How will *hukou* reform affect the city system in China?

Newest Version

December 28, 2018

By Yijiao Liu

This paper applies the novel spatial dataset, China Spatial Administrative Unit Coding System (CN-SAUCS), constructed by Liu (2018) to quantify the consequences of *hukou* reform trials prior to the national 2014 *hukou* reform Opinions on Further Reform of the *Hukou* System, on migration flows to different types of destinations in China. The 2014 reform was designed to encourage migration to small or medium size cities, and away from large urban centers. By using a discrete choice model, the estimates show that there is more than one possible strategies of *hukou* reform will achieve the aim, while also reveal the general preferences of the Chinese migrants: people prefer richer places which are closer to their *hukou* registration locations. For the already reformed zones, though they can be more attractive to migrants from further geographical locations than before, averagely, those zones would be impoverished in the beginning of the implementation of the opening-up migration policies.
JEL: J1, O0, R0
Keywords: hukou reform, urbanization, migration, discrete choice, spatial geocoding

1. Introduction

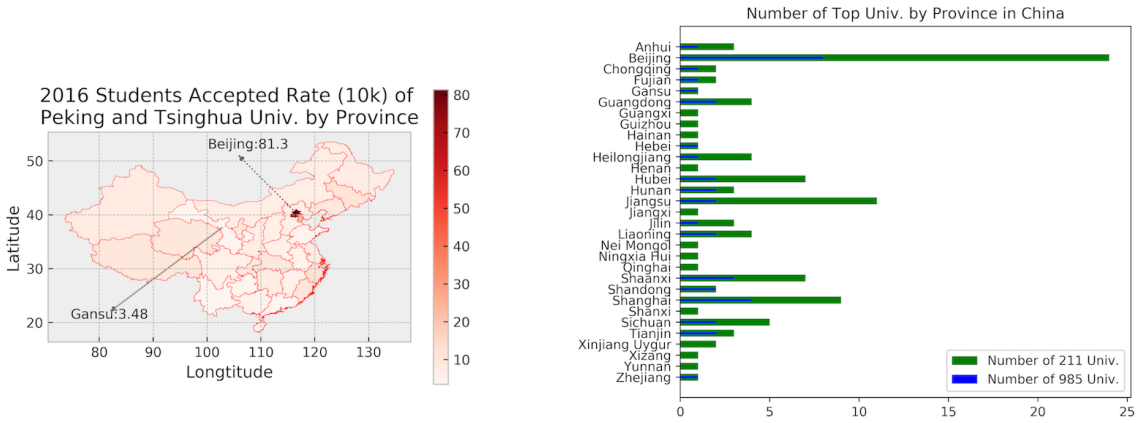
In China, one' s permanent residence is registered in a *hukou* system. This well-known 'internal passport system' was built at the beginning of the establishment of the People' s Republic of China and was used as a tool to tightly control and monitor migration. This project leverages the novel spatial dataset constructed in Liu (2018) to quantify the consequences of *hukou* reform trials prior to 2014, on migration flows to different types of destinations in China.

Historically, one' s *hukou* location has been profoundly connected to the level of public services and entitlements one receives. (National People's Congress Standing Committee, 1958). Residents who are registered with different types of *hukou* –both in terms of location and the urban or rural designation of that location–will receive different social benefits which vary along with the policies of the local government. However, it is difficult to change one' s *hukou* from rural to urban areas, or from smaller cities to larger cities (Chan, 2013).

With the largest population in the world, China has 245 million people who are referred to as the 'floating population' ; these people do not hold a *hukou* in the area where they currently live. The 'floating population' are the *de facto* but not *de jure* population (NBS National

Bureau of Statistics, 2017) [1]. These internal migrants, mostly rural migrant workers, started to migrate to places without having a local *hukou* during the economic reform in 1978. The tremendous demographic shift of these rural migrants has mainly driven the urbanization process over the past few decades (Chen, Davis, and Landry, 2017). Rural migrant workers are typically ineligible for the full scope of local government services (e.g. pension, health insurance, higher education opportunities for dependents) due to their *hukou* being held in different locations.

FIGURE 1.1. DIFFERENCE IN LIKELIHOOD OF COLLEGE ACCEPTANCE CAUSED BY DIFFERENT *Hukou* LOCATION



Note: The first chart shows the variance of acceptance chances by the top two universities in China. The quota of college acceptance is given differently to each province, and students have to take national entrance exams back in the location of their hukou. The second chart shows the unbalanced education resources distributed across provinces by providing the number of '985' and '211' universities that each province contains.

On the one hand, since natural urban population growth rate is slow in large cities, to achieve the goal of augmenting the shares of consumption, China needs to shift the rural population to cities to expand the middle-class population of the city. However, the recent shortage of rural migrant workers in the coastal manufactories shows that the current generation of rural migrant workers is seeking equal rights and non-discriminatory wage rates from other urban residents(Friedman and Kuruvilla, 2015). Ultimately, it is their non-localized *hukou* that is the main obstacle preventing them from residing permanently in the urban areas. This phenomenon has further delayed urbanization progress in China (Song, 2014).

On the other hand, mega cities such as Beijing and Shanghai, are confronted with congestion issues. The 2017 China Urban Research Report of Baidu Maps shows congestion cost in Beijing is the highest, on average, which is one of the leading concerns of the local government about opening up its door to outsiders.According to the Opinions of Beijing Municipal People's Government (2016), the government aims to strictly control and decrease the number of *de facto* residents of the municipal districts by limiting the registration of new *hukou* of the migrants. It is not hard to tell that the main concern of local governments in mega cities is that opening up *hukou* policy would make congestion more severe.

In 2014, the State Council of China issued the 'Opinions on Further Reform of the *Hukou* System' aiming to accelerate the urbanization process. The reform was designed to encourage

[1] *De facto* population: the residents live in the current locations; *de jure* population: the residents register their permanent living address at the current locations. (Chan, 2013)

migration to small or medium size cities, but away from large urban centers. It is expected that the floating population could be motivated by the reformed *hukou* policies and shift to the opened-up city zones with the gained rights from the post-reform *hukou*.

Though the opening-up policies would encourage rural migrants, the variances in the development or the level of social benefits that one could gain in smaller cities and larger cities may weaken the effectiveness of *hukou* policy reform' s ability to redistribute the population (Chen, Davis, and Landry, 2017). In the meantime, there are 57.18 million rural migrants working and living in other rural places, leaving a lack of labor support for agricultural work in their own *hukou* villages (NBS, 2018). No previous study has analyzed the impact of this reform on the potential migration patterns into the less congested urban zones.

The task is to evaluate whether or not the goal of the 2014 *hukou* Opinions would be met, and to examine if the post-reform migration inflows would rebalance the population distribution of the city hierarchies and accelerate urbanization progress. It is essential to quantify the effect of *hukou* reform on the population of each type of place using accurate detailed records of the population, city hierarchy and spatial information.

The discussion of *hukou* reform comes from two perspectives. The government' s intention is to utilize the reform of *hukou* policy to accelerate urbanization progress and alter population distribution. The migrants' concerns focus on maintenance or enhancement of utilities after updating their *hukou*. Failing to address and achieve either side would not break migration barriers in the quick and smooth fashion that is intended by the reform.

First, we need to find the answer to whether or not changing *hukou* policies will impact migration, both across urban areas and from rural to urban locations. Since the 2008 global financial crisis, labor shortages have become noticeable in the manufactories located along the coastal zones (Chan, 2010). Some researchers argued that China already reached the Lewis' turning point, meaning that the country has exhausted cheap rural migrant labor (Das and N'Daiye in 2013)). Other researchers evaluate the phenomenon as a paradox of the co-existence of a shortage of labor in urban areas in China with an abundant supply in rural areas. Previous studies agree on the fact that the urban labor market of China has faced a shortage of rural migrant workers in recent years. As one of the major institutional constraints, the *hukou* system is an obstacle to the free flow of labor from rural to urban areas (Cui, Meng, and Lu, 2018). However, there is still a need to quantify how much the change of the *hukou* policies could alter population distribution: if the reformed *hukou* policy is strong enough to advance urbanization growth, while controlling the population of crowded cities.

Secondly, we need to address the feasibility of implementing the *hukou* system reform policies. Governments, institutions, and researchers have debated heavily over the types of reform policies. As shown in the research report of Sun (2017)), well before the 2014 Opinions, starting in the early 2000s, local governments had implemented different *hukou* policies regarding migration inflows. In Sun' s research, he also quantified the extent to which major cities of China had opened up *hukou* policies, and found that larger cities with a higher rate of *de facto* floating population had more restricted policies against *hukou* registrations submitted by outsiders. He also argued that *hukou* reform policies need to incorporate the retention of preceding land property rights for rural migrants to encourage urbanization progress. The research proposes strategies to guide further *hukou* reform, however, without access to the *hukou* policy data, this study is unable to dissect the reasons for those proposals.

This study is the first one to quantify the impact of different *hukou* reform policies on shifting rural migration inflows into different types of urbanized areas. Based on the experimental

hukou reforms carried out in some provinces before 2014, using a conditional logit discrete choice model, I combine data from the 2010 Chinese General Social Survey (CGSS) and the China Spatial Administrative Unit Coding System (CN-SAUCS, Liu, 2018) to study how *hukou* reform affects location choices in various regional levels within China (in village scale). The CN-SAUCS is a comprehensive spatial census dataset that I built and have maintained since 2017. It relies on Google/Baidu/Tencent/Gaode Maps API Web-services, which includes information on village-level administrative units and yields a link to cities, towns, and rural partitions of local populations in China. My new comprehensive dataset (CN-SAUCS) on the population geographic characters of China allows me to estimate the impact of *hukou* policy on the residential choices of (potential) migrants.

