhood type 2 (0.38) is slightly higher than the whole study area (0.37). The maximum residuals in the model for neighborhood type 1 are smaller than the global model. Moreover, most model residuals of neighborhood type 1 fall in the range of -4 and 8 instead of -9.1 and 9.1 in the global model (Figure VI-8; Figure VII-4). The residual improvements in the model of neighborhood type 2 are not apparent.



VII-4a Neighborhood type 1



VII-4b Neighborhood type 2

Figure VII-4 Regression residuals for migration status: staying at the same house

7.4 Conclusions

Re-defining the physical boundaries of neighborhoods has improved the performance of predictive models. On average, the models built on a specific neighborhood perform better than on the whole study area. For the between-states migration, the model built on neighborhood type 2 strongly increases the adjusted R^2 than the entire study area. Similarly, a dramatic increase of model fit occurs in building in-state migration for neighborhood type 1.

Regression analysis on individual neighborhoods has revealed spatial patterns hidden from a global model. Neighborhood type 2 (low in socioeconomic status and Stable index) is the primary residential choice for immigrants from other states. They have a high ratio of college degrees. Jobs are a primary moving motivation. On the contrary, neighborhood type 1 (high in socioeconomic status and Stable index) has more within-state immigrants. They are high in the ratio of owning a business, and many of them were born in Hong Kong. Stratifications also exist in the Chinese who moved into the US from abroad.

In summary, regression models on individual neighborhoods have revealed spatial patterns of various migration behavior. The results further illustrate the relationship between people's migration behavior patterns and their demographic and socioeconomic status. From Chapter V, neighborhood classification, we know that spatial autocorrelations in areas exist from the global Moran's I index. However, such spatial correlations have not been considered in the modeling process of this chapter. In the next chapter, regression models taking into account spatial autocorrelations will be presented.

CHAPTER VIII

GWR AND MGWR

8.1 Introduction

As the third regression method in this study, Geographically weighted regression (GWR) and multiscale geographically weighted regression (MGWR) are applied to answer Research Question 3: How do local factors in the New York-Newark-Jersey City MSA impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially? This chapter focuses more on the second half of my third research question. GWR and MGWR are tools for exploring the spatial non-stationarity of migration behaviors and the underlying processes. While GWR offers a measure to examine the spatial scales of relations and assumes all relations in a model operate at the same spatial scale (Fotheringham et al. 1998). This is a critical limitation of GWR. MGWR improves GWR by allowing processes to work at different scales (Fotheringham et al. 2017; Oshan et al. 2019).

In this study, I include significant predictor variables selected from OLS models. One note here is that a significant predictor variable in an OLS model may not be as significant in an MGWR model because of the differences in their weighting schemes. Therefore, the predictor variables in

a GWR or MGWR regression model could differ from an OLS model. MGWR software does not have a built-in function to select the most valuable predictors. Although MGWR software cannot determine the most significant variables, it provides a tool for evaluating each predictor variable's influence level in a GWR or MGWR model, serving as a variable-filter method.

An adaptive bandwidth has been applied in GWR and MGWR models. In this study, relationships between the response variable and predictor variables generally vary at three spatial scales: a local scale of around 50 neighbors, a small regional scale of about 100 neighbors, and a broad regional scale of around 150 neighbors.

There are four migration categories: between-states migration, within-state migration, moving from abroad, and staying at the same house. The model fit is poor for building the regression model of within-state migration, with an R^2 value of around 0.1 for both GWR and MGWR models. The reason needs further investigation. With such little variation of the response variable being caught by the predictor variables, it isn't meaningful to discuss a variable's predictive strength. Thus, this chapter's subsequent sections focus on three out of the four migration categories: between-states migration, moving from abroad, and staying at the same house.

8.2 Migration status: moved between states

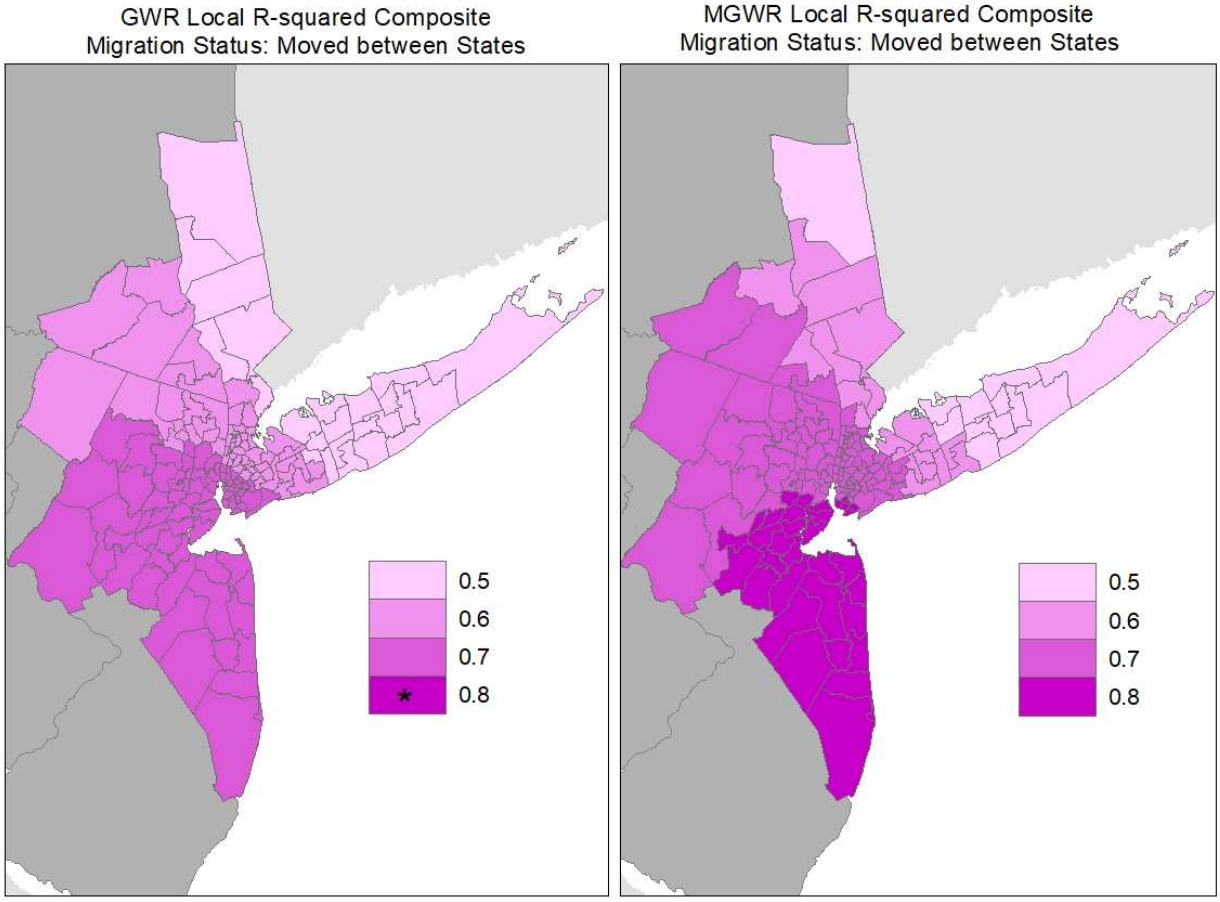
8.2.1 Model fit

Table VIII-1 summarizes model fit statistics across OLS, GWR, and MGWR models. GWR model and MGWR model have substantial improvements in adjusted R^2, AIC, and AICc. Figure VIII-1 is the distribution maps of the adjusted R^2 in the GWR and MGWR models, which apply the same R^2 legend to ease the comparison between them. The R^2 distribution patterns in the GWR model and MGWR model are similar. There is an increase in the R^2 values from the north towards the south on both maps. PUMAs in New Jersey state have a better model fit in the study area than those in New York state. The highest R^2 values on the MGWR map are above 0.8, which is a vast improvement

compared to an OLS model. The magnitudes of adjusted R^2 in the OLS model are around 0.3. Studies involving human activities are usually low in the R^2 due to the difficulties in coding social factors into an analysis. Regression results applying MGWR models provide another explanation of the poor data fit. Those studies may have ignored the spatial context of the study problem.

	OLS	GWR	MGWR
Adjusted R^2	0.376	0.598	0.633
AIC	365	309	301
AICc	368	316	314

Table VIII-1 Model fit statistics



Note: No records in the category marked by a ∗.

Figure VIII-1 GWR and MGWR local R2

8.2.2 Variable evaluation

The software MGWR offers hypothesis testing to evaluate each predictor variable's uncertainty in a model. At a 95% confidence level, the critical t-value of 1.96 (absolute values) is the threshold for

rejecting the null hypothesis that a parameter estimate is not significantly different from zero. The t-value of the model intercept is about zero, but it is expected due to data standardization. After standardization, all variables have a mean of zero and variance of the unit, facilitating the comparison of model contributions and bandwidths (in an MGWR model) across variables.

The thresholds in a GWR or an MGWR model are more conservative since their hypothesis testing is multiple-dependent. The GWR software produces an adjusted threshold of 2.293. The threshold in the MGWR model varies across each covariate relationship in the model (see Appendix A.1). They are all greater than the traditional threshold of 1.96. A greater value is more conservative to include a variable in the model. There are no changes in the statistically significant predictors in MGWR models based on the adjusted thresholds. Therefore, the following analysis will not include the three variables whose parameter estimates are not significantly different from zero: self-employment, married, and speaking only English.

Except for the three variables whose parameter estimates are not significantly different from zero, all other variables but naturalized citizenship positively correlate with the response variable. Based on the parameter estimates, the most influential predictor is single, followed by college degrees and separated. The last variable that has a positive relationship with the response variable is employment. The variable naturalized citizenship is the only variable that has a negative association with the response variable.

8.2.3 Parameter estimates and bandwidths

Table VIII-2 lists each parameter's bandwidth in the GWR and MGWR models of between-states migration. The optimal bandwidth derived in the GWR model is 130 nearest neighbors. The bandwidths in the MGWR model vary from one parameter to another, indicating that they operate at different spatial scales. Two predictor variables of college degrees and single impact the response variable at a broad regional scale of 148 neighbors within a total number of 150 PUMAs. The

variable of naturalized citizenships operates at a bandwidth of 136 nearest neighbors, and the variable of separated has a bandwidth of 134. The optimal bandwidth for the variable employment is 109 nearest neighbors.

Bandwidths for between-states migration models		
Predictor variable	Bandwidth	
	GWR	MGWR
% naturalized citizens	130	136
% college degree or above	130	148
% people separated from spouse	130	134
% wage workers	130	109
% people single	130	148

Table VIII-2 Bandwidths in the GWR and MGWR models for between-states migration

Figure VIII-2 is the composite map of parameter estimates where warm colors indicate a positive relationship and cold colors represent a negative relationship. The darker the color, the stronger a relationship is. The maps' interpretations are mainly on two aspects, with a particular focus on the MGWR maps. The first effort is to compare each covariate relationship between a GWR model and an MGWR model, indicated by the spatial patterns on the maps. Next, parameter estimates in a GWR model and an MGWR model will be compared to the global model.

The parameter estimate for naturalized citizenship in the OLS model is -0.241. It is the only statistically significant variable with a negative impact on the dependent variable. The GWR and MGWR models agree with the OLS model in the sign of the parameter estimate. The parameter estimates' magnitudes (absolute values) are smaller than the OLS model for most areas on both GWR and MGWR maps. Moreover, on the GWR and MGWR maps, naturalized citizenship has a more profound influence around Long Island.

The parameter estimate for college degrees is 0.315 in the OLS model. The estimates on the GWR map are not too far away from the OLS model, with three legend groups below the global estimate

and two legend groups cover or above the global estimate of 0.315. However, the parameter estimates on the MGWR model are smaller than the global estimate for all PUMAs in the study area.

On both GWR and MGWR maps, the parameter estimates for the variable separated vary smoothly from the north to the south, with larger values towards the south. While the relationship's sign agrees with the global model, the parameter estimates' magnitudes in most PUMAs are smaller than the global model (0.302) on the GWR map. All areas on the MGWR map have a parameter estimate smaller than the global model.

The employment variable exhibits a core and periphery spatial pattern on the MGWR map, with the biggest values around the center of the map and decreasing gradually to the periphery. The OLS model has a parameter estimate of 0.18 for employment. The variable has a more significant impact on both the GWR and the MGWR models than the global model, with many PUMAs having a bigger estimate than the global model.