Estimates from the conditional logit discrete choice model of a big choices dataset imply that regions that undertook reforms prior to 2014 are, on average, poorer. Generally, migrants prefer more prosperous destinations in China and more helpful neighbors, and the close relationship of neighbors discourages inflows. The counterfactual effects indicate that *hukou* reform would accelerate urbanization progress if the fundamental rights of the rural migrants that were attached to their before-reformed *hukou* are retained.

This paper contributes to the literature by showing the power of *hukou* reform on migration decisions, and alterations of city hierarchy, in a quantitative and qualitative study. It also expands practical strategies for policymakers by showing the effects of different *hukou* policies on different aftereffect of the population distribution of China. Moreover, as the first application of CN-SAUCS, it demonstrates the necessity for and efficiency of distinguishing places and population in China by using correct and consistent census data.

2. Migration and *Hukou* Reform Background in China

According to the report of national rural migrants (NBS, 2018), the trend of the total amount of rural migrants reached to 286.52 million population, including 137.10 million rural migrants residing in different types of urban destinations. The majority of the 'floating population', those who hold a *hukou* that differs from their residential location, has limited access to the public services provided by the city to which they have migrated. The report also shows that only 38% of rural migrants feel 'localized' by living in a city without holding that city's *hukou*. This phenomenon is more severe in larger cities than in less urbanized regions. About 44% of rural migrants do not feel satisfied by their current living status in the city, and more than half of the children of rural migrants have difficulties enrolling in local schools.

Those facts motivate this research to answer the question of how reformed *hukou* could affect people's choice of residential location. The portion of the floating population who have already moved out of their original *hukou* locations could be most affected by the variance of *hukou* policies in different places, since *hukou* is the directly correlated to one's social benefits and the corresponding public services offered by the local government. Under the 2014 *hukou* reform, and the sequential *hukou* reforms by different local governments in recent years, people might change their residential location based on the possibility that their related social benefits may be improved or restricted as compared to their current living location. As a result, opening *hukou* registration in a city could attract more a greater flow of migrants, while the restrictions of *hukou* registration in a city could make the life of rural migrants more difficult there. This research focuses on how changes to a location's *hukou* registration policies would affect people's likelihood to migrate into that zone.

Starting in 2003, some local governments in China started *hukou* reform trials to address inequities faced by the floating population, as compared to *hukou* holding residents in that area. By 2011, more than ten provinces implemented some *hukou* reform trials by canceling the urban/rural designation of the type of *hukou* (People.cn, 2013). By holding a 'residential *hukou*' instead of 'urban/rural' *hukou*, discrimination towards incoming migrants should be diminished. However, the fiscal burden of accepting a rural migrant to a city could impact the efficiency of the process and may decrease the government's commitment to the project, ultimately, resulting in a decision to return to a more restricted *hukou* registration policy. On the other hand, rural migrants are not willing to give up ownership of their rural land, along with their entitlement within their rural *hukou*. To sum up, *hukou* decides the majority of one's public services, and different *hukou* types have different public services and entitlements attached to each person in China.

Generally, a *hukou* type is defined by its registration's location, at the village-level unit:

- (1) urbanization degree (urban/rural), and
- (2) relation to the upper-level governments,
  - i.e., if this village-level unit belongs to the core part of the city or the townships that are governed by the small county residing within the big city zone. The levels of social security or entitlement attached to this village-level community vary based on the ruling by those different governments from different administrative levels.

A reformed *hukou* policy generally means that people have less/no restrictions to register their *hukou* at that reformed region. Urban/rural distinction would be canceled either by holding a *hukou* at a location within the city boundary without labeling its urban/rural designation, or by giving up the original *hukou* to register at a new location within that reformed region. In the *hukou* reform trial, people have the freedom to choose to either keep their before-reformed *hukou*, or change to a new post-reform *hukou*.

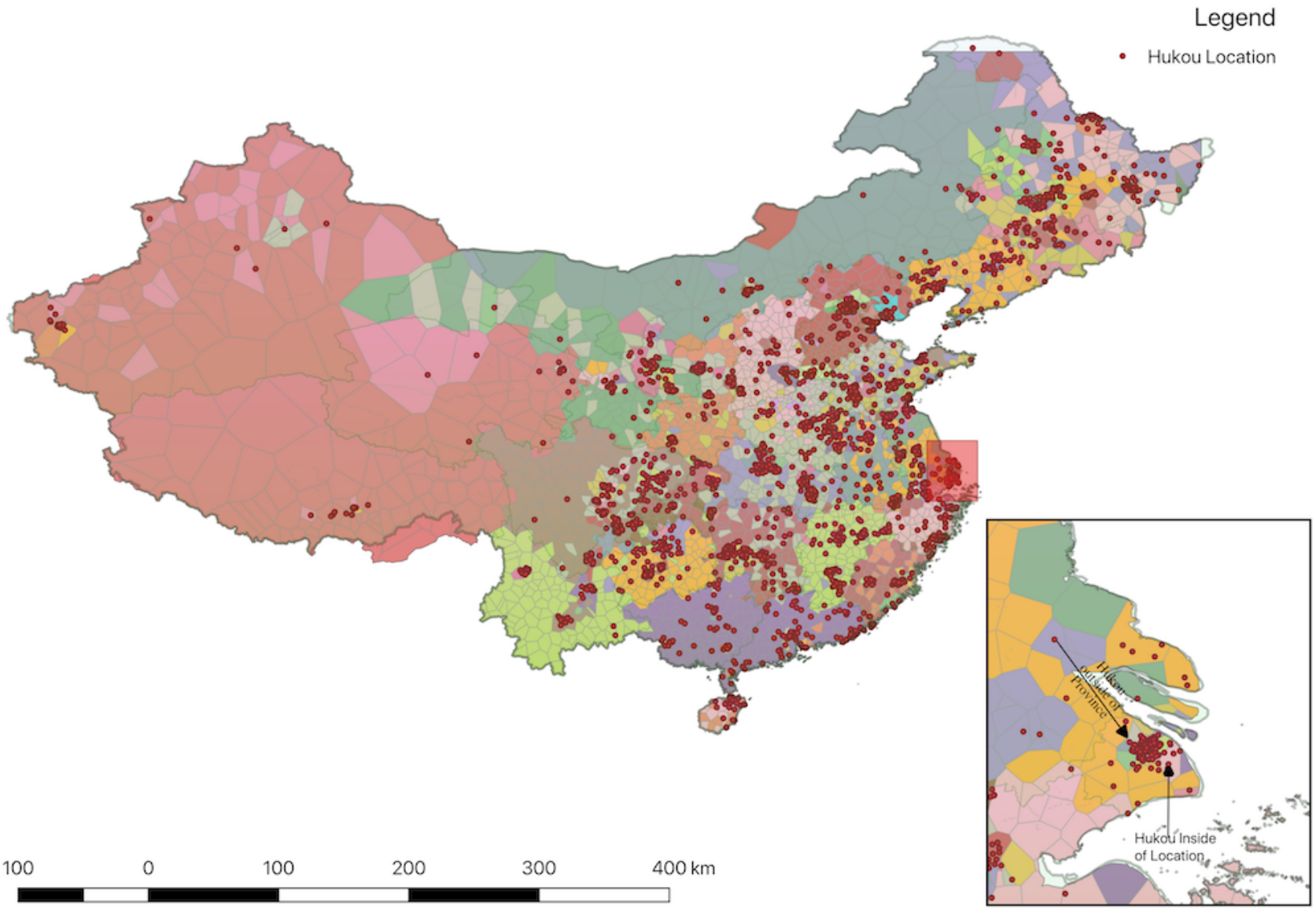
To understand how the changing policies of *hukou* could affect one's decision on where to live, or the inflows of migrants into different regions, I am going to combine survey data, comprehensive spatial census data, and data from statistics yearbooks to check the past experimental *hukou*-reform in the pre-2014 era. From the perspective of individual choices and accurate population distribution, this research is aiming to explain the causal effect of opening up *hukou* policies, while forecasting migration inflows into different types of cities based on different *hukou* policies, i.e., the counterfactual effect of an opening-up policy vs. a more strictly-controlled policy.

3. Data

3.1. Overview

I am using the 2010 wave of CGSS survey data, CN-SAUCS, and 2010 statistics yearbooks to partition China into 469 separate location-choice zones. In Figure 3.1, each location choice alternative is represented by a color. A person's *hukou* from the survey data is only represented in one location, indicated by red coordinates. Each person's *hukou* location is treated

Figure 3.1. Location Choices (Partition) V.S. *Hukou* Locations



as an individual feature that does not vary across location choices. In practice, a person could hold their *hukou* in the location where they currently live, or in another place across China.

In the following subsections, I will demonstrate the method to get the location choices and their related variables. First, I am going to introduce CGSS survey data and make a compression of the other famous migration survey of China. Second, I am going to show how I combine the CGSS survey data with CN-SAUCS dataset to create partitions of location alternatives of China. As a complement to describe each location more, the next subsection with outline the variables I am choosing from statistical yearbooks. In the end, the full description of the variables is shown with statistics visualization tools and tables.

3.2. Migration Survey Data–CGSS Survey

3.2.1. Cross-Sectional Data Selection

To investigate the migration choices made by each potential migrant in China under the influence of their current *hukou* status, I apply the survey data with a detailed record of his or her current living residential locations with the corresponding *hukou* status (location and location-type). An individual' s choice of location is assumed to be affected by either the features of that location alternative or by personal interactions between individual characteristics and features of that location. We assume that no single individual features (that do not vary across choices) could affect residential preferences. [2] However, the interactions between individual features and location choices, such as people' s social securities tied to the location choice, are affecting his/her choices. In the conditional logistic discrete choice model, we do not necessarily observe every interaction between individual characteristics and the location choices. By including them in the unobserved utility part, the differences that result from the possibility of an individual choosing a location because of the variation of their interactions could be captured.

The main features we need to collect for this study are the current living locations of the individual, their *hukou* location (which determines the *hukou* type), some features of each of the location choices, and some features of the types of interactions people may have at each different location.

For this purpose, I firstly employ survey data obtained from the Chinese General Social Survey (CGSS), which are collected and maintained by the Department of Sociology of Renmin University and 26 other universities in China. [3] Established in 2003, the CGSS project is the earliest national representative continuous survey project which reflects the radical transition of Chinese society and changes in Chinese behavior and attitude. There has been two phases to the project in order to collect the strictly probability-sampling data, with over 10,000

[2]There are 469 choices across China in this study; it is hard to believe that individual features, such as gender, would have an effect on different location choices. For example, if gender always has the same effect on various location alternatives, the effect will be canceled out because the difference of the utilities raised by gender is always the same across alternatives. If gender generates a different partial effect on location choices, it would mean that being a male/female going to one $place^{469}$will have a different partial effect of being a male/female going to another $place^{468}$. However, to calculate that we would need to calculate 468 coefficients just for one variable. It is hard to believe that the results of that large amount of coefficient would be significant. In general, we can assume that since all types of locations are represented among the 469 location choices across China, the non-changing individual features would not generate a large difference between locations when one is facing a choice.

[3]Description from the official website of CGSS project: http://www.chinagss.org/index.php?r=index/introduce

weighted respondents (adults that were at least 18 years old at the time of each wave) from cities, small townships and rural villages across all 31 provinces in mainland China.

Notably, in each wave of the finished 2010, 2011, 2012, 2013, and 2015 surveys of the second phase, a person' s last living location and current *hukou* status (including nature and location of that *hukou*) were recorded together, along with current investigated location as his/her location choice. CGSS data keeps records of the residential location of each investigator with detailed village-level information; this is the smallest boundary unit to designate urban/rural property of a place by the government, and also the smallest administrative unit in China to rule its residents. [4]

Considering that the surveys from the second phase of the CGSS are rotating panel datasets with 50% of respondents quitting in the next wave, and that different individuals are assigned different spatial weights in different waves, this research is focused on the cross-sectional data for one wave. It is notable that China implemented the sixth national census in 2010. At the same time, some provinces started to implement experimental *hukou* reform trials, while others remained unreformed. Examining the influence of *hukou* reform on different individuals would follow a 'natural experiment' before the 2014 *hukou* reform. Incorporating Census data and the records with individuals' locations detailed in village-level administrative units (either residents' committees or villagers' committees) would help to conduct nice counterfactual studies to forecast the population for each type of city with various *hukou* reform policies.

3.2.2. Data Description of CGSS 2010

The 2010 survey has 11783 weighted individuals located in 469 village-level administrative units, covering 31 provinces and all types of places (mega-, super-large-, large-, small- sized cities, townships and rural regions) in China. The respondent from each household is randomly selected, in which are chosen from 480 SSU out of 140 PSU by a multi-stage stratified sampling design. We can observe a person's current living location, his/her last living location (if migrated), his/her *hukou* location, his/her *hukou* type (before reform *hukou* or not reform *hukou*), and so on. We assume that a *hukou* reform trial (before 2014) was at least implemented by prefectural-level city government. Within the same *hukou*-experimented province boundaries, the *hukou* policies are comparable to similar types of the locations within the region. The basic migration pattern of the respondents is shown in the following graphs.

In Figure 3.2,the CGSS 2010 survey shows that, among the current residents of a location, there are two groups of residents: one with local *hukou* (*de facto* and *de jure*), and one without a local *hukou* (*de facto* but not *de jure*). In each group, there are people who have moved to their current location either before or after the point at which *hukou* reform trials started in various regions across China. Meanwhile, Figure 3.3 indicates whether people also update their *hukou* location after they start living in their current location. Findings show that 5% of the *de jure* local residents have changed their *hukou* from other places to their current living locations since *hukou* reform trials started in China. It is important to consider that *hukou* reform trials had been implemented less than seven years at the time when 2010 CGSS survey was investigated. This was much shorter than the history of strictly-controlling *hukou* registration policies. Along with rapid urbanization facts, 5% of those who migrate to the new places after the *hukou* reform trials fit the motivation of this study. The lifetime

[4]However, the village-level unit does not set up a formal local government. Instead, it is a self-governed grassroots organization that helps the upper-level government implement the rules and regulations within its political power.

migration records are not available from CGSS survey, but it still shows the relationship between migration choices and different *hukou* policies.

Figure 3.2. Point at Which People Move to Current Locations

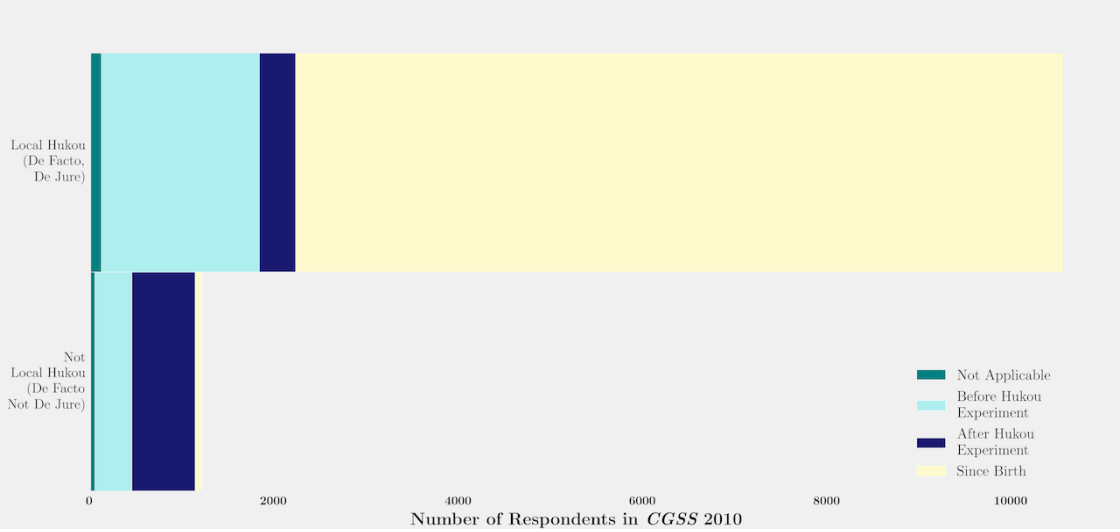
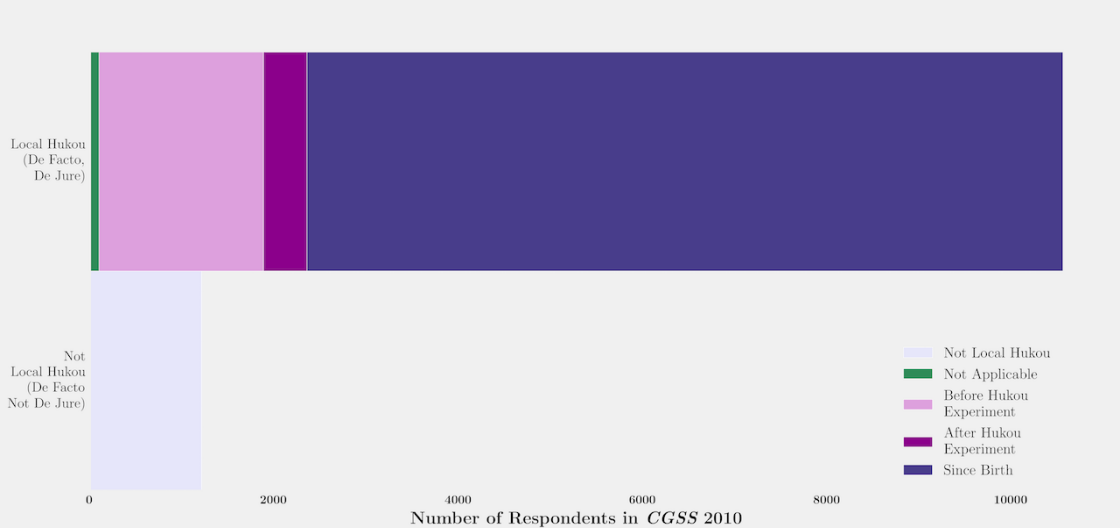


Figure 3.3. Point at Which People Change *Hukou* to Current Locations



Next, using the CGSS 2010 data, I construct the *hukou* reformed regions dummy for each location by observed *hukou* type (i.e. either before or after *hukou* reform). If a record of an post-reform *hukou* location within that city is observed, the whole region of the city is assumed to have implemented the reform of the *hukou* policy.

Moreover, to collect local features which demonstrate the heterogeneity across 469 locations, I select some variables that describe neighbors. The CGSS survey asked the respondent questions such as air pollution, neighbor safety level, and neighbor infrastructure level to help the researchers better understand the local social environment. Those questions are designed

to be self-reported, but could be used as a measure to demonstrate the local environment, in detail.