The variable single is more influential to a between-state migration behavior in the MGWR model than the global model. Most areas on the MGWR map have a parameter estimate greater than the global estimate of 0.33. The MGWR map has a low spatial heterogeneity. Most PUMAs are dark brown, with the top north and Long Island areas in the NY state in a lighter brown color.

For each parameter estimate, the GWR map almost always has a higher spatial heterogeneity level than the MGWR map, while the visual pattern on the MGWR model is smoother. The GWR and MGWR maps display similar spatial patterns of naturalized citizenship. The parameter estimate's magnitudes increase from the south (NJ) to the north (NY). The main difference between the two maps is that the pattern is smoother on the MGWR map. One reason could be the overfitting issue brought about in a GWR model (Oshan et al. 2020).

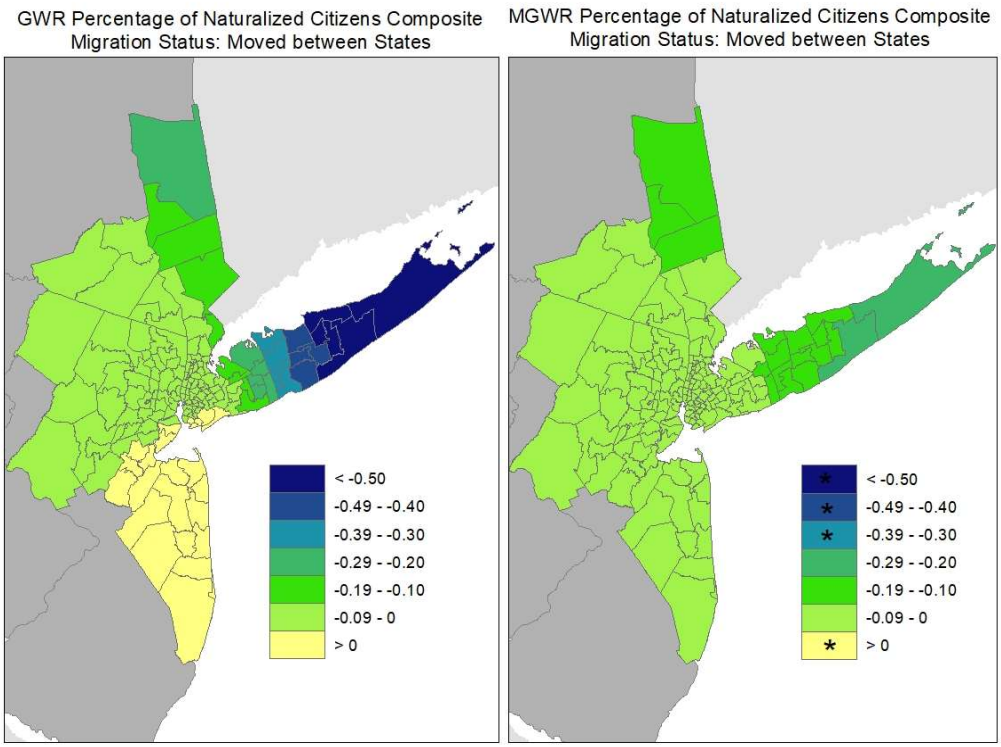


Figure VIII-2a Naturalized citizenship

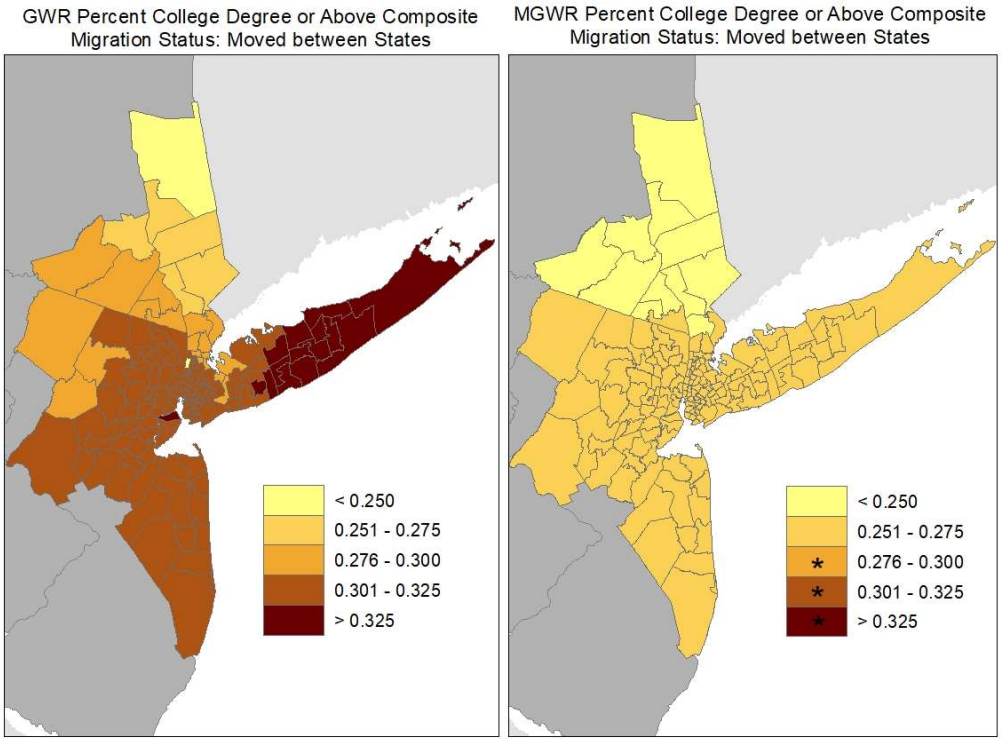


Figure VIII-2b College degrees

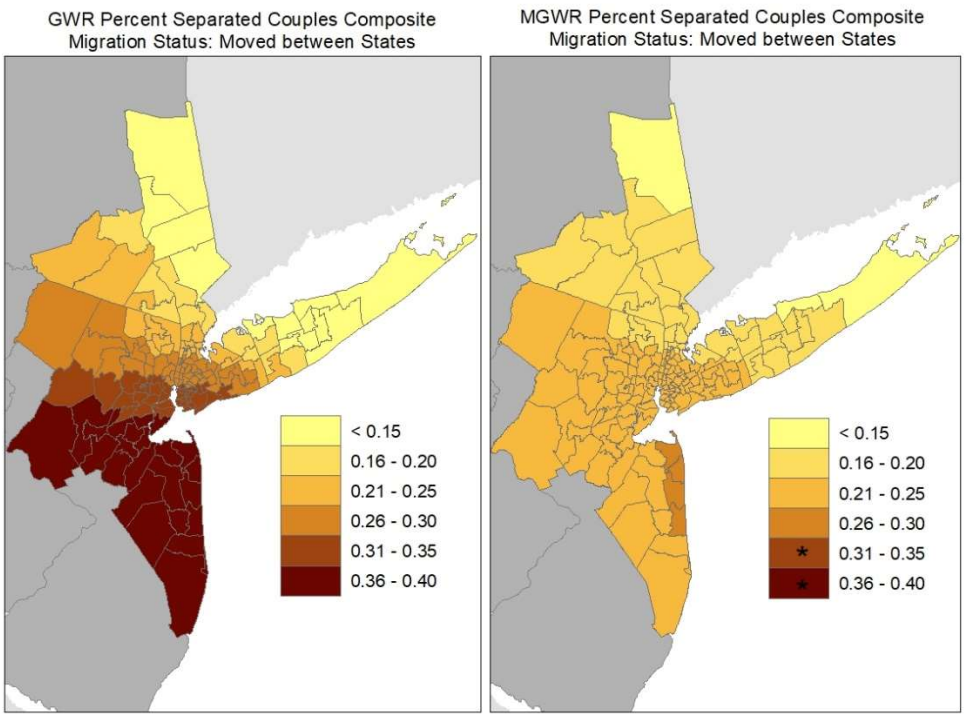


Figure VIII-2c Separated

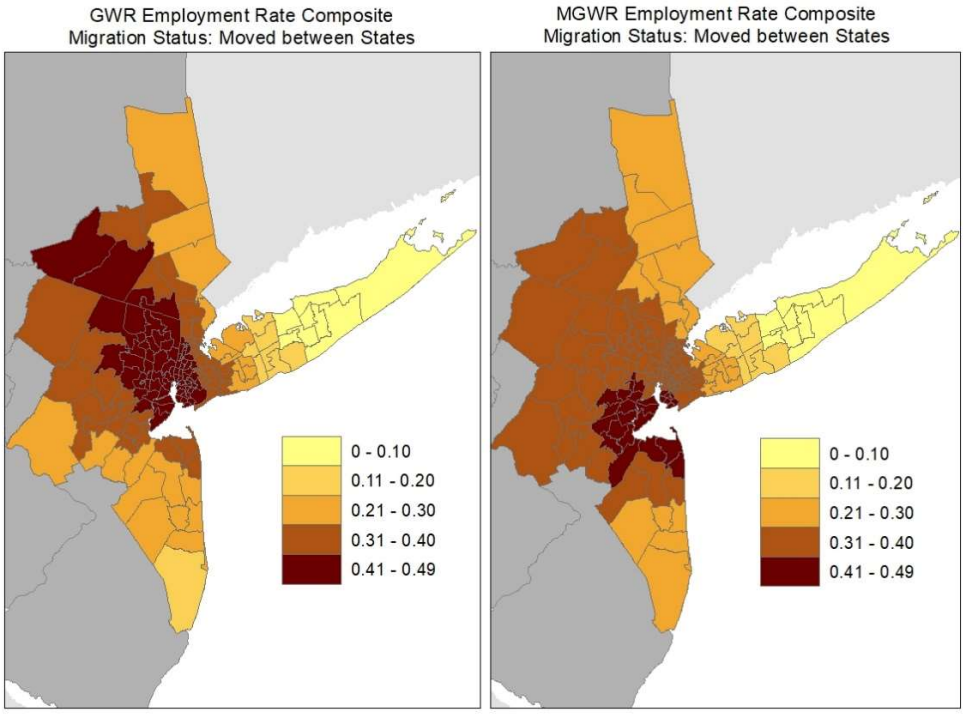


Figure VIII-2d Employment

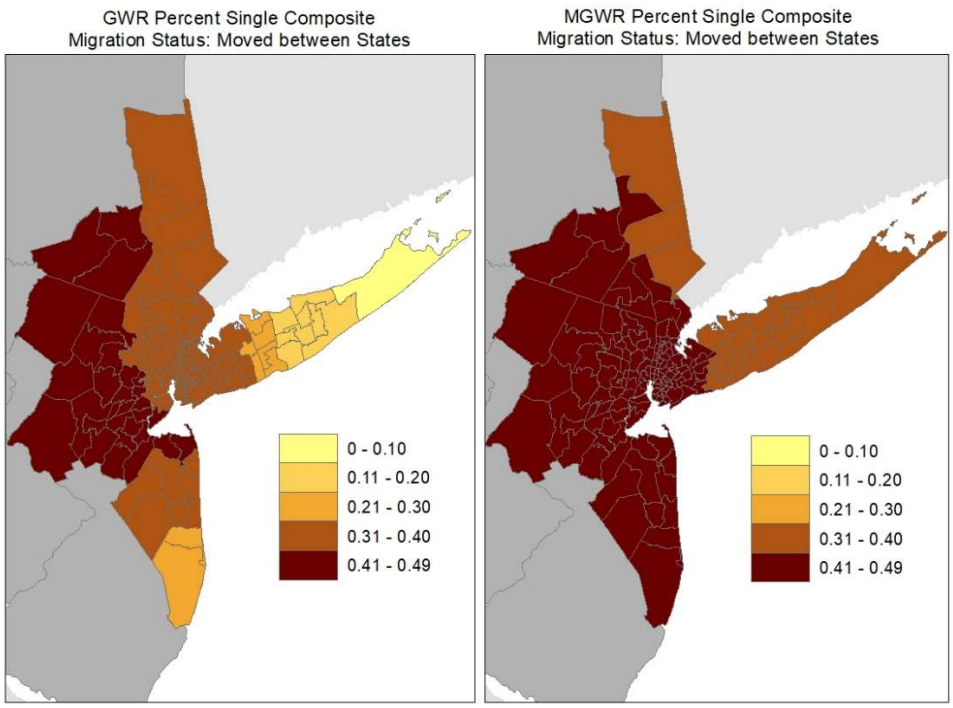


Figure VIII-2e Single

Note: No records in the category marked by a ∗.