3.2.3. Sample Design and Weights for Respondents

Though the CGSS survey carefully follows the sample design rules as a multi-layer stratified sampling study, the bias is generated by the wrong definition of city zones with its corresponding population, when designing the weights for each. The weight of each individual follows the formula

$$w = \frac{City_{pop}}{Village_{pop}} \frac{Village_{hhs}}{HH}$$

where $City_{pop}$ is the total population of the city each individual is in, $Village_{pop}$ is the population of his current village-level unit, $Village_{hhs}$ is the number of households within that village-level unit, and HH is the number of investigated households in this village-level unit.

The problem generated from the $City_{pop}$, where the *hukou* registration (*de jure*) population was used to generate the weights, rather than the population of actual living residents (*de facto* population). This could be solved just by correcting the population data on the numerator. To get the correct city *de facto* population for each place, we need to refer to the information provided by the CN-SAUCS data. The specific method will be introduced in the next (model) section.

3.2.4. Comparison: CLDS Survey

Another possible survey data resource is from the China Labor-force Dynamics Survey (CLDS) by Zhongshan University in China, established in 2011. This series of surveys are also rotating panels with 50% individuals substituted in the next round. It covers most of the regions in China, except for Tibet, and sets up multi-layer stratified sampling based on different levels of geographic regions (eastern, middle, western parts of China) and provincial, county-level, and village-level administrative boundaries.

There are flaws with the CLDS survey. First, the information about a person' s current living location and *hukou* location are masked. We do not know this person' s exact *hukou* location, and we only know the city-level of a person' s current living location. There are features of the locations that help to identify the type of the place it could be. However, missing the real geographical locations hurts our ability to define vital information: what is the local level of social security insurance that is covered by this place, and is it the same as other places in the same city region that are observed in the survey? Without a clear answer to that, one' s *hukou* related social benefits could not be incorporated into the model, and we could barely conduct the causal structural model of how much the reform of *hukou* registration policy could impact post-reform migration inflows into different regions.

Moreover, in the 2011, 2012, 2014 waves, we cannot observe any record of a reformed-type *hukou*. It does not consider the fact that some of the province regions already implemented the reform on *hukou* registration or *hukou* types by 2011. It is hard to understand if this problem is caused by a misunderstanding by the survey designer about the types of *hukou*, or

by the fact that they did not investigate samples within the reformed *hukou* regions. Either way, it causes a sample bias problem.

However, there is still useful information from this survey to support the argument that *hukou* may affect a person' s migration choices. For example, the 2012 wave of the CLDS survey shows that among the 16,253 investigated respondents, more than 10% of them held a non-local *hukou*, and 89% had been living in their current location without updating their current *hukou* for more than half a year. The other 90% of the respondents were holding their local-*hukou* within the same city or county of their residence, but 74% of them had updated their previous *hukou* location to the current *hukou* location. Approximately 9% of updates happened between 2003 and 2012.

It is not difficult to tell from the evidence and facts that while there is a substantial 'floating population', there are also many people adjusting their *hukou* location during the experiments of the *hukou* reform. This implies that studying the relationship between location choices and *hukou* related policies could provide more guidance for policymakers about migration inflows into different zones in China.

3.3. Spatial Census Data–CN-SAUCS

We have randomly selected the 469 location choices to represent different types of places all over China. We built the choice set based on 469 village-level units to separate China into 469 choice zones, which are defined as the set of location choices for each Chinese citizen.

We divide mainland China, geographically, into 469 separate sections by using the spatial data of each level of administrative boundaries, census data (which correctly indicates village type, e.g. large city, rural area within a city, etc.), and the corresponding *de facto* population in each place. The CN-SAUCS data set is the comprehensive spatial census data that keeps records of the data mentioned above. First, we split all village-level units of China into five categories: mega-, super large or large-, medium or small- sized cities, townships, and rural areas. Second, we combine sites that are close, geographically, into village level units. Upon completion of these steps, I partition China into 469 regions, which are distinguished on the map by different colors. Each of the regions represent the choice variable in the model.

We also need the geographic location information of one' s *hukou* location. There are two reasons this is important. First, a *hukou* location provides the village-level urban/rural designation and the corresponding local social insurance level covered by its governing cities. We need to capture this *hukou* location as an individual-level characteristic that could interact with different location choices. We need to understand if this *hukou* location resides within the same administrative boundaries as their current residence, and how far away it is from their current living choices (zones). The second reason to apply *hukou* location from CN-SAUCS data is to fill-in incomplete information about one' s *hukou* location. I assigned a person' s *hukou* location to the same type of the place (mega-, super large or large-, medium or small-sized cities, townships and rural areas) within the administrative boundaries of the *hukou* location that we could observe.

Throughout the course of this study, I choose 469 coordinates to represent the location choice set, which are considered the residential sites. In the map, each *hukou* location is drawn as a point, which resides within one of the established 469 region choices. The *hukou* location of

a person could be in the current choices (colored) region but have a distance to the center of the choice region. It means, a choice region must contain the same type of the village-level units but may contain regions from different provinces.

The goal is to see migration inflows into different types of places, rather than different administrative boundaries. During the experimental *hukou* reform, as well as to this day in many places, people have the freedom to choose where to register their *hukou*, which makes *hukou* registration more detached from residency choices. The difference between *hukou* positions and residential sites only captures the different combination of the interaction between different types of personal *hukou* within or across different region options.

3.4. Aggregate Data from Statistical Year book

When choosing a place to live, it is not only the economic and social factors of a place that matter, but also the location itself. Two-dimensional location coordinates could provide the geographic distance between each of the residential neighbors but realistically, more factors are considers when making a decision of this nature. In China, generally, there are different social and economic environments generated and separated by northern/southern and eastern/western China. I select average yearly temperature and average yearly rainfall to capture the meteorological characteristics of the big geographical regions. On average eastern regions experience more rainfall than the west, and southern regions are warmer than the northern part. Data are collected by 2010 national and provincial statistical yearbook of China. The two columns that highlight this data will help to convey information in the model such as: is a location' s general weather condition going to affect migrant' s choices on where to live? Do warmer places within a region that shares the same *hukou* type?

Also, the statistical yearbook provides useful information for the regions within each location. I selected average disposable income level and average housing square meters for each person of that region as additional information to help capture the differences across location choices.

3.5. Construction of the dataset

Finally, a discrete choice 'panel data' is set up for this study. Each case (individual) is faced with 469 choices all across China, if we assume that everyone has the right to choose where he wants to live. The variables I am choosing for location alternatives are: Average Rainfall, Average Temperature, Average Housing Area, Average Disposable Income, *Hukou* Reform Dummy, Air Pollution, Water Pollution Environment, Food Access, Infrastructure, Safety, Neighbor Relations, and Support from Neighbors. The data statistics are shown in Figure 3.4 , Figure 3.5, and Table 1.

In Figure 3.4, the histograms of each continuous variables of the location alternatives, and the boxplots by the *hukou* reform dummy groups are shown.

In Figure 3.5, the histograms help visualize self-reported ordered discrete variables of the location alternatives, and the boxplots by the *hukou* reform dummy groups. As mentioned above, the ordered categorical variables are actually the mean values for the individuals within the same location; they are rounded to 2 digits after taking the mean values.

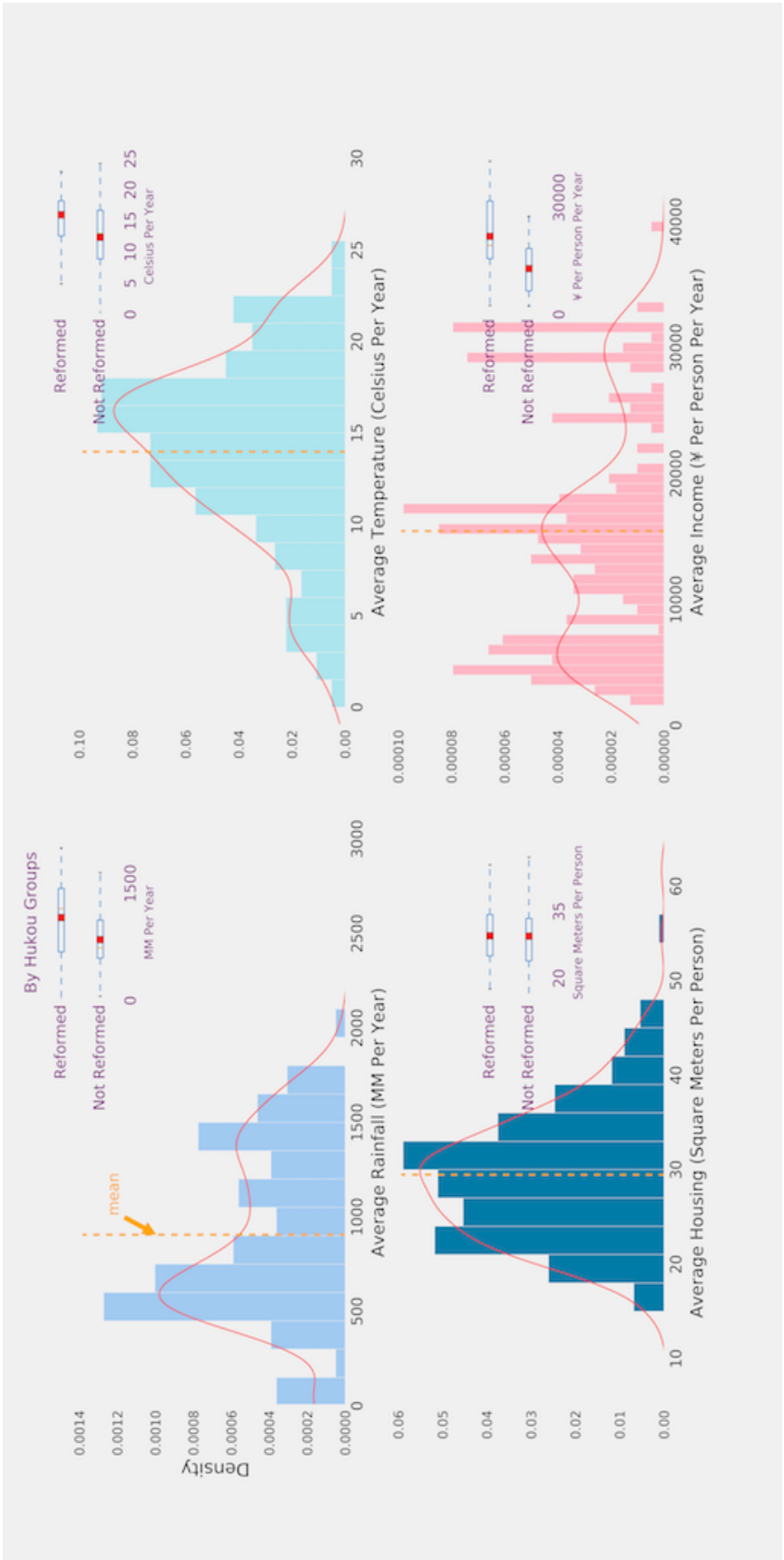


FIGURE 3.4. CONTINUOUS VARIABLES OF LOCATION ALTERNATIVES

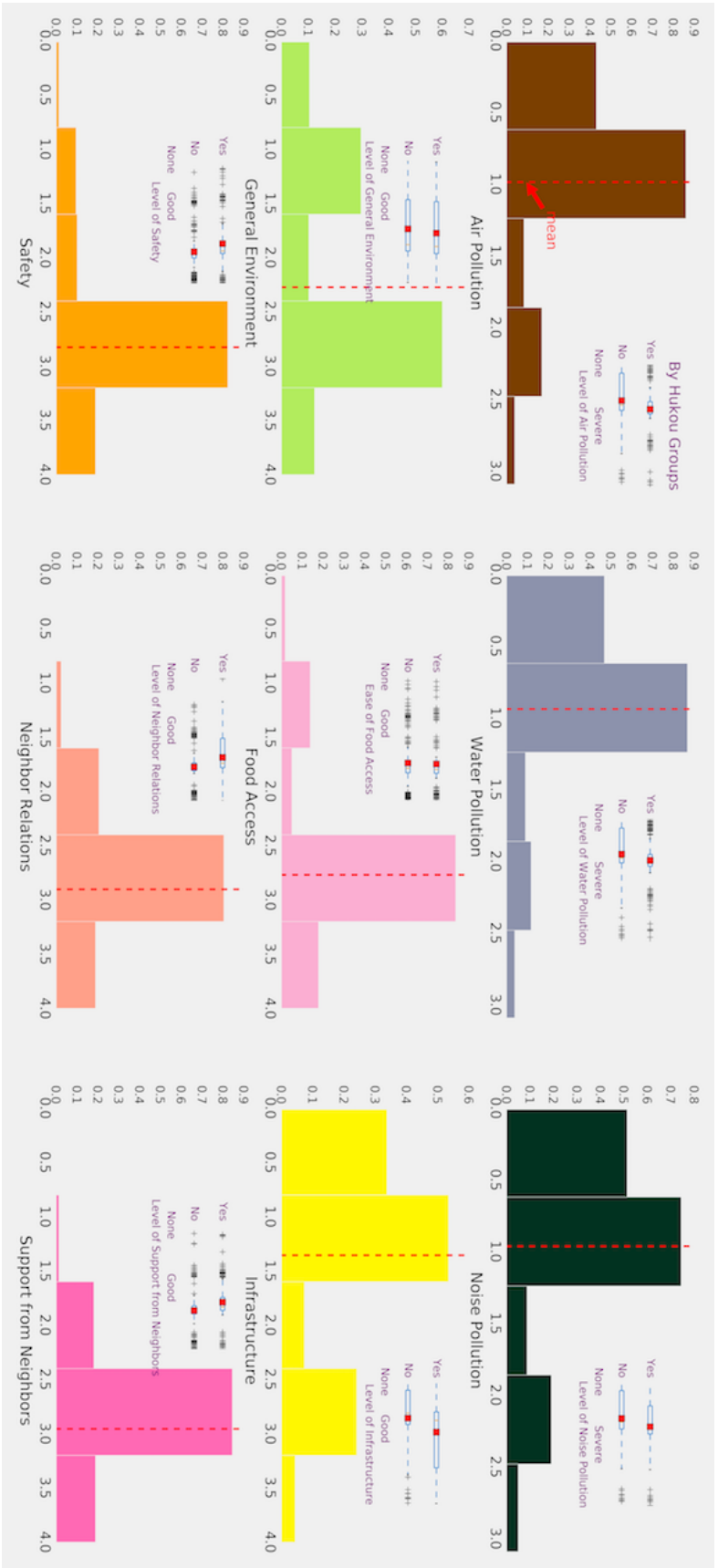


FIGURE 3.5. SELF-REPORTED VARIABLES OF LOCATION ALTERNATIVES

Table 1—Statistics for Variables of Location Alternatives

Variables	Mean	Standard Deviation
Air Pollution	0.99	-0.64
Average Disposable Income	15427.78	-9028.68
Average Housing Square Meters	29.47	-7.07
Average Rainfall	905.82	-462.53
Average Temperature	13.98	-5.07
Environment	2.27	-0.99
Food Access	2.77	-0.77
Hukou Reformed Dummy	0.44	-0.5
Infrastructure	1.36	-1.04
Support from Neighbors	2.96	-0.51
Neighbor Relations	2.9	-0.58
Noise Pollution	0.96	-0.68
Safety	2.83	-0.7
Water Pollution	0.94	-0.62

The interactive characteristics for each person with different locations are: Distance Between *Hukou* and Choice (in kilometers), and a dummy for *Hukou* and Choice being within the same city. Figure 3.6 shows the statistics of these two variables. Per the discrete choice 'panel data' design, we need to expand the survey data to create the panel choices dataset. In the original survey, most of the residents held their *hukou* within the same city of the current living location. However, if we allow the residents to choose every possible location in China, as the right-top subplot of Figure 3.6 shows, most of the people will not have their *hukou* in the same city as their current living location.