Figure VIII-2 Composite maps of GWR and MGWR models

8.3 Migration status: abroad one year ago

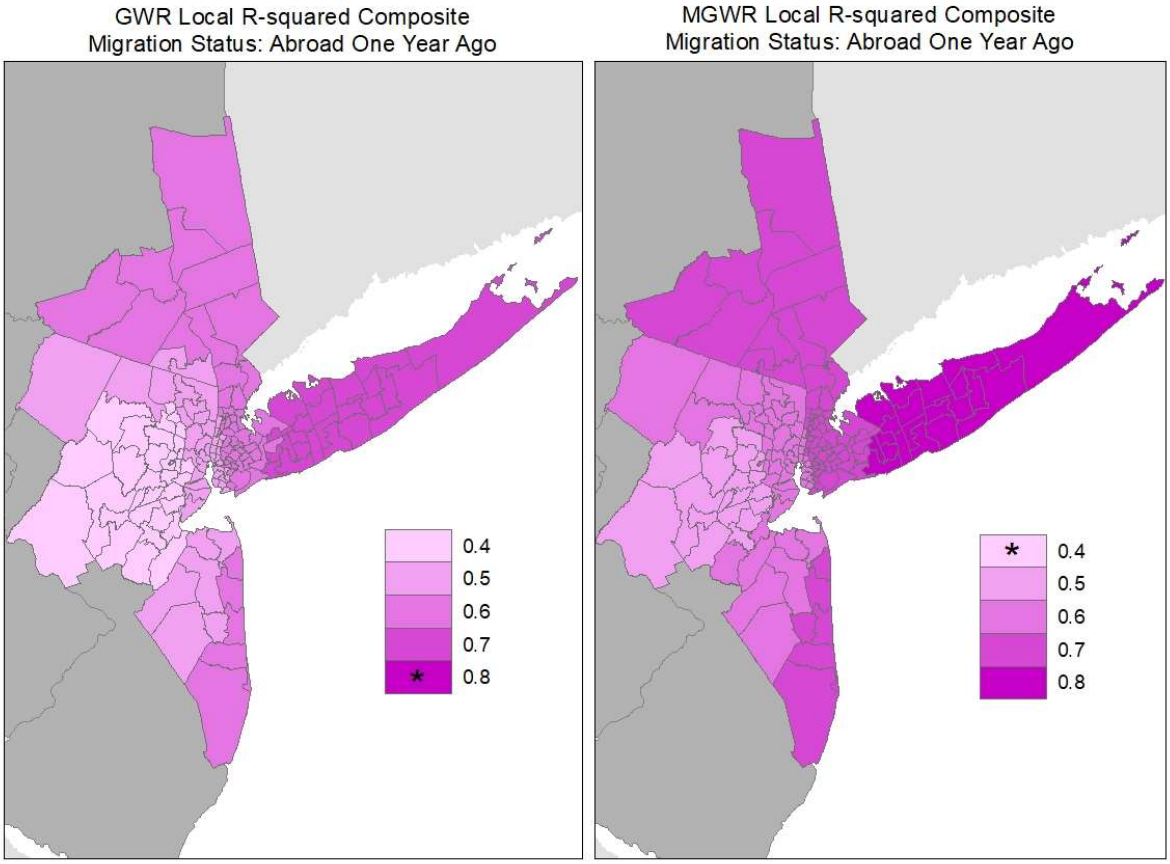
8.3.1 Model fit

An OLS model, a GWR model, and an MGWR model are used to study Chinese immigrants' migration behavior who moved into the US within one year of the data collection. Table VIII-3 summarizes model fit statistics in the above models. The AIC and AICc in the GWR and MGWR models are smaller than those in the OLS model, indicating a better model fit. Moreover, the GWR model and the MGWR model significantly increase the adjusted R^2 than the OLS model of 0.442. The R^2 values in the GWR and MGWR models are 0.601 and 0.690, respectively.

The distributions of local R^2 are shown in the maps of Figure VIII-3. The GWR map and the MGWR map display similar spatial patterns. Both models perform better in the periphery areas than in the center. The R^2 maximum and minimum values are 0.74 and 0.42 in the GWR model and 0.83 and 0.57 in the MGWR model.

	OLS	GWR	MGWR
Adjusted R2	0.442	0.601	0.690
AIC	351	312	284
AICc	355	323	314

Table VIII-3 Model fit statistics (migration status: abroad)



Note: No records in the category marked by a ∗.

Figure VIII-3 GWR and MGWR local R2 (migration status: abroad)

8.3.2 Variable evaluation

The OLS model includes eleven predictor variables derived from the OLS models. It is a combination of predictor variables for both neighborhood type 1 and neighborhood type 2. The two neighborhoods vary in their migration patterns and related factors, which results in a relatively large number of variable selections for models in this section. However, not all included variables are statistically significant based on the hypothesis testing on the parameter estimates (see Appendix A.2). Without splitting the study area into two neighborhood types, there are only three predictor variables whose parameter estimates are statically different from zero at a confidence level of 95%. These three

variables are speaking English well, not in the labor force, and age. The GWR and MGWR models produce the same hypothesis testing results of the parameter uncertainty as in the OLS model. The same three predictor variables' parameter estimates are statically different from zero at a confidence level of 95% (see Appendix A.3). The first two variables positively correlate with the response variable. The variable age has a negative relationship with the response variable. People newly moved into the US are generally young and not in the labor force. The capability of speaking English well is a crucial skill for them.

8.3.3 Parameter estimates and bandwidths

In analyzing the migration behavior of people moved into the US from a foreign country, the GWR model produces an optimal bandwidth of 134 nearest neighbors. It indicates an impact from predictor variables at a broad regional scale. The relationships vary at different scales in the MGWR model. Two variables, age and not in the labor force, impact the response variable at 93 and 64 neighbors, respectively. The predictor variable speaking English well affects the migration pattern at a scale of 144 nearest neighbors.

Bandwidths		
Predictor variable	Bandwidth	
	GWR	MGWR
% people speaking English well	134	144
Age	134	93
% people not in labor force	134	64

Table VIII-4 Bandwidths in the GWR and MGWR models (migration status: abroad)

Figure VIII-4 contains parameter estimates maps for the GWR and the MGWR models. For the variable speaking English well, both models show a core and periphery pattern. The values are the smallest around the center of the map and growing larger outwards. The pattern on the MGWR map is smoother than the GWR map. The GWR model and MGWR model agree with the global model in the direction of the relationship. The parameter estimate for speaking English well in the OLS model is 0.25. Most PUMAs on the MGWR map have a smaller parameter estimate.

The predictor variable age is negatively related to the response variable, and the magnitudes in the three models (OLS, GWR, and MGWR) are similar. Compared to the GWR map for the age variable, two key features emerge on the MGWR map. The impact from age is the strongest on the east end of Long Island (colored in yellow) and the weakest around the center (colored in navy) on the MGWR map. The parameter estimate in the global model is -0.287. On the MGWR map, the parameter estimates (absolute values) are greater in some PUMAs and smaller in others.

For the variable not in labor force, its parameter estimate surface has a medium to high spatial heterogeneity in both the GWR and the MGWR models. The two maps' spatial patterns are very similar, with parameter estimates increasing from the south to the north. The OLS model has a parameter estimate of 0.44, which is around the middle point of the estimate ranges on the GWR and the MGWR models.

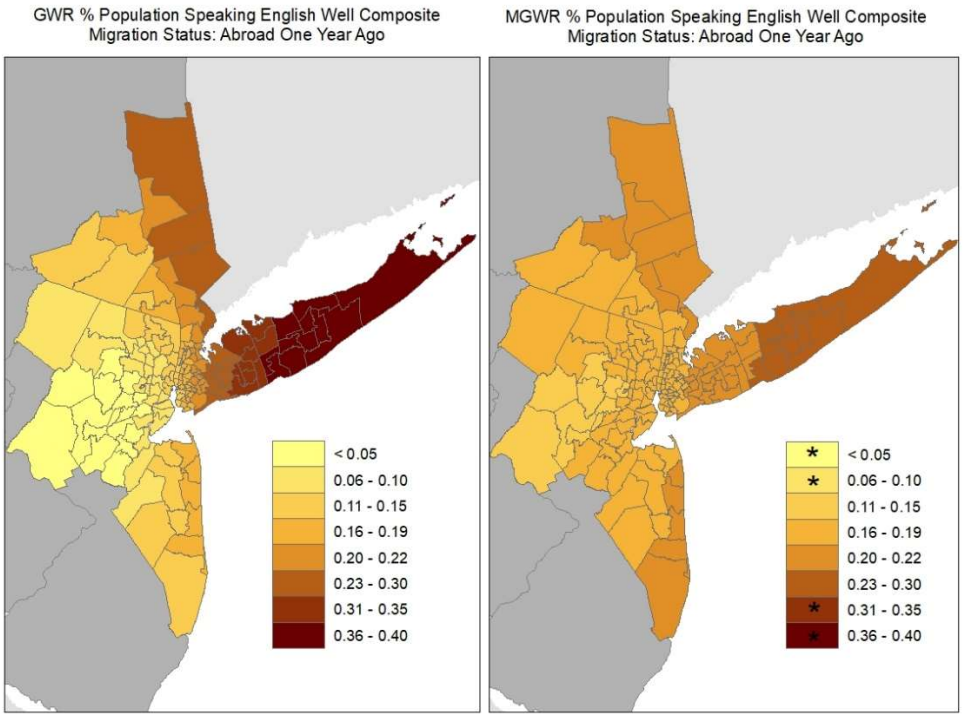


Figure VIII-4a Speaking English Well

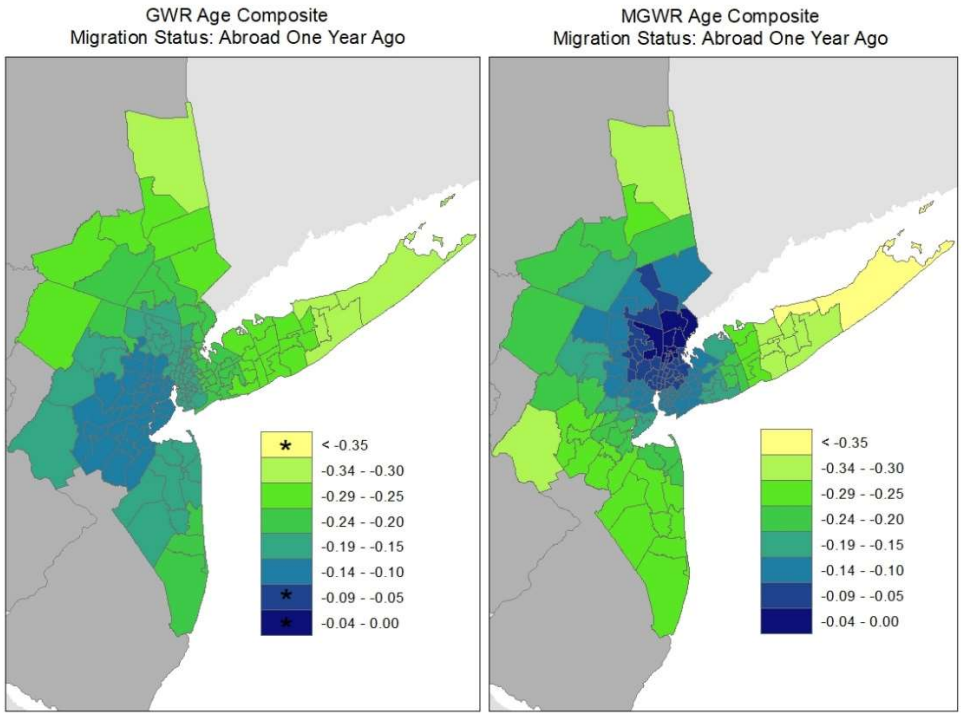


Figure VIII-4b Age

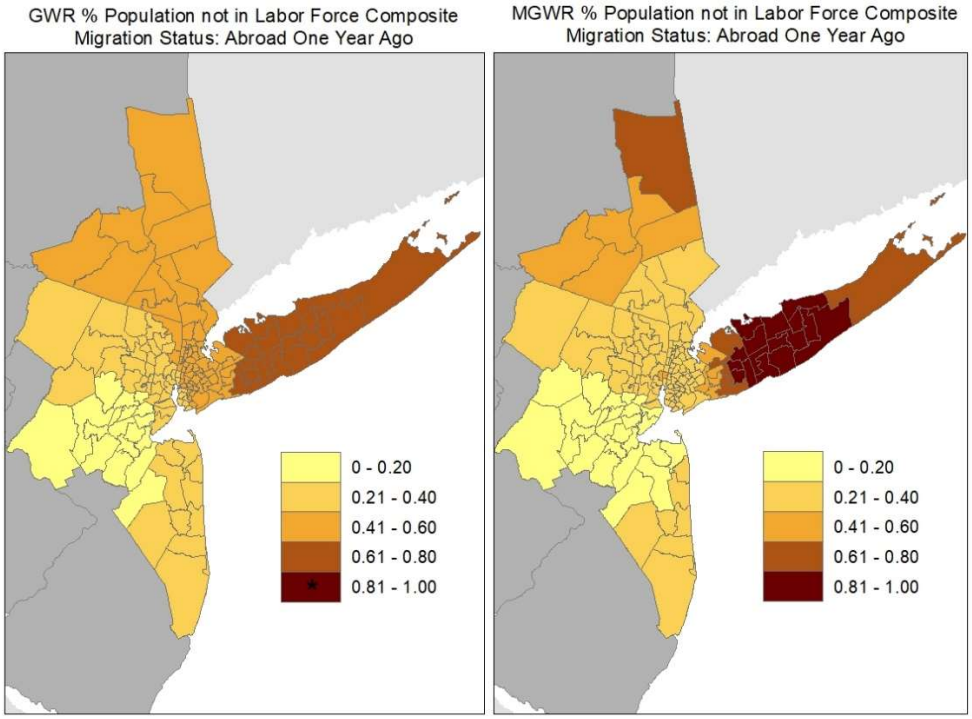


Figure VIII-4c Not in labor force

Note: No records in the category marked by a ∗.