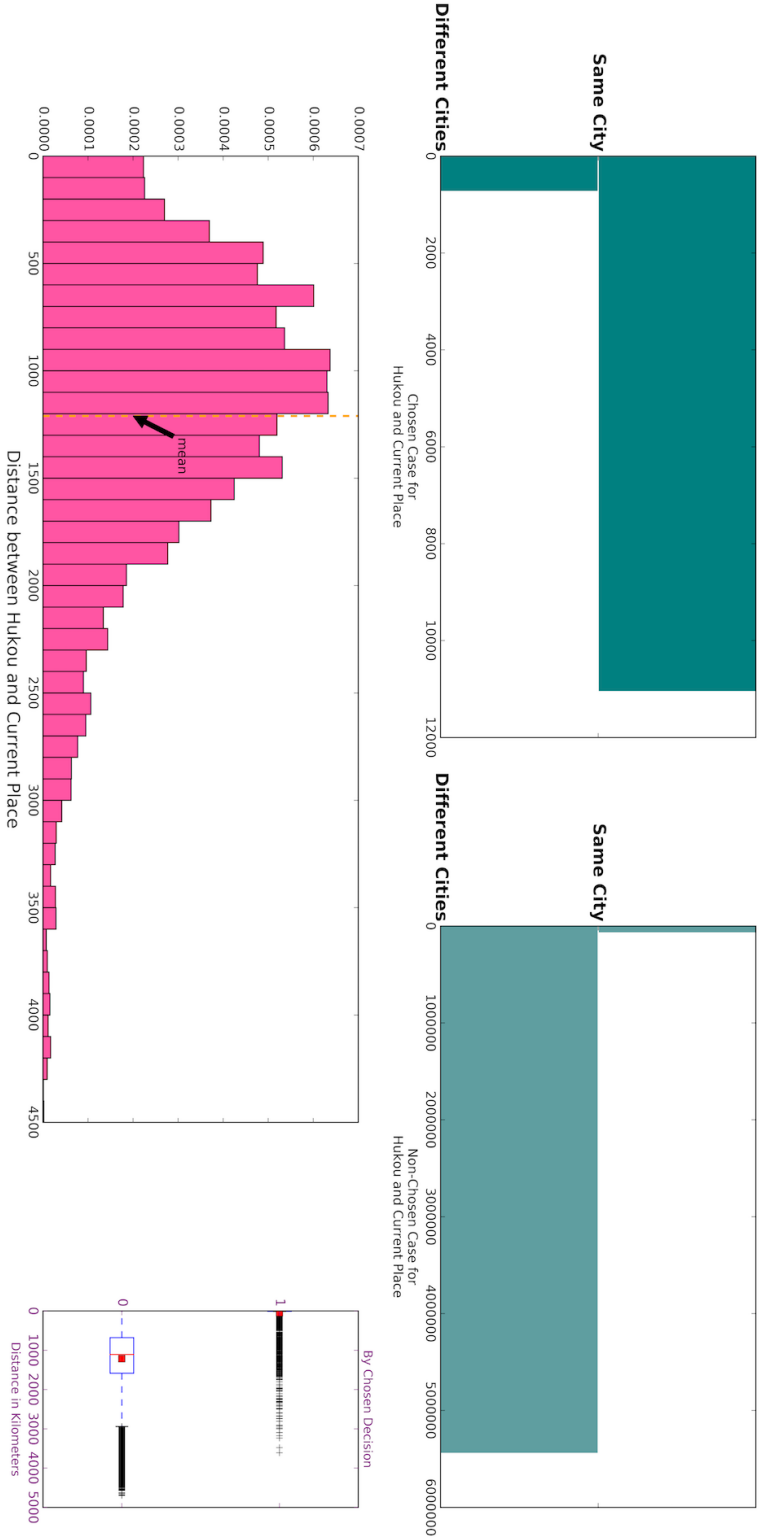
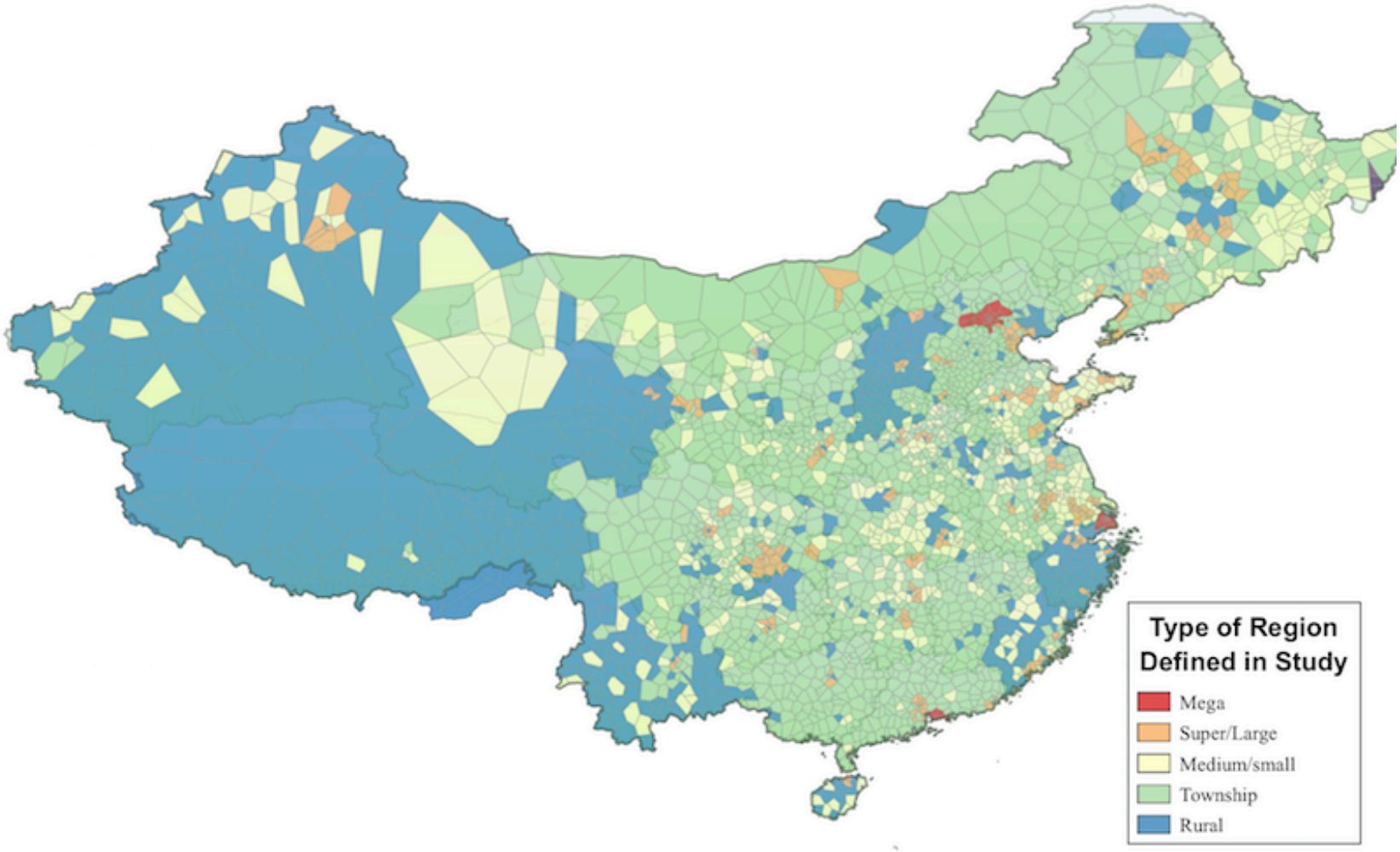


FIGURE 3.6. INTERACTION VARIABLES BETWEEN LOCATIONS AND INDIVIDUALS

FIGURE 3.7. TYPES OF PLACES CONTAINS IN EACH PROVINCE

FIGURE 3.8. TYPES OF PLACES DEFINED IN THE PAPER



Note: For instances when certain province cities were not represented by the survey' s respondents, I merged them with a neighboring province that shares similar characteristics. In the mapped visualization of the location choices, the regions of some locations j could be overdrawn or underdrawn due to survey samples; ultimately, this does not affect the qualitative results or basic special conclusions.

4. Model: Conditional Logit Model for Discrete Choices

4.1. Settings of the discrete choice model

The main structure of the models in this paper is inspired by the methods mentioned in Train (2009). In this section, I will follow those methods to build the models for this paper.

4.1.1. Basic conditional (multinomial) logistic model

The model for the discrete location choices of each Chinese citizen is set up by using conditional logit models. Specifically, the model for the choice of a location j of a person n is derived under an assumption of utility-maximizing behavior by the migrant n, which can also be seen as simply describing the relation of explanatory variables to the outcome of a location choice, without referencing exactly how that choice is made. Locations alternatives js would be the partitions that exhaust mainland China, and the decision maker n is every Chinese, who is ultimately seen as a potential migrant.

This random utility model (RUM) makes the utility U_{nj} of a migrant n choosing location j composed by: (1)V_{nj}: the utility that is observable by the researchers; (2)ε_{nj}: the utility that is not observable by the researchers, but is known by the decision maker. We have the following linear RUM

$$\begin{aligned} U_{nj} &= V_{nj}(X_{nj};\beta) + \varepsilon_{nj} \\ &= \beta^T X_{nj} + \varepsilon_{nj} \end{aligned}$$

The observed location alternative of a person maximize the utility of him/her, meaning that n chooses alternative i if and only if $U_{ni} > U_{nj}\ \forall j \neq i$. The RUM means that we treat the unobserved utility ε_{nj} as a random term, which follows a independent, identically distributed Gumbel and type extreme value distribution:

$$P.D.F \quad f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}$$

and

$$C.D.F \quad F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}$$

The following assumptions fit for all the logit based models in this paper. The reform of the *hukou* policy is captured by a dummy variable, 1 means the opened up location provides free *hukou* registration to migrants and 0 means that it holds a restrictive *hukou* policy to new migrants. In reality, different levels of opening-up policies in reformed *hukou* regions exist, and will even vary over time. Since we are applying the cross-sectional data, and the conditional logit model, the difference in the opening-up policies across locations j would be captured in the unobserved utility term ε_{nj} for each migrant n and each location j. The *hukou* reform dummy of each location j is an exogenous variable and is only decided by the policy makers.

The probability for person $n \in N$ to choose location $i \in J$ is

$$\begin{aligned} P_{ni} &= Prob(U_{ni} > U_{nj} \quad \forall j \neq i) \\ &= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \quad \forall j \neq i) \\ &= Prob(V_{ni} - V_{nj} > \varepsilon_{nj} - \varepsilon_{ni} \quad \forall j \neq i) \\ &= \int_{\varepsilon} (\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \forall j \neq i) f(\varepsilon_n) d\varepsilon_n \\ &= \int (\prod_{j \neq i} \varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \forall j \neq i) f(\varepsilon_{ni}) d\varepsilon_{ni} \\ &= \frac{e^{V_{nj}}}{\sum_j e^{V_{nj}}} \end{aligned}$$

where P_{ni} is the logistic choice probability, and with the location choice sets partitioning China into disjoint zones, $\sum_{i=1}^{J=469} P_{ni} = 1$.

The variables of each choice would at least vary across choices (alternatives), or vary across both location choices and personal associations with locations (interaction). It is clear from the above probability formula that if the choice only varies across different interactions, the difference in the utilities choosing different alternatives will cancel those unchanged individual effects in basic logit models. In this model, the parameter of each choice feature or each interaction feature is constant.

To estimate the conditional logit discrete choice model, we will apply the traditional maximum-likelihood approach. Assuming that peoples' decisions are independent from each other, the log-likelihood function for every Chinese citizen is

$$LL(\beta) = \sum_n \sum_i y_{ni} ln P_{ni}(X_{ni}; \beta)$$

where y_{ni} is the dummy variable to indicate if location i is chosen as a residential location by person n ($y_{ni} = 1$, if n chose i).

Since coefficient β of independent variables do/does not vary across alternatives j, it makes the basic logit choice model hold the independence from irrelevant alternatives (IIA) property, which does not allow one' s taste to vary among location choices.

For instance, given the average disposable income level of the locations choices, the IIA property of the logit model indicates that people generally prefer a richer zone (if $\beta_{disposable\_income} > 0$):

- As long as it is richer, people prefer this rich rural village to CBD zones in cities.
  - In reality, villages such as Huaxi Village in Jiangsu Province, which is among the richest rural villages in China, would show the consistency of the IIA property from the model. According to a report from the *Independent*, the village even imported migrant works from other places.
- As long as it is richer, the level of how much people prefer that zone is the same for everyone.

– If people prefer to live in Beijing, then it has the same effect on increasing the utility for every Chinese to live in Beijing, regardless of whether or not the person holds a Beijing local *hukou* or not.

The limitation of the application of the standard logit model would occur if we are interested in the variation of an individual's tastes across locations based on certain features. I will introduce the mixed logit model in the following part to solve this problem.

4.1.2. Mixed Logit Model

Suppose now that we allow for individual's tastes to vary across location alternatives. Based on the basic logit model, the probability of a mixed logit follows the form

$$\begin{aligned} P_{ni} &= \int L_{ni}(\beta) f(\beta) d\beta \\ &= \int \frac{e^{V_{ni}(\beta)}}{\sum_j e^{V_{nj}(\beta)}} f(\beta) d\beta \end{aligned}$$

which can be treated as a weighted average of the logit model with different values of β.

Especially, if we write down the RUM for a mixed logit model

$$U_{nj} = \beta^T x_{nj} + \alpha_n^T r_{nj} + \varepsilon_{nj}$$

where x_{nj} and r_{nj} are groups of variables that vary across location js and may/may not vary across individuals at the same time. People are assumed to have personal taste over r_{nj}s. We can see r_{nj}s as the random components that compose the stochastic part of the utility, which have coefficient $\alpha_n = E(\alpha_n) + \sigma_{nj} = \alpha + \sigma_\alpha \eta_{nj}$, where $\sigma_{nj} \sim i.i.dN(0, \sigma_\alpha)$ $\eta_{nj} \sim i.i.dN(0,1)$. Then the RUM can be expressed as

$$U_{nj} = \beta^T x_{nj} + \alpha^T r_{nj} + \sigma_{nj}^T r_{nj} + \varepsilon_{nj}$$

where the random coefficient α_n are seen to vary randomly with mean α and the standard deviation σ_α around the mean. The IIA will not be held since individual's variation term enters to the P_{nj}.

The estimation will come from the simulation of the variation of the random coefficients, i.e., σ_{nj}.We will first randomly draw Q times of η_{nj} from its distribution, and take the average of the simulated probability

$$\check{P_{ni}} = \frac{1}{Q} \sum_{q=1}^{Q} L_{ni}(\eta^q; \beta, \alpha, \sigma_\alpha)$$

. Then we insert this simulated probability to the log-likelihood function for a simulated log likelihood

$$SLL = \sum_n \sum_j y_{nj} ln \check{P_{ni}}$$

where $y_{nj} = 1$ if n chose j, and zero otherwise. We will then estimate β, α, σ_α by taking the maximum simulated likelihood (MSLE) , which are the values that maximizes SLL. If the

results show that σ_{alpha}, the standard deviations of the coefficients, are significant, it implies that the mixed logit model behaves significantly better than the basic logit model. Moreover, in our case, it means that migrants personal preference will generate different effects on their utilities, even when location features are the same, since the difference of the utility now is

$$U_{ni} - U_{nj} = \beta^T(x_{ni} - x_{nj}) + \alpha_n^T(r_{ni} - r_{nj}) + (\varepsilon_{ni} - \varepsilon_{nj}) \; \forall i \neq j$$

even when $r_{ni} = r_i \; \forall i \in J$.

4.1.3. Endogeneity Test: A Control Function Approach in Mixed Logit Model

With the construction of a mixed logit model, we can try to test and solve the potential endogeneity in our explanatory variables. Particularly, we would like to know if one's *hukou* related variables are correlated with the unobserved utility. On the one hand, one's *hukou* is directly connected with the level of social benefits that he/she could receive; on the other hand, one's *hukou* type or *hukou* location only varies over different individuals (cases) but not the location choices, which means even when faced with different locations choice options, the effect of different *hukou* that only vary across individuals, and not locations, would be cancelled out.

However, it would be interesting to consider the possible effects from the the interaction variable, such as the distance between the *hukou* location and one's current living location choices, on the unobserved utilities. The distance between each person's *hukou* and their current living locations may capture not only geographic (spatial) distance between the two places, but also affects things such as social network and administrative boundaries, which could possibly be correlated with the unobserved utility.

To check the potential endogeneity, I apply the control function approach here. Specifically, based on the mixed logit model, we revise the RUM as follows

$$U_{nj} = \beta^T x_{nj} + \alpha_n^T r_{nj} + \theta^T s_{nj} + \varepsilon_{nj}^1 + \varepsilon_{nj}^2$$

where the new entered s_{nj} is group of potential endogenous variables, correlated with the unobserved utilityε_{nj}^1, and uncorrelated with the unobserved utility $\varepsilon_{nj}^2 \sim i.i.d \; extreme \; value$. We can find its instrument z_{nj} which gives that

$$s_{nj} = \gamma^T z_{nj} + \mu_{nj}$$

where z_{nj} independent with ε_{nj}^1, $cov(\mu_{nj}, \varepsilon_{nj}^1) \neq 0$, and μ_{nj} independent with ε_{nj}^2. More specifically, the control function of ε_{nj}^1 is $CF = E(\varepsilon_{nj}^1|\mu_{nj})$, which makes $\varepsilon_{nj}^1 = E(\varepsilon_{nj}^1|\mu_{nj}) + \widetilde{\varepsilon}_{nj}^1 = \lambda\mu_{nj} + \widetilde{\varepsilon}_{nj}^1$. Together, the RUM is now

$$U_{nj} = \beta^T x_{nj} + \alpha_n^T r_{nj} + \theta^T s_{nj} + \lambda\mu_{nj} + \widetilde{\varepsilon}_{nj}^1 + \varepsilon_{nj}^2$$

where $\widetilde{\varepsilon}_{nj}^1 = \sigma_{\widetilde{\varepsilon}}\eta_{nj}^{\widetilde{\varepsilon}} \sim N(0, \sigma_{\widetilde{\varepsilon}})$, and $\eta_{nj}^{\widetilde{\varepsilon}} \sim N(0, 1)$. We can now treat the error term apart from the control function of ε_{nj}^1 as a random coefficient. The choice probability of each j now is mixed over both α_n , and $\sigma_{\widetilde{\varepsilon}}\eta_{nj}^{\widetilde{\varepsilon}} \; \forall j$. Given the random draws of $\eta_{nj}^{\widetilde{\varepsilon}}$, we can estimate the standard deviation of $\widetilde{\varepsilon}_{nj}^1$, which is $\sigma_{\widetilde{\varepsilon}}$. If the $\sigma_{\widetilde{\varepsilon}}$ is not statistically significant, we will reject that endogeneity in variables s_{nj}.

4.2. Partial Effects

The estimated coefficients are not the partial effects in the logit models, though the sign of the coefficients would basically give us a sense of the direction of the effect(Magazzini, 2014). The partial effects indicate how changing features in any location would affect the probability of a location alternative being chosen. There are two cases for each individual n. If the feature in the data is continuous, first, we can check how the change of the features of the choice location i affect the probability for n to choose i, which is

$$\frac{\partial P_{ni}}{\partial x_{ni}} = \frac{\frac{\partial e^{V_{ni}}}{\partial \sum_k e^{V_{nk}}}}{\partial x_{ni}} = \frac{\partial V_{ni}}{\partial x_{ni}} P_{ni}(1 - P_{ni})$$

In a standard logit model, assumed linear utility in $V_{ni}, \frac{\partial V_{ni}}{\partial x_{ni}} = \beta$ for continuous variable. In a mixed logit model, if x_{ni} is the variable with random coefficient, then $\frac{\partial V_{ni}}{\partial x_{ni}} = \beta + \sigma_\beta \eta_{nj}$, where $\eta_{nj} \sim N(0, 1)$. The partial effect should be

$$\begin{aligned}\frac{\partial P_{ni}}{\partial x_{ni}} &= \int \frac{\partial V_{ni}(\eta_{nj})}{\partial x_{ni}} P_{ni}(\eta_{nj})(1 - P_{ni}(\eta_{nj})) f(\eta_{nj}) d\eta_{nj} \\ &\approx \frac{\sum_r}{R} \frac{\partial V_{ni}(\eta_{nj})}{\partial x_{ni}} P_{ni}(\eta_{nj})(1 - P_{ni}(\eta_{nj}))\end{aligned}$$

The average partial effect of location i on x over all the people n should be $\frac{1}{n} \sum_n \frac{\partial P_{ni}}{\partial x_{ni}}$ in both cases $\forall\ i$.

In the second case, similarly, the partial effect of the features of the choice location j with respect to the probability for n to chooses i is

$$\frac{\partial P_{ni}}{\partial x_{nj}} = \frac{\frac{\partial e^{V_{ni}}}{\partial \sum_k e^{V_{nk}}}}{\partial x_{nj}} = -\frac{\partial V_{ni}}{\partial x_{nj}} P_{ni} P_{nj}$$

The standard logit model assumes linear utility in V_{ni}, $\frac{\partial P_{ni}}{\partial x_{nj}} = -\beta P_{ni} P_{nj}$ for continuous variables. As long as we know the sign of β, the increase on the feature of another location j will decrease the probability of choosing i, if $\beta > 0$, and vise versa. In a mixed logit model, if x_{ni} is the variable with random coefficient, then $\frac{\partial V_{ni}}{\partial x_{nj}} = \beta + \sigma_\beta \eta_{nj}$, where $\eta_{nj} \sim N(0, 1)$.

$$\begin{aligned}\frac{\partial P_{ni}}{\partial x_{nj}} &= \int -\frac{\partial V_{ni}(\eta_{nj})}{\partial x_{nj}} P_{ni}(\eta_{nj}) P_{nj}(\eta_{nj}) f(\eta_{nj}) d\eta_{nj} \\ &\approx \frac{\sum_r}{R} - \frac{\partial V_{ni}(\eta_{nj})}{\partial x_{nj}} P_{ni}(\eta_{nj}) P_{nj}(\eta_{nj})\end{aligned}$$

Since our model has $J = 469$, I am only going to report the partial effects of the feature on its own location, otherwise there would be $469 * 468$ partial effects for each variable of x_{nj}.