Figure VIII-4 Composite maps of GWR and MGWR models (migration status: abroad)

8.4 Migration status: same house

8.4.1 Model fit

Table VIII-5 summarizes the OLS, GWR, and MGWR statistics for modeling people who stay at the same house. The two geographically weighted models are not superior to the OLS model. The GWR model and the MGWR model have increased the adjusted R^2 from 0.37 in the OLS model to 0.42. The small decrease of AIC and AICc in the geographically weighted models is ignorable. The distributions of local R^2 are in Figure VIII-5. Two maps display similar patterns, with R^2 values range between 0.42 and 0.53.

	OLS	GWR	MGWR
Adjusted R2	0.371	0.421	0.419
AIC	365	360	361
AICc	368	363	365

Table VIII-5 Model fit statistics (migration status: staying at the same house)

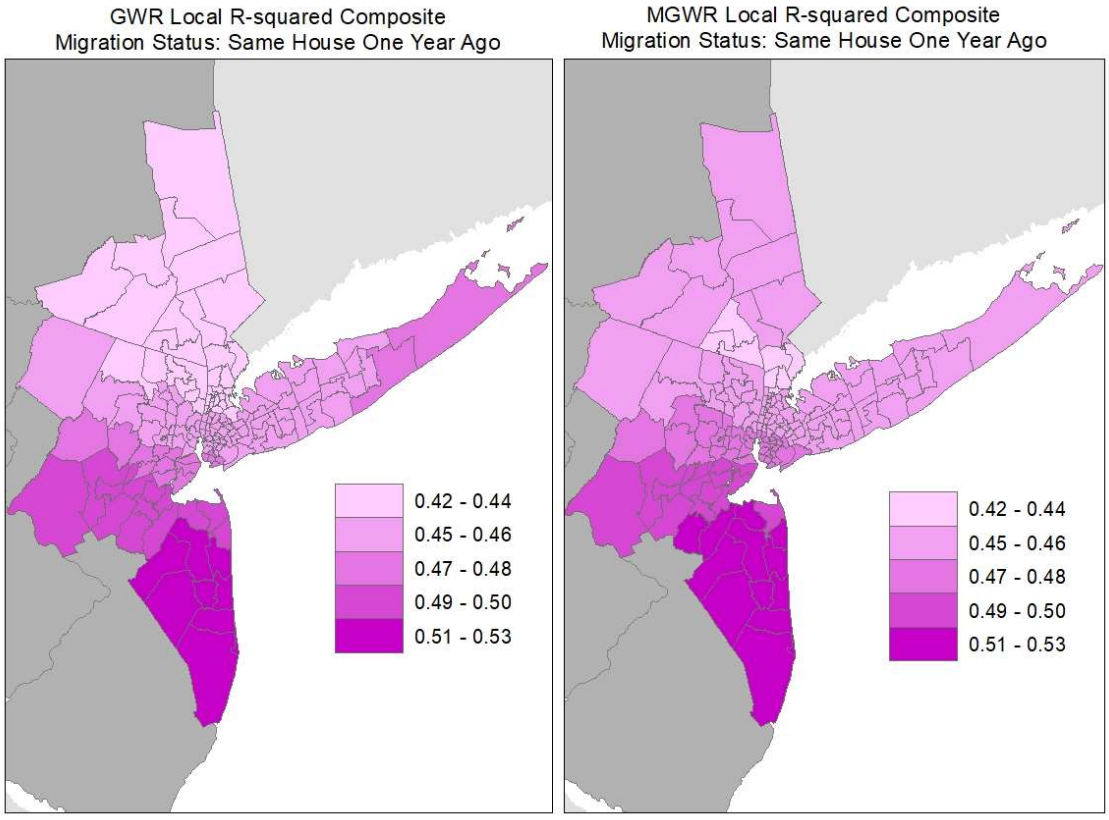


Figure VIII-5 GWR and MGWR local R2 (migration status: staying at the same house)

8.4.2 Variable evaluation

Four variables are statistically significant in the OLS model. The uncertainty test results for the parameter estimates in the MGWR models are consistent with the OLS model (see Appendix A.3). The most influential predictor is marital status (being married specifically), followed by naturalized citizenship, age, and college attainment. Three out of the four variables are positively related to the response variable. The first three variables positively correlate with the response variable, while the variable college degrees is negatively related to the response variable. The results suggest that married people, especially those with naturalized citizenship, are less likely to move to a new place than other Chinese immigrants in the study area. At the same time, being married and with naturalized citizenship indicate a relatively older age. PUMAs with a high ratio of people staying at the same house are associated with a low percentage of college degrees.

8.4.3 Parameter estimate and bandwidths

As indicated in Table VIII-6, the bandwidth differences of parameter estimates in the GWR model and the MGWR model are minor. The final MGWR models include four variables: naturalized citizenship, age, married, and college degrees. In the GWR model, the optimal bandwidth is 141, while the bandwidths in the MGWR model range from 142 to 148 for the four variables left in the model.

Regression model for % Stay at the same house		
Predictor variable	Bandwidth	
	GWR	MGWR
% naturalized citizens	141	148
Age	141	142
% people married	141	142
% college degree or above	141	148

Table VIII-6 Bandwidths in the GWR and MGWR models (migration status: same house)

In Figure VIII-6, the GWR composite maps demonstrate higher spatial heterogeneity than the corresponding MGWR maps. The spatial pattern in the MGWR map is smoother, with less variety in

each parameter estimate. Values of the variable married increase from the north to the south of the study area on the MGWR map. Most areas' parameter estimates fall into the category of 0.31 and 0.40 (colored in light brown), close to the estimate of 0.310 in the OLS model. The parameter estimates of naturalized citizenship and age increase from the south to the north on the MGWR maps. For the variable of naturalized citizenship, the OLS model has a parameter estimate of 0.275, which is consistent with the first two legend categories on the MGWR map. The global parameter estimate for the variable age is 0.186 in the OLS model. Only a few PUMAs indicated in one legend category on the MGWR map have parameter estimates around 0.186. All other areas have greater parameter estimates than the OLS model. Other than a north-south spatial pattern as in the above three variables, the variable college degrees has a west-east pattern. Its local parameter estimates decrease from west to the east on the MGWR map. The parameter estimates on the MGWR map vary between -0.2 to -0.279, while the global parameter estimate in the OLS model is -0.24, which is around the middle point of the local estimates.

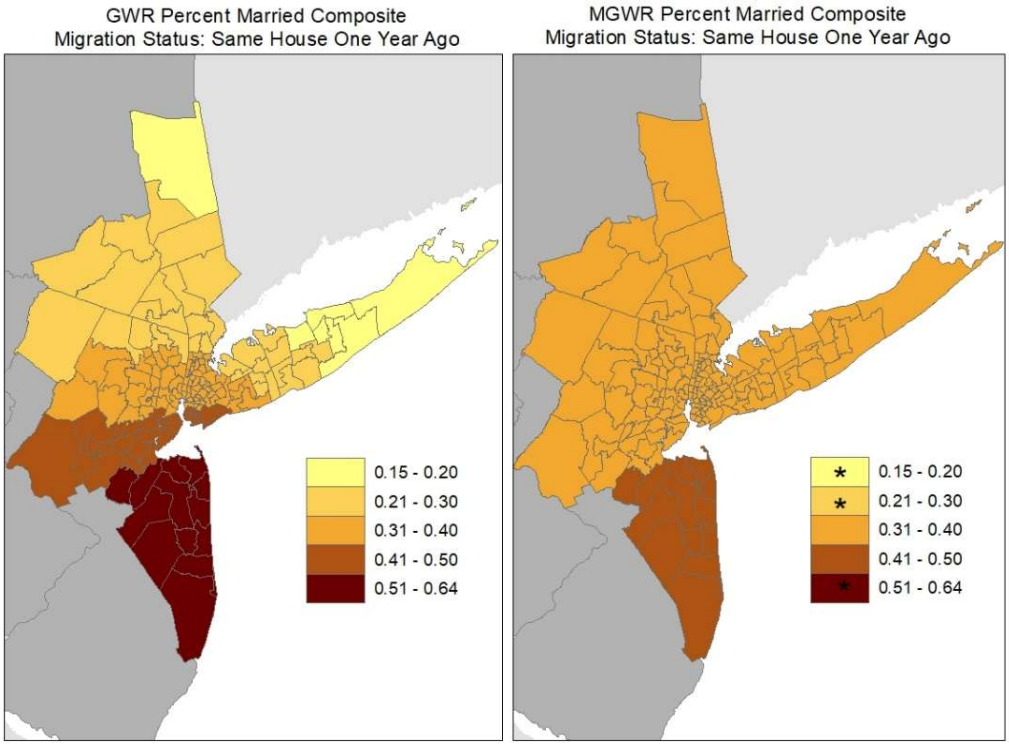


Figure VIII-6a Married

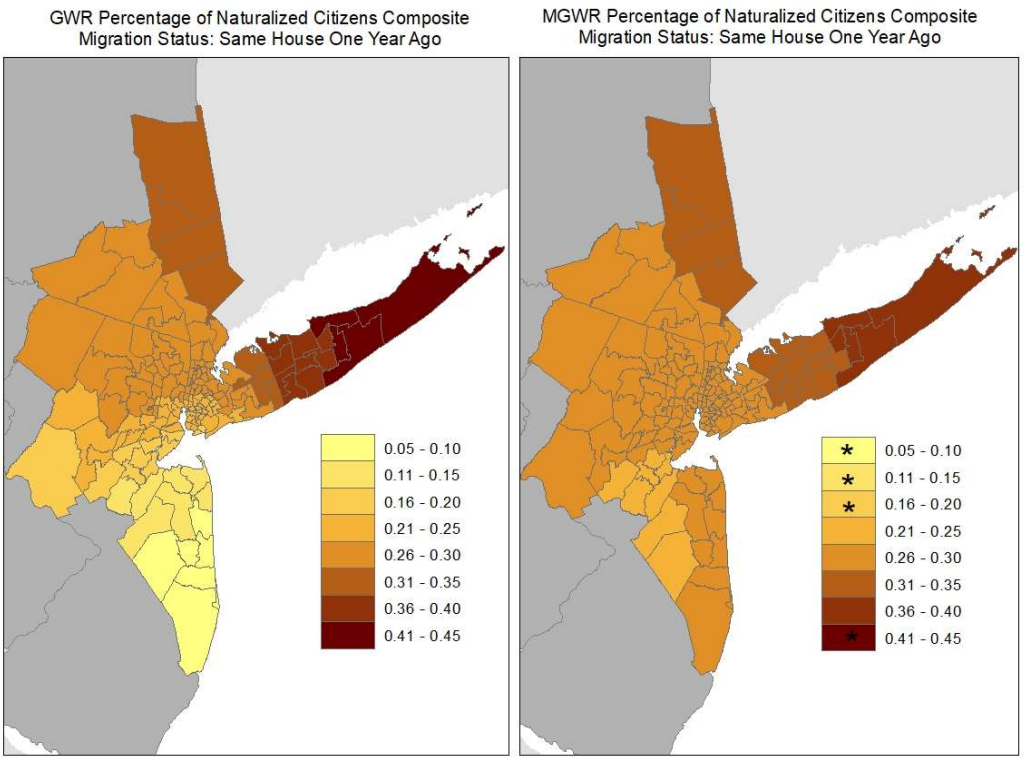


Figure VIII-6b Naturalized citizenship

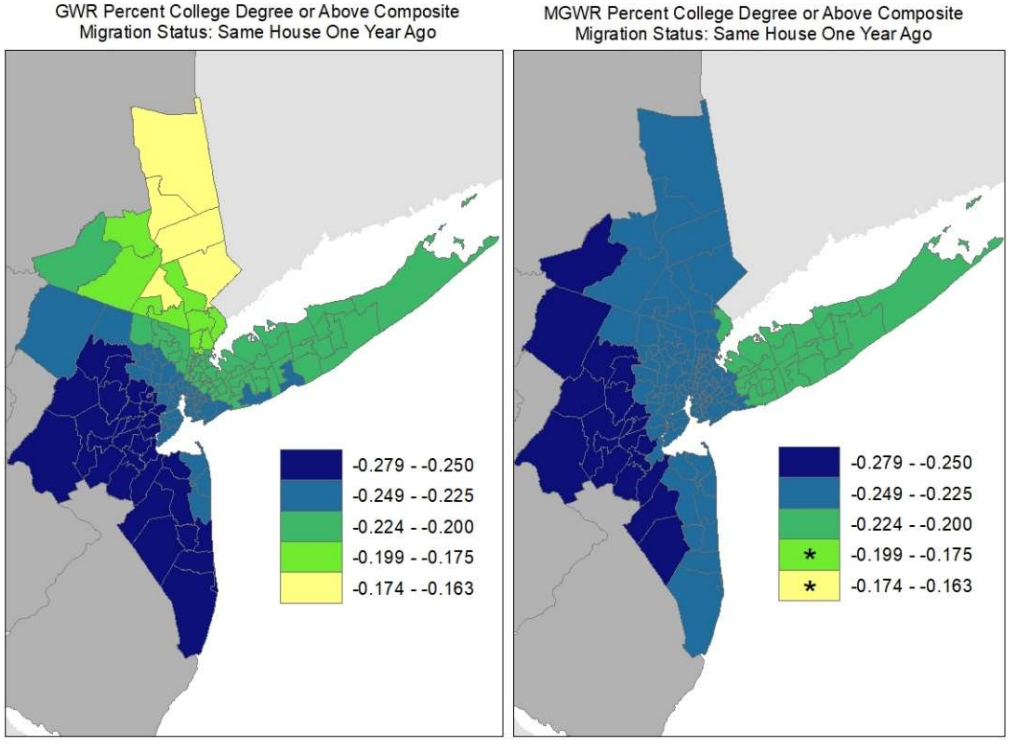


Figure VIII-6c College degrees

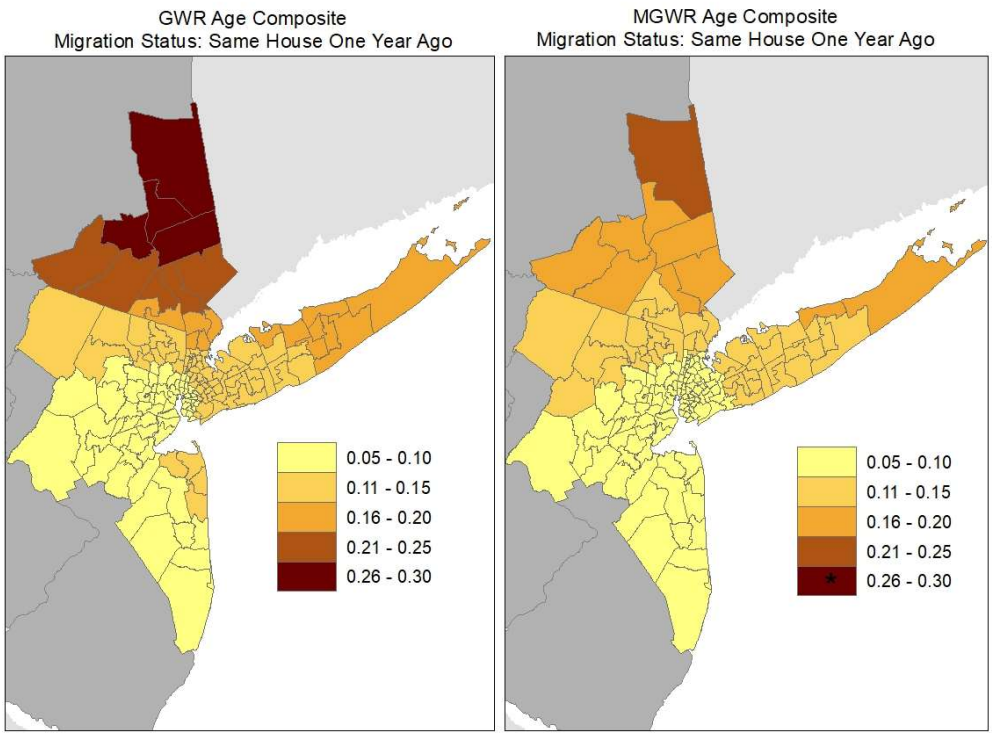


Figure VIII-6d Age

Note: No records in the category marked by a *.

Figure VIII-6 GWR and MGWR Composite maps (migration status: same house)

8.5 Conclusions

GWR and MGWR models have stronger predictive strength than the global model. For example, the GWR and MGWR models for between-states migration increase the adjusted R^2 from 0.38 in the OLS model to 0.60 and 0.63, respectively. The GWR and MGWR models for migration from abroad increased the adjusted R^2 from 0.44 in the OLS model to 0.60 and 0.69. However, the geographically weighted regression did not improve the model for within-state migration when the model fit is poor in the OLS regression. For diverse migration categories, MGWR models have a better model fit than GWR models.

GWR and MGWR models produce local model fit statistics (R^2 and residuals) and parameter estimates. The two geographically weighted models display similar spatial patterns of the relationships between a predictor and the dependent variable, either in a core-and-periphery or north-

to-south pattern. However, the MGWR models' distribution patterns are generally smoother. The MGWR maps help locate areas where the MGWR models perform the best. For example, the MGWR model for between-states migration performs better in New Jersey than New York, with the highest R^2 over 0.8 in New Jersey. For migration from abroad, the maximum R^2 in the MGW model is 0.83 along Long Island.

Though the MGWR software does not have a function to select the most significant predictors, it can test whether or not an independent variable is statistically significant. For between-states migration, there are five variables whose parameter estimates are tested to be significantly different from zero. These variables are single, college degrees, separated, employment rate, and naturalized citizenship. Among the independent variables, the employment rate has the smallest bandwidth of 109 nearest neighbors. The bandwidths of the other four variables range between 134 and 148, very close to a global model of all 150 PUMAs. Three variables are statistically significant for migration from abroad: speaking English well, not in the labor force, and age. The variable speaking English well has a bandwidth of 144, which is almost global. The bandwidths for not in the labor force is 64 neighbors, which indicating a local effect. The variable age has a bandwidth of 93 neighbors, signifying a regional impact.

In summary, accounting for the spatial associations of relationships has dramatically increased the model fit. However, this improvement would not happen when the model fit is too poor, such as in the case of within-state migration. Moreover, GWR and MGWR can pinpoint the locality properties in a migration study. GWR and MGWR models also identify the scales of underlying processes in a migration study, which helps understand the spatial variation of migration behavior.

CHAPTER IX

CONCLUSIONS

9.1 Introduction

In studying the migration of Chinese immigrants, I examine underlying processes and their spatial variations. In the research process, I have examined three research questions.

Question 1. What contextual factors may affect a Chinese immigrant's migration behavior in the New York-Newark-Jersey City MSA?

Question 2. What are the distribution patterns and characteristics of each neighborhood type in the New York-Newark-Jersey City MSA?

Question 3. How do local factors in the New York-Newark-Jersey City MSA impact the migration behavior of the Chinese population? Particularly, how do relationships vary spatially?

First, this chapter recaps how the analysis results in this study explore each research question. Some analysis results are discussed together since contextual factors cannot be addressed independently of their relationship to migration behavior. Subsequently, I discuss the practical

implications of this work. The last section explains the limitations of this project and suggestions for future research.

9.2 Results

9.2.1 Contextual factors

In examining Research Question 1, decision tree analysis has been applied to reveal significant indicators at the individual scale. Wages, citizenship, and self-employment status are the top three crucial variables in differentiating movers from non-movers. Housing situations are only a concern for self-employed people. For immigrants from mainland China, having a college degree is an extra critical indicator.

Among several attempts to apply decision trees on the Chinese microdata, two attempts succeeded. The first attempt was to differentiate movers from non-movers. Employed people and naturalized citizens have a higher moving possibility (either within the state or out of the state) compared to their counterparts. Since the microdata is un-balanced with substantially more movers within the state than out of the state, the within-state migration trend is more evident than between-states migration in the decision tree analysis results. Self-employment is a unique factor in the ethnic migration of the Chinese population. Compared to self-employed people, individuals working for wages are more likely to make a between-states migration.

In the decision tree, there are also housing statistics at a lower rank. These variables include the number of bedrooms and gross monthly rental cost. These indicators belong to the "branch" of self-employed people. Based on rent and number of bedrooms, self-employed people are assigned into four subgroups. The first subgroup includes people whose rental cost per month is between $667 and $1,669. They are more likely to make a within-state move. The other three subgroups are mainly new immigrants from abroad. The second and third subgroup's rental costs are less than $667 and greater than $1,669, respectively. Many people in the fourth subgroup live in a large house, which is

indicated by the number of bedrooms in the housing unit. People who identified themselves as self-employed are more diverse than people who work for wages.

The second successful attempt in applying the decision tree analysis was to capture key migration factors associated with people from mainland China. The first two significant indicators are employment status and citizenship. People from mainland China tend to move within the same state when employed and have naturalized citizenship. The third splitting variable is college degrees. Among the new Chinese immigrants in the survey year, two subgroups appear. One group of people have a college degree. The other group of people does not have a college degree, and more than 80% identify themselves as self-employed.

9.2.2 Neighborhood classification

Research Question 2 relates to neighborhood classification. After experimenting with assigning census tracts into different neighborhoods, six clusters (neighborhood types) best depict the distribution of the Chinese population in relation to a place's local characteristics. However, assigning PUMAs into two neighborhood types is better due to the data size issue after transferring the neighborhood classification into the PUMA level.

To identify each neighborhood's characters, I selected variables from demographics, socioeconomic conditions, and living environment. After studying the structure of these variables, three variable clusters appear, with each representing a specific aspect of a census tract: SES (socioeconomic status), Stability index, and Chinese percentages.

A high SES is an area high in college degrees, occupations in MBSA, median earnings, rental costs, and house value. These areas usually are low in the unemployment rate and the number of jobs in service. One note here is that data are normalized with a mean of zero and a standard deviation of one. The Stability index describes percentages of people who are married, naturalized citizens, or white. An area high in these statistics is more stable than an area with a high percentage of renters and

people under poverty. A high SES does not necessarily mean a high Stability index and vice visa. However, the two indexes are consistent in their signs. An area with a positive SES always has a positive Stability index. A negative SES always corresponds to a negative Stability index. The differences in the Chinese percentages in the census tracts are significant. The combination of SES, Stability index and Chinese percentages split areas into different neighborhood types.

When splitting areas into six neighborhood types, three of them have a Chinese percentage greater than the mean, and the other three are below the mean. The focus is on the first three neighborhood types. The neighborhood type with the highest Chinese percentage of 48.48% is low in both SES (-1.51) and Stability index (-1.37). Both indexes are the second to the lowest among all six clusters. The second highest Chinese percentage is 27.97%. This cluster corresponds to a close-to-average SES (0.26) and Stability index (0.08). There is a third cluster whose Chinese percentage is greater than the mean, which is 4.12%. This cluster has the highest SES and the second-highest Stability index value.

PUMAs are assigned into two neighborhood types to make sure each neighborhood type has enough data for analysis. Being stable and holding high SES characterizes neighborhood type 1. On the contrary, neighborhood type 2 areas have low Stability and low SES. Neighborhood type 2 has a greater Chinese concentration than type 1, with the former (0.13) slightly above the study area's mean percentage and the latter (-0.13) marginally lower than the mean. The two neighborhood types generally display a core-and periphery pattern, with neighborhood type 2 around the center (except Manhattan) and neighborhood type 1 disperses in the periphery areas. The neighborhood clusters are one embodiment of the contextual factors' spatial variations.

9.2.3 Regression results

To address Research Question 3, I built regression models (OLS, regression on neighborhoods, and GWR and MGWR) on the migration behavior of the immigrant Chinese at the PUMA scale. Based on

migration patterns, there are four independent variables: within-state migration, between-states migration, migration from abroad, and stay at the same house.

For the within-state migration, the OLS model fit is poor, with an R^2 of 0.1. The poor model fit may be a result of the "impureness" in the within-state data. From the decision tree analysis, we can see that within-state movers vary in their employment status, naturalized citizenship, college degrees, class of workers, rental costs, and the number of bedrooms in their housing units. With such wide variations, it is hard to catch the migration pattern of any subgroup.

After splitting PUMAs into two neighborhood types, the model R^2 increased to 0.44 in neighborhood type 1 and decreased to 0.04 in neighborhood type 2 for the within-state migration. One group of people has been identified in neighborhood type 1. In neighborhood type 1, migration within the same state positively relates to people born in Hong Kong, self-employed, single, and spouse absent. The within-state migration is negatively associated with the percentage of people who speak only English and the number of bedrooms in their housing units.

While the model for neighborhood type 1 has more predictive strength for the within-state migration, neighborhood type 2 dramatically improves the model fit for between-states migration. The R^2 in the model increases from 0.38 in the OLS model to 0.73 for neighborhood type 2. The results agree with the decision tree analysis. The variables employed, college degrees, and separated play a positive role in the between-states migration. Other variables including married, speaking only English, and self-employed are negative factors in between-states migration. It is clear that neighborhood type 2 is more of a residence choice for movers who migrate between states, and neighborhood type 1 is dominated by movers who migrate within the same state.

Compared to the above two migration categories, the neighborhood clustering approach only slightly improves model fit for migration behavior of people who moved from abroad or those who did not change their residence in the survey year. There are no apparent differences in the migration pattern

or mover characters between the two neighborhoods. Across the global model, models for individual neighborhood types, there are some common traits. For new immigrants from abroad, the variables college degrees, self-employed, single, and speaking English well have a positive role in increasing the likelihood of migration. On the contrary, married, age, and naturalized citizenship decrease the possibility.

For people who did not change their residence, regression analysis results reinforce the differences between movers and non-movers. Age, married, naturalized citizenship, speaking only English have a positive influence for staying at the same house. Conversely, college degrees and self-employed are two main negative factors for not changing residence.

MGWR software does not have a function to select the most significant variables. However, the software can evaluate whether a variable's estimate parameter is statistically different from zero. After performing the variable-filter function, MGWR supports the significance of some factors. The variables employment rate, college degrees, naturalized citizenship positively affect a PUMA's between-states migration rate. Being single and separated decrease the obstacles to making a move. MGWR software performs poorly on modeling the within-state migration. In modeling the migration behavior of those who previously lived abroad, speaking English well is a positive factor, while age is a negative factor. Not all factors selected in the OLS model have passed the significance test. The reason may be that there are various migration motivations among the new immigrants. The data set is a mixture of various groups of people, and their impact is canceled out by each other. PUMAs with a high percentage of people who did not change their residence in the survey year are positively related to high naturalized citizenship, age, and married percentage. These PUMAs also have a relatively low percentage of college degrees. Next, I will discuss other spatial variations in the migration study: the locality of underlying processes.

9.2.4 Spatial variations

From the neighborhood clustering analysis, we can see the locality nature of contextual factors and their effect on migration. Neighborhood type 1 is high in movers who migrated within the state. For migration within the state, two key indicators are born in Hong Kong and self-employed. Neighborhood type 2 is dominated by movers from other states. For migration between states, two key indicators are college degrees and being employed (specifically working for wages). However, the two internal migration categories (either between-states or within-state) are not entirely different in their predictors. The two migration categories have common variables related to marital statuses, such as being single, spouse absent, or married. It is not hard to understand the relationship between an individual's marital status and migration behavior. For married people, there are always more considerations from the family aspect. By building regression models on individual neighborhood types, the model fit has dramatically increased compared to modeling the whole study area. Compared to preexisting administrative boundaries, the neighborhoods derived in this study are more helpful for understanding the research problem.

GWR and MGWR models reveal spatial variations within the underlying processes of migration as well. The two geographically weighted regression models have dramatically improved the model fit compared to the global model for between-states migration. The OLS model had an adjusted R^2 of 0.38 for the between-state migration. This statistic increased to 0.63 as the average, with the minimum R^2 of 0.54 and the maximum R^2 of 0.83 in the MGWR model. The GWR model has improved the model fit as well, just not as dramatically as the MGWR model. The GWR and MGWR models display a similar spatial pattern with the maximum R^2 in the southern areas and decreasing towards the north.

9.3 Practical implications

The results in this study have offered some insights into migration theories: spatial assimilation, ethnic enclave, resurgent ethnicity, and heterolocalism. These theories are not exclusive of each other. One explanation for the co-existence of different theories is the stratifications within the immigrant Chinese. Each subgroup population has its migration pattern and underlying processes. For example, while it is evident to see a growing immigrant population in suburban neighborhoods, the neighborhoods with a large Chinese population have experienced growth at the same time. The former process offers evidence of the spatial assimilation process, but the latter is a sign to support the ethnic enclave theory. With the influx of the population, laborers, and fortune, Chinese enclaves are thriving and resurging. Immigrant Chines may not cluster in traditional Chinatowns, but their ties are suggested to be tighter. The ethnic economy plays a positive role for the Chinese immigrants in building their identity as a whole in the US. It is hard to separate the influences of the above theories from each other in a migration study. Migration factors emphasized in them are intertwined (Fang and Brown 1999; Wang 2007).

It is important to dismiss the concept that individual Chinese immigrants follow the same assimilation pattern. Assimilation is not simply a spatial process. True assimilation involves the identification and acceptance of social norms and values (Zhou 2010). Stratifications exist in the Chinese population, contributing to the "branches" in the social nature of assimilation. Only a small portion of the immigrant Chinese goes through assimilation in the traditional way: receiving education in colleges and getting a job through which their socioeconomic status increases. Such educational purpose immigration is more evident among the immigrants from mainland China. Education is important. As indicated in the regression analysis for neighborhood type 2 (relatively low in Stability and SES, but slightly high in Chinese percentage), a college degree (or above) and being employed are the two significant indicators of between-states migration.

Self-employment offers an alternative path for the immigrant Chinese to increase their status in the social hierarchy. In the decision tree analysis, self-employment has a broader influence than education in differentiating people of different migration patterns. Self-employment is the third most significant splitting variable for the whole microdata set, whereas education is only significant in the decision tree for people from mainland China. In the PUMS data, 30% of the population define themselves as self-employed. A subgroup has come to the nation with fortune, of which a substantial portion is from Hong Kong. They live in big and nice houses, which can be seen from the number of bedrooms in their housing unit and the rent. This subgroup stands out in neighborhood type 1 in the within-state migration. In the regression analysis, being self-employed and obtaining a college degree are two significant variables. Housing conditions are their primary concern to make a move, mainly within the state. These entrepreneurs may not live in communities with large ethnic populations, but they do maintain connections in the ethnic enclave. The connections are from various ethnic resources in immigrant relocation, such as a familiar working environment and a channel for employment and housing information (Forbes 1984; Fang and Brown 1999; Lobo and Mellander 2020; Zhou 2010).

Not every self-employed Chinese was wealthy or an entrepreneur before they moved into the US. There are more immigrant Chinese who were not wealthy at the time they arrived in the US. Due to language and cultural obstacles, their skills cannot be easily transferred to obtain a job equivalent to the socioeconomic status in their home country. Self-employment offers a shortcut to achieve their goals. In some cases, they open ethnic-related stores to meet the needs of immigrant Chinese. Within ethnic enclaves, there is a large Chinese population, which offers abundant consumers and cheap labor from their co-ethnics. As indicated in previous research, self-employment has a particular meaning in the immigrant economy. Self-employed immigrants may provide their co-ethnics job opportunities or housing information since they can access more people, which benefits from residing in ethnic concentrations (Kritz and Nogle 1994; Chen 2017; Wang 2010).

It is also important to dispel the negative image attached to the enclave (Zhou 2010). The Chinese immigrant population has been growing. The profiles of the immigrant Chinese unveiled in this study are consistent with findings in previous research (Lee 2018; Li 1998; Bai 2015; Hooper and Batalova 2015; Kadarik 2019). Compared to their predecessors, the new immigrants are more diverse and with higher educational attainment. A college degree, English-language ability, and self-employment are evident in the regression analysis for people who moved to the US from abroad. Chinese enclaves are no longer clustering places of low classes. The enclave helps the Chinese population sustain their identity (Bodenner 2014). Moreover, the enclave economy forms a small community for the co-ethnics and offers them a shortcut to advance their socioeconomic status (Zhou 2010). With the constant inflow of people and capital, the Chines enclave is thriving, not declining.

Public assistance, neighborhood renovation, immigrant relocation, and job training are examples of conventional government assistance to ethnic minorities. Recognition and encouragement of ethnic economy and culture is government assistance as well. Compared to the former assistance methods, the latter would have a broader and more profound impact on the immigrant Chinese. For example, Chinese immigrants who worked in ethnic-labeled jobs are viewed as low class. However, Chinese culture puts the family ahead of the individual. It is not rare to see some Chinese parents work in the ethnic enclave for low wages while their children receive a good education in reputed institutions. The families may own a house in suburban areas. These parents are respectable and successful in Chines people's eyes. Ethnic populations and white people could have the same goals yet different paths to their goals. Acknowledging alternative paths to success is a means to racial equality. An inclusive society is more robust.

9.4 Limitations

This study has some data limitations. Though the Chinese concentrations in the NY-N-JC MSA are among the largest in the US, the population is still small compared to the white population. The small

population results in minimal access to data. A systematic examination of ethnic migration behavior requires many variables. Therefore, aggregated data tables on the US Census website do not work well since these tables keep only some general population statistics and do not include detailed attributes.

PUMS data have both advantages and disadvantages. The data set contains detailed information on Chinese immigrants facilitating research. However, it only has one geographical dimension, PUMA, within the state level, which limits the scale choice in the study. For example, PUMA is the basis for defining the physical boundaries of neighborhoods and the spatial unit for regression analyses. It limits the possibilities of expanding or comparing to migration patterns at other spatial scales. There are also limitations concerning the data quality. In this study, subgroups of the population have been shown based on the IPUMS data. However, bias may exist in the survey. One reason is the language obstacle. The accuracy of survey answers is questionable for respondents whose native language is not English. Moreover, Chinese people who are not comfortable with English may choose not to respond to the survey. Another reason is related to the migration status. People without a legal migration status have a high possibility of not answering the survey. Therefore, the data could be biased initially and not representative of the population.

9.5 Future research

Future research may add other human capital variables to consider the pulling effect from the ethnic enclave. It is not always the case that ethnic people obtain benefits from residing in the same community. Immigrants may reside dispersedly, but they are socially tight through specific ways, such as churches or other ethnic-related organizations (Gurak and Kritz 2000; Portes and Bach 1985). It would be beneficial to take into account social capital from those organizations.

Two other types of ethnic-specific data could be added in future research. Based on research results, self-employment is a crucial characteristic among the immigrant Chinese. Business information such

as locations, incorporated or not, gross avenue would provide a more detailed examination of the role of self-employment in their lives. In Chinese culture, family achievements hold a greater weight than personal accomplishments (mainly in career). Some parents are willing to sacrifice their careers for their children. Therefore, a study including the second even third generation of the immigrant Chinese could offer different angles to better understand the population.

Last but not least, this dissertation has some methodological implications. Data consists of geographic and non-geographic information. This study provides a more complete picture of the Chinese immigrants by adding a geographic perspective to the migration research. The methodology can apply to other areas and other ethnic groups. A method that simultaneously considers both geographic and non-geographic attributes in the ethnic migration study is promising.

REFERENCES

Alba, R. D., J. R. Logan, B. J. Stults, G. Marzan, and W. Zhang. 1999. Immigrant groups in the suburbs: A reexamination of suburbanization and spatial assimilation. *American Sociological Review* 64: 446-60.

Allen, J. P., and E. Turner. 2005. Ethnic residential concentrations in United States metropolitan areas. *Geographical Review* 95: 267-85.

Allen, J. P., and E. Turner. 2009. Ethnic residential concentrations with above-average incomes. *Urban Geography* 30(3): 209-38.

Andersson, F., M. Garcia-Perez, J. Haltiwanger, K. McCue, and S. Sanders. 2014. Workplace concentration of immigrants. *Demography* 51(6): 2281-2306.

Bagchi-Sen, S., T. Schunder, and X. Tai. 2020. An analysis of employment patterns of domestic migrants and immigrants in a Rustbelt city: A study of Buffalo-Niagara Falls. *Growth and Change* 51(1): 123-43.

Bai, W. 2015. A Portrait of Chinese Americans: From the Perspective of Assimilation. PhD diss., University of Arkansas.

Barabantseva, E. 2016. Seeing beyond an 'ethnic enclave': the time/space of Manchester Chinatown. *Identities* 23(1): 99-115.

Beekman, D. 2011. The changing Chinatowns: Move over Manhattan, Sunset Park now home to most Chinese in NYC. NY Daily News. Last accessed June 19, 2021. https://www.nydailynews.com/changing-chinatowns-move-manhattan-sunset-park-home-chinese-nyc-article-1.948028.

Bezdek, J. C., and R. J. Hathaway. 2002. VAT: A tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks*. IJCNN'02 (Cat. No. 02CH37290) 2002 May 12 (Vol. 3, pp. 2225-2230). IEEE.

Bhandari, A. 2020. Key difference between R-squared and adjusted R-squared for regression analysis. *Analytics Vidhya*. Last accessed June 5, 2021. https://www.analyticsvidhya.com/blog/2020/07/difference-between-r-squared-and-adjusted-r-squared/.

Bodenner, Z. J. 2014. "Knowing who you are": The role of ethnic spaces in the construction of hmong identities in the Twin Cities. PhD diss., Ohio University.

Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Belmont, California : Wadsworth International Group.

Budiman, A. 2020. Key findings about U.S. immigrants. Pew Research Center. Last accessed June 5, 2021. https://www.pewresearch.org/fact-tank/2020/08/20/key-findings-about-u-s-immigrants/.

Burt, J. E., G. M. Barber, and D. L. Rigby. 2009. *Elementary Statistics for Geographers*. New York: Guilford Press.

CGTN America. 2019. 'ChinaTown' in New York expands far beyond Manhattan. Last accessed June 18, 2021. https://www.youtube.com/watch?v=hB3jsH-aW4o.

Chacko, E., and M. Price. 2020. (Un) settled sojourners in cities: the scalar and temporal dimensions of migrant precarity. *Joural of Ethnic and Migration Studies*: Special Issue: (Un)Settled Sojourners in Cities: 1-18. doi: 10.1080/1369183X.2020.1731060.

Chavent, M., V. Kuentz, B. Liquet, and L. Saracco. 2011. ClustOfVar: An R package for the clustering of variables. arXiv preprint arXiv:1112.0295.
Chen, C.C.Y. 2017. Transformation of a New Chinese Immigrant Community in the United States: A Case Study in Flushing, New York (美國新華人移民社區的轉型—以紐約法拉盛為探討中心). *Translocal Chinese: East Asian Perspectives* 11(2): 208-29.
Claritas. 2021. Claritas PRIZM premier segment narratives 2020. Last accessed June 10, 2021. https://claritas.com/prizm-premier/.
Crawley, M. J. 2012. *The R book*. John Wiley & Sons.
Davis, M. 1992. Chinatown revisited? The internationalization of downtown Los Angeles. In *Sex, Death and God in L.A.* ed. D. Reid, 54-71. New York: Pantheon Books.
Dhillon, I., E. Marcotte, and U. Roshan. 2003. Diametrical clustering for identifying anticorrelated gene clusters. *Bioinformatics* 19(13): 1612-19.
Dymski, G. A., and J. M. Veitch. 1996. Financing the future in Los Angeles. In *Rethinking Los Angeles*, ed. M. J. Dear, H. E. Schockman, and G. Hise, 35-55. Thousand Oaks, CA: Sage Publications.
Ellis, M., R. Wright, and V. Parks. 2007. Geography and the immigrant division of labor. *Economic Geography* 83(3): 255-81.
Emeka, A. 2020. Free and clear: national origins and progress toward unencumbered homeownership among post-civil rights era immigrants in the US. *Journal of Ethnic and Migration Studies* 46(18): 3808-28.
Fang, D., and D. Brown. 1999. Geographical mobility of the Chinese born in large metropolises, 1985-1990. *International Migration Review* 33(1): 137-55.
Flynt, A., and N. Dean. 2016. A survey of popular R packages for cluster analysis. *Journal of Educational and Behavioral Statistics* 41(2): 205-25.
Forbes, S. S. 1984. Residency patterns and secondary migration of refugees. *Migration News 34*(1): 3-18.
Fotheringham, A. S., M. E. Charlton, and C. Brunsdon. 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning* A30(11): 1905-27.
Fotheringham, A. S., W. Yang, and W. Kang. 2017. Multiscale geographically weighted regression (MGWR). *Annals of the American Association of Geographers* 107(6): 1247-65.
Gabriel, S. A., J. P. Mattey, and W. L. Wascher. 2003. Compensating differentials and evolution in the quality-of-life among US states. *Regional Science and Urban Economics* 33: 619-49.
Gibson, C., and K. Jung. 2006. *Historical Census Statistics on the Foreign Born Population of the United States, 1850 to 2000*. Washington, DC: Population Division, US Census Bureau.
Grayson, G. W. 1995. *The North American Free Trade Agreement: Regional Community and the New World Order* (Vol. 3). Lanham: University Press of America.
Greenacre, M., and R. Primicerio. 2014. *Multivariate Analysis of Ecological Data*. Fundacion BBVA.
Gurak, D. T., and M. M. Kritz. 2000. The interstate migration of U.S. immigrants: Individual and contextual determinants. *Social Forces* 78(3): 1017-39.
Hall, M. 2009. Interstate migration, spatial assimilation, and the incorporation of U.S. immigrants. *Population, Space and Place* 15(1): 57-77.
Hall, M. 2013. Residential integration on the new frontier: immigrant segregation in established and new destinations. *Demography* 50: 1873-96.
Hartigan, J. A., and M. A. Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28 (1):100-08.
Hipple, S. F., and L. A. Hammond. 2016. *BLS Spotlight on Statistics: Self-Employment in the United States.* Washington, DC: Bureau of Labor Statistics.
Hooper, K., and J. Batalova. 2015. Chinese immigrants in the United States. *Migration Policy Institute* 28.

Hubert, L., and P. Arabie. 1985. Comparing partitions. *Journal of Classification* 2(1): 193-218.
IndexMundi. 2021. Pennsylvania Asian population percentage by county. Last accessed June 26, 2021. https://www.indexmundi.com/facts/united-states/quick-facts/pennsylvania/asian-population-percentage#chart.
James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.
Johnston, R., M. Poulsen, and J. Forrest. 2009. Research note—Measuring ethnic residential segregation: Putting some more geography in. *Urban Geography 30*(1): 91-109.
Julia, M. 2010. Northeast China Branches Out in Flushing. *The New York Times*. February, 9. Last accessed June 19, 2021. https://www.nytimes.com/2010/02/10/dining/10chine.html.
Kadarik, K. 2019. Moving out, moving up, becoming employed: Studies in the residential segregation and social integration of immigrants in Sweden. PhD diss., Uppsala University.
Kaplan, D. H. 1998. The spatial structure of urban ethnic economies.*Urban Geography* 19(6): 489-501.
Kaplan, D. H., and F. Douzet. 2011. Research In Ethnic Segregation III: Segregation Outcomes. *Urban Geography* 32: 589-605.
Kasarda, J. D. 1988. Jobs, Migration, and Emerging Urban Mismatches. In *Urban Change and Poverty,* ed. M.McGeary and L. E. Lynn, 148-98. National Academy Press.
Kassambara, A. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Scotts Valley, CA: CreateSpace.
Kaufman, L., and P. J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
Kritz, M., and D. T. Gurak. 2015. U.S. immigrants in dispersed and traditional settlements: national origin heterogeneity. *International Migration Review* 49(1): 106-41.
Kritz, M. M., D. T. Gurak, and M. A. Lee. 2011. Will They Stay? Foreign born Out-Migration from New U.S. Destinations. *Population Research and Policy Review* 30(4): 537-67.
Kritz, M. M., and J. M. Nogle. 1994. Nativity concentration and internal migration among the foreign born. *Demography* 31(3): 509-24.
Kuhn, M. 2013. Predictive modeling with R and the caret package. Last accessed June 5, 2021. https://www.r-project.org/conferences/useR-2013/Tutorials/Kuhn.html.
Lawson, R. G., and P. C. Jurs. 1990. New index for clustering tendency and its application to chemical problems. *Journal of chemical information and computer sciences* 30(1): 36-41.
Lee, C. A. 2018. The role of race/ethnicity in the spatial construction of neighborhoods and housing choice. PhD diss., UCLA.
Lee, E. 1966. A theory of migration. *Deomography* 3 (1): 47-57.
Li, W. 1997. Spatial transformation of an urban ethnic community from Chinatown to Chinese ethnoburb in Los Angeles. PhD diss., University of Southern California, Los Angeles.
Li, W. 1998. Anatomy of a new ethnic settlement: The Chinese ethnoburb in Los Angeles. *Urban Studies* 35(3): 479-501.
Light, I. H., and E. Bonacich. 1988. *Immigrant Entrepreneurs: Koreans in Los Angeles, 1965-1982*. Berkeley: University of California Press.
Lin, J., and P. Robinson. 2005. Spatial disparities in the expansion of the Chinese ethnoburb of Los Angeles. *GeoJournal 64*(1): 51-61.
Ling, H. 2005. Reconceptualizing Chinese American Community in St. Louis: From Chinatown to Cultural Community. *Journal of American Ethnic History 24*(2): 65-101.
Linoff, G. S., and M. J. Berry. 2011. *Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.
Liu, C. Y., and E. J. van Holm. 2019. The geography of occupational concentration among low-skilled immigrants. *Economic Development Quarterly* 33(2): 107-20.
Liu, C. Y., and G. Painter. 2012. Immigrant settlement and employment suburbanisation in the US: Is there a spatial mismatch?.*Urban Studies* 49(5): 979-1002.

Liu, J. M., and L. Cheng. 1994. Pacific Rim development and the duality of post-1965 Asian immigration to the United States. In *The New Asian Immigration in Los Angeles and Global Restructuring*, ed. P. Ong, E. Bonacich and L. Cheng, 74-99. Philadelphia: Temple University Press.

Lobo, J., and C. Mellander. 2020. Let's stick together: Labor market effects from immigrant neighborhood clustering. *Environment and Planning A*: *Economy and Space* 52(5): 953-80.

Logan, J. R., W. Zhang, and R. D. Alba. 2002. Immigrant enclaves and ethnic communities in New York and Los Angeles. *American Sociological Review* 67(2): 299–322.

Logan, J. R., S. Spielman, H. Xu, and P. N. Klein. 2011. Identifying and bounding ethnic neighborhoods. *Urban Geography* 32(3): 334-59.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1 (14): 281-97.

Maechler M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2005. Cluster Analysis Basics and Extensions. *R package version* 1.12.1.

McGlinn, L. A. 2002. Beyond Chinatown: dual immigration and the Chinese population of metropolitan New York City, 2000. *Middle States Geographer* 35(4): 110-19.

McGrew Jr, J. C., and C. B. Monroe. 2009. *An Introduction to Statistical Problem Solving in Geography.* 2nd ed. Waveland Press.

Massey, D. S., and B. P. Mullan.1984. Processes of Hispanic and Black Spatial Assimilation. *American Journal of Sociology* 89(4): 936-74.

Min, P. G., ed. 2006. *Asian Americans: Contemporary trends and issues* Vol. 174. Pine Forge Press.

Mukherjee, A., and B. K. Pattnaik. 2020. Assimilation, heterolocalism and ethnic capital: The case of an immigrant Indian community in America. *Sociological Bulletin* 70(1): 24-41. doi: 0038022920956737.

Newbold, K. B. 2010. *Population Geography: Tools and Issues*. Rowman and Littlefield.

Newbold, K. B., and M. Foulkes. 2004. Geography and segmented assimilation: examples from the New York Chinese. *Population, Space and Place* 10(1): 3-18.

Newman, M. E., A. L. E. Barabási, and D. J. Watts. 2006. *The Structure and Dynamics of Networks.* Princeton university press.

Ong, P., and J. M. Liu. 1994. U.S. immigration policies and Asian migration. In *The New Asian Immigration in Los Angeles and Global Restructuring*, ed. P. Ong, E. Bonacich and L. Cheng, 45-73. Philadelphia: Temple University Press.

Oshan, T. M., Z. Li, W. Kang, L. J. Wolf, and A. S. Fotheringham. 2019. Mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information* 8(6): 269.

Oshan, T. M., J. P. Smith, and A. S Fotheringham. 2020. Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. *International Journal of Health Geographics* 19: 1-17.

Ostrow, D., and D. Ostrow. 2008. *Manhattan's Chinatown*. Arcadia Publishing.

O'brien, R. M. 2007. A caution regarding rules of thumb for variance inflation factors.*Quality & Quantity* 41(5): 673-90.

Penninx, R. 2003. Integration: The Role of Communities, Institutions, State. Migration Information Source. Last accessed June 18, 2021. https://www.migrationpolicy.org/article/integration-role-communities-institutions-and-state.

PennState. 2018. Lesson 11: Model building. The Pennsylvania State University. Last accessed June 19, 2021. https://online.stat.psu.edu/stat462/node/197/.

Pieterse, J. N. 2003. Social capital and migration Beyond ethnic economies. *Ethnicities* 3(1): 29-58.

Portes, A., and R. L. Bach. 1985. *Latin Journey: Cuban and Mexican Immigrants in the United States.* Los Angeles: University of California.

Poston Jr, D. L., and H. Luo. 2007. Chinese student and labor migration to the United States: Trends and policies since the 1980s. *Asian and Pacific Migration Journal* 16(3): 323-55.
R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336): 846-50.
Robbins, L. 2015. With an influx of newcomers, little Chinatowns dot a changing Brooklyn. *New York Times*. Last accessed June 19, 2021. https://www.nytimes.com/2015/04/16/nyregion/influx-of-chinese-immigrants-is-reshaping-large-parts-of-brooklyn.html.
Roleke, J. 2019. Elmhurst in Queens, NY: Neighborhood Profile. Last accessed June 18, 2021. https://www.tripsavvy.com/elmhurst-queens-neighborhood-profile-2819270.
Ross, Z. 2017. Predictive modeling and machine learning in R with the caret package. ZevRoss: Know Your Data. Last accessed June. 5, 2021. http://zevross.com/blog/2017/09/19/predictive-modeling-and-machine-learning-in-r-with-the-caret-package/.
Ruggles, S., S. Flood, S. Foster, R. Goeken, J. Pacas, M. Schouweiler, and M. Sobek. 2021. IPUMS USA: Version 11.0 [2015 ACS-5year]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V11.0.
Sassen, S. 1995. Immigration and local labor markets. In *The Economic Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship*, ed. A. Portes, 87-127. Russell Sage Foundation.
Schiller, N. G., and A. Caglar, ed. 2011. *Locating Migration: Rescaling Cities and Migrants*. Cornell University Press.
Schuch, J. C., and Q. Wang. 2015. Immigrant businesses, place-making, and community development: a case from an emerging immigrant gateway. *Journal of Cultural Geography* 32(2): 214-41.
Scott, A. J. 1988. *Metropolis from the Division of Labor to Urban Form*. Berkeley: University of California Press.
Semple, K. 2009. In Chinatown, sound of the future is Mandarin.The New York. Last accessed June 18, 2021. https://archive.nytimes.com/www.nytimes.com/2009/10/22/nyregion/22chinese.html.
Shalizi, C. 2009. Distances between clustering, hierarchical clustering. *Lectures notes*. Carnegie Mellon University.
Shircliff, J. E. 2020. Is Chinatown a place or space? A case study of Chinatown Singapore. *Geoforum* 117: 225-233.
Shirkhorshidi, A. S., S. Aghabozorgi, and T. Y. Wah. 2015. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one* 10(12). doi: e0144059.
Sinha, B. R. K. 2005. Human migration: concepts and approaches. *Foldrajzi Ertesito* 3(4): 403-14.
Stillwell, J., and O. Duke-Williams. 2005. Ethnic population distribution, immigration and internal migration in Britain: what evidence of linkage at the district scale. In *British Society for Population Studies Annual Conference, University of Kent at Canterbury*, 12-14.
Storper, M., and R. Walker. 1989. *The capital imperative: territory, technology and industrial Growth*. Cambrideg: Blackwell Publishers.
Therneau, T. M., B. Atkinson, B. Ripley, and M. B. Ripley. 2019. Package Rpart. Last accessed June 26, 2021. https://cran.r-project.org/web/packages/rpart/rpart.pdf.
Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology* 63(2): 411-23.

Tobler,W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234-40.
US Census Bureau. 2016. Core-Based Statistical Areas. Last accessed June 26, 2021. https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html.
US Census Bureau. 2019. Census Tract. Last accessed June 26, 2021. https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13.
US Census Bureau. 2019. Hispanic Or Latino Origin By Specific Origin. Last accessed June 18, 2021.https://data.census.gov/cedsci/table?q=Race%20and%20hispanic%20or%20latino%20origin&tid=ACSDT1Y2019.B03001.
US Census Bureau. 2020. Census Tract Relationship Files. Last accessed June 6, 2021. https://www.census.gov/geographies/reference-files/2010/geo/relationship-files.html#par_list_0.
US Census Bureau. 2020. Foreign Born Population. Last accessed June 6, 2021. https://www.census.gov/topics/population/foreign-born/about.html#par_textimage.
US Census Bureau. 2021. Public Use Microdata Areas (PUMAs). Last accessed June 6, 2021. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html.
USAFACTS. 2021a. Immigrant population. Last accessed June 6, 2021. https://usafacts.org/data/topics/people-society/immigration/immigration-and-immigration-enforcement/immigrants/.
USAFACTS. 2021b. Population. Last accessed June 6, 2021. https://usafacts.org/data/topics/people-society/population-and-demographics/population-data/population/.
Vigneau E, and E. M. Qannari. 2003. Clustering of variables around latent components. *Communications in Statistics Simulation and Computation* 32(4): 1131-50.
Walton, E. 2012. Resurgent ethnicity among Asian Americans: Ethnic neighborhood context and health. *Journal of Health and Social Behavior* 53(3): 378-94.
Walton, E. 2017. Spatial assimilation and its discontents: Asian ethnic neighborhood change in California. *Urban Geography* 38(7): 993-1018.
Wang, Q. 2006. Linking home to work: Ethnic labor market concentration in the San Francisco consolidated metropolitan area. *Urban Geography* 27(1): 72-92.
Wang, Q. 2007. How does geography matter in ethnic labor market segmentation process? A case study of Chinese immigrants in the San Francisco CMSA. US *Census Bureau Center for Economic Studies Paper No. CES-WP-07-09*.
Wang, Q. 2010. How does geography matter in the ethnic labor market segmentation process? A case study of Chinese immigrants in the San Francisco CMSA. *Annals of the Association of American Geographers* 100(1): 182-201.
Weeks, J. R., A. Getis, A. G. Hill, S. Agyei-Mensah, and D. Rain. 2010. Neighborhoods and fertility in Accra, Ghana: An AMOEBA-based approach. *Annals of the Association of American Geographers* 100(3): 558-78.
Wen, M., D. S. Lauderdale, and N. R. Kandula. 2009. Ethnic neighborhoods in Multi-Ethnic America, 1990-2000: Resurgent ethnicity in the Ethnoburbs? *Social Forces* 88(1): 425-60.
Yu, S. 2018. Mobilocality. *Urban Geography* 39(4): 563-86.
Zelinsky, W. 1971. The hypothesis of the mobility transition. *Geographical Review* 61(2): 219-49.
Zelinsky, W., and B. A. Lee. 1998. Heterolocalism: an alternative model of the sociospatial behaviour of immigrant ethnic communities. *International Journal of Population Geography* 4(4): 281-98.
Zhou, M. 2004. Revisiting ethnic entrepreneurship: convergencies, controversies, and conceptual advancements. *International Migration Review* 38(3): 1040-74.
Zhou, M. 2010. *Chinatown: The Socioeconomic Potential of an Urban Enclave*. Temple University Press.

Zhou, M., and M. Lin. 2005. Community transformation and the formation of ethnic capital: Immigrant Chinese communities in the United States. *Journal of Chinese Overseas* 1(2): 260-84.
Zhou, M., and J. R. Logan. 1991. In and out of Chinatown: Residential mobility and segregation of New York City's Chinese. *Social Forces* 70(2): 387-407.
Zhou, Y. 1998. How do places matter? A comparative study of Chinese ethnic economies in Los Angeles and New York City. *Urban Geography* 19(6): 531-53.

APPENDICES

A.1a Variable evaluation for between-states migration

Variable	Est.	SE	t(Est/SE)	p-value
Intercept	-0.000	0.065	-0.000	1.000
naturalized	-0.241	0.083	-2.892	0.004
college	0.315	0.068	4.653	0.000
separated	0.302	0.077	3.944	0.000
selfEmp	-0.105	0.070	-1.492	0.136
married	0.016	0.139	0.112	0.911
onlyEng	-0.149	0.090	-1.649	0.099
employed	0.180	0.071	2.533	0.011
single	0.328	0.142	2.320	0.020

A.1b Adjusted critical t-values in GWR and MGWR models

Regression model for between-states migration		
Predictor variable	Adj. critical t (95%)	
	GWR	MGWR
% naturalized citizens	2.293	2.189
% college degree or above	2.293	2.073
% people separated from spouse	2.293	2.163
% people self-employed*	2.293	2.493
% people married*	2.293	2.812
% people speaking only English*	2.293	2.519
% people employed	2.293	2.435
% people single	2.293	2.051

A variable marked with a * is one whose parameter estimate is not significantly different from zero.

A.2a Variable evaluation for migration from abroad

Variable	Est.	SE	t(Est/SE)	p-value
Intercept	-0.000	0.061	-0.000	1.000
naturalized	-0.038	0.079	-0.487	0.626
wellEng	0.254	0.068	3.728	0.000
notWellEng	-0.093	0.094	-0.990	0.322
AGE	-0.287	0.093	-3.086	0.002
married	-0.058	0.120	-0.484	0.628
single	0.213	0.129	1.653	0.098
college	0.040	0.104	0.390	0.696
selfEmp	0.033	0.067	0.489	0.625
nLabor	0.440	0.072	6.086	0.000
wage	0.118	0.093	1.267	0.205
rent	-0.013	0.071	-0.178	0.859

A.2b Adjusted critical t-values in GWR and MGWR models

Regression model for % moved abroad		
Predictor variable	Adj. critical t (95%)	
	GWR	MGWR
% naturalized citizens*	2.275	2.081
% people speaking English well	2.275	2.147
% people speaking English, but not well*	2.275	2.404
Age	2.275	2.468
% people married*	2.275	2.059
% people single*	2.275	2.760
% college degree or above*	2.275	2.047
% people self-employed*	2.275	2.129
% people not in labor force	2.275	2.704
Wage*	2.275	2.809
Rent*	2.275	2.104

A variable marked with a * is one whose parameter estimate is not significantly different from zero.

A.3a Variable evaluation for stay at the same house

Variable	Est.	SE	t(Est/SE)	p-value
Intercept	0.000	0.065	0.000	1.000
naturalized	0.275	0.089	3.074	0.002
AGE	0.186	0.090	2.068	0.039
married	0.310	0.089	3.472	0.001
college	-0.240	0.071	-3.395	0.001
selfEmp	-0.101	0.073	-1.386	0.166
nLabor	-0.065	0.076	-0.861	0.389
onlyEng	0.141	0.089	1.590	0.112

A.3b Adjusted critical t-values in GWR and MGWR models

Regression model for % Stay at the same house		
Predictor variable	Adj. critical t-value (95%)	
	GWR	MGWR
% naturalized citizens	2.229	2.117
Age	2.229	2.217
% people married	2.229	2.199
% college degree or above	2.229	2.095
% people self-employed*	2.229	2.377
% people not in labor force*	2.229	2.333
% people speaking only English*	2.229	2.194

A variable marked with a * is one whose parameter estimate is not significantly different from zero.

VITA

Yanxia Wu

Candidate for the Degree of

Doctor of Philosophy

Dissertation:
SPATIAL ANALYSIS OF ETHNIC MIGRATION BEHAVIOR:
A CASE STUDY OF CHINESE IMMIGRANTS IN THE NEW YORK-NEWARK-JERSEY CITY METROPOLITAN AREA

Major Field: Geography

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Geography at Oklahoma State University, Stillwater, Oklahoma in July, 2021.

Completed the requirements for the Master of Science in Civil Engineering at Beijing University of Technology, Beijing City, China in 2010.

Completed the requirements for the Bachelor of Science in Civil Engineering at Yanshan University, Qinhuangdao City, Hebei Province, China in 2005.