For dummy variable, I calculate the probability with zeros and ones for each person n at location j, and take average of the difference between the two values of the dummy variables for n of each location j ($\forall\ j$), to get the partial effect of the dummy variables.

4.3. Total Derivatives of Utility with respect to Two Features

For the logit model, or the variables without random coefficients of the mixed logit models, the total derivative with respect to two continuous variables should be

$$dU_{nj} = \beta_k dx^k_{nj} + \beta_l dx^l_{nj} \ \forall \ n, \ j$$

To maintain at least the same utility level of each individual n by choosing j, the change of variable x^l_{nj} would have a $-\frac{\beta_l}{\beta_k}$ effect on the change of the variable x^k_{nj}.

If the variable x^l_{nj} is a dummy variable, then the change of the value of that from 0 to 1 will cause $-\frac{\beta_l}{\beta_k}$ change of the variable x^k_{nj} to maintain at the same utility level.

In this research, it is particularly interested to see how the change of the *hukou* reform policies would generally affect the other variables, such as the average disposable income of the region, to let everybody at least maintain the same utility level as before.

4.4. The Estimation of Population for Each Choice Location, and the Counterfactual Effects of Hukou Reform

The decision to proceed with *hukou* reform trials in China prior to 2014 was generally exogenously made by the central government and local governments. During this period, potential migrant workers were facing the decision to choose from each of the locations , and may or may not have implemented *hukou* reform policies. Generally, a region (prefectural-level city) that implemented the *hukou* reform policy would allow the *de facto* but not *de jure* residents (who hold *hukou* from other cities) to update their *hukou* registration location to the region where they were current living. In the rural areas within a region that implemented *hukou* reform policies, changing their *hukou* from a rural one to an post-reform *hukou* type of the *hukou*-reform-implemented urban region would still allow the rural migrants to keep some of their entitled rights in the rural *hukou*, such as rural land ownership from Contracted Management Right.

With the estimation of the coefficients of the logit/mixed logit models, the probability P_{ni} is given $\forall \ n, \ i$. I would like to check the counterfactual effects of *hukou*-reform implementation on different types of the regions. Inspired by the 2014 *Hukou* Reform Opinion, I divide the location choice sets J into three groups:

* Mega cities, super-large cities, large cities
* Mediums cities, small cities and townships
* Rural villages

The counterfactual *hukou* reform variable would be the reform dummy for each choice. There will be eight groups of counterfactual effects ($2^3 = 8$) to check the different migrants inflows to different types of places in China. We can compare and check how the different implementations of *hukou* policies would change the population distribution in each type of city in China.

4.5. Adjusting Weights of Each Observation

With the estimated probability of each person on each choice, we can simply introduce the associated weights of each observation, and do the summation over the weighted probability of each person at each choice i, which is

$$\hat{P}_i = \frac{\sum_n w_n P_{ni}}{N = 11783}$$

to get the estimated population proportion of each of the choices, $\hat{P}_i$. However, as introduced in the data section, there is a problem with the way the stratified sampling calculates w_n for each observation. Fortunately, since $w_n = \frac{City_{pop}}{Village_{pop}} \frac{Village_{hhs}}{HH}$, and the only inaccurate population is $City_{pop}$. Given that each of the samples are randomly selected, we can directly adjust the exogenous sample weights, which is w_n.

We can see w_n composed of two parts, the accurate population part and the inaccurate population part. Given that each n must belongs to a location choice j, then we can write

$$w_{n(n\ at\ j)} = accurate\_part * H_j$$

where H_j is the proportion of the population of location choice j is calculated in w_n. However, the true population proportion of location j should be Q_j. Then the corrected weights for each individual n at its choice location j should be

$$\begin{aligned} \tilde{w_{n_j}} &= accurate\_part * Q_j \\ &= accurate\_part * H_j * \frac{Q_j}{H_j} \\ &= w_{n(n\ at\ j)} * \frac{Q_j}{H_j} \end{aligned}$$

This method is inspired by the adjusted weight in the weighted exogenous sample maximum likelihood (WESML) method (Mabit, 2008; Mcfadden, 1999). But we do not need to do the weighted maximum likelihood in this study, since the sampling method is exogenous stratified sampling from CGSS 2010 survey. Correcting the wrongly calculated weights due to inaccurate city population data would give accurate estimation results, as we can see in the following section.

To get Q_j, we need to apply the accurate census data, which distinguishes each type of the place in China by connecting the village-unit level administrative hierarchy of China. I am using the dataset CN-SAUCS to get this accurate $Q_j\ \forall\ j$, considering the type of place of each place j in China.

Moreover, we need to adjust the average partial effect by multiplying the adjusted individual weights for each individual

$$APE = \frac{1}{N = 11783} \sum_n \frac{\partial P_{ni}}{\partial x_{ni}} \tilde{w_{n_j}}$$

5. Results and Discussion

5.1. The Random Utility Models with the Data

The RUM for the standard MNL model has the following realization from the data

$$\begin{aligned} U_{nj}^{MNL} = {} & \beta^1 \log rain + \beta^2 temperature + \beta^3 d_{location\_reform} + \beta^4 d_{hukou\_at\_same\_city} \\ & + \beta^5 \log noise\_pollution + \beta^6 \log neighbor\_relationship + \beta^7 \log neighbor\_helpfulness \\ & + \beta^8 \log distance\_to\_hukou + \beta^9 \log average\_disposable\_income + \varepsilon_{nj} \end{aligned}$$

The RUM for the mixed logit model has the following realization from the data

$$\begin{aligned} U_{nj}^{Mixed} = {} & \beta_n^1 \log rain + \beta_n^2 temperature + \beta^3 d_{location\_reform} + \beta^4 d_{hukou\_at\_same\_city} \\ & + \beta^5 \log noise\_pollution + \beta^6 \log neighbor\_relationship + \beta^7 \log neighbor\_helpfulness \\ & + \beta^8 \log distance\_to\_hukou + \beta_n^9 \log average\_disposable\_income + \varepsilon_{nj} \end{aligned}$$

where $\varepsilon_{nj} \sim iid\ type\ I\ extreme\ value$, and

$$\begin{pmatrix} \beta_n^1 \\ \beta_n^2 \\ \beta_n^9 \end{pmatrix} = \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix} + \begin{pmatrix} \varepsilon_{\beta^1} \\ \varepsilon_{\beta^2} \\ \varepsilon_{\beta^9} \end{pmatrix} = \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix} + L \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} = \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix} + \begin{bmatrix} s_{11} & & \\ s_{21} & s_{22} & \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix}$$

, $\eta \sim i.i.d\ N(0,1)\ \forall\ \eta$ in the estimation. We could have the expectation of $\begin{pmatrix} \beta_n^1 \\ \beta_n^2 \\ \beta_n^9 \end{pmatrix} = \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix}$, since

$$E \begin{pmatrix} \beta_n^1 \\ \beta_n^2 \\ \beta_n^9 \end{pmatrix} = E \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix} + E \left(L \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} \right) = \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^9 \end{pmatrix}$$

A Choleski factor of the variance matrix of the random coefficient $\begin{pmatrix} \beta_n^1 \\ \beta_n^2 \\ \beta_n^9 \end{pmatrix}$ is defined as a lower-triangular matrix L such that

$$\begin{aligned} var \left(\begin{pmatrix} \beta_n^1 \\ \beta_n^2 \\ \beta_n^9 \end{pmatrix} \right) &= E(L \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} \left(L \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} \right)') \\ &= LE \left(\begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} \begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix}' \right) L' \\ &= LVar \left(\begin{pmatrix} \eta^1 \\ \eta^2 \\ \eta^3 \end{pmatrix} \right) L' \\ &= LIL' = LL' = \Omega \end{aligned}$$

We should have the standard deviation of the estimated random coefficients calculated as $\sigma_{\beta_n^1} = \sqrt{s_{11}^2} = |s_{11}|$, $\sigma_{\beta_n^2} = \sqrt{s_{21}^2 + s_{22}^2}$, and $\sigma_{\beta_n^9} = \sqrt{s_{31}^2 + s_{32}^2 + s_{33}^2}$, in the report.[5]

[5]The reported value of each s could be either positive or negative, since the estimation process from the codings are unconstrained.

The estimation results from control function are not significant, that is, there is no significant endogeneity in the distance between *hukou* location and destination with the unobserved utility of each person. Results are attached to the appendix in this paper.

In Table 2, the results with the comparison between the standard logit model and the mixed logit models are shown. . The first two mixed variables are rainfall and temperature of a location, which capture the meteorological features of that location. The other mixed variable is average disposable income of a location j.

To compare the two models, first let us focus on the signs and values of the coefficients. Except for rainfall and temperature, the other coefficients of the two models have similar values and same signs. While the absolute value for the coefficients of rainfall and temperature are close to each other, the signs of the two parameters are opposite. We can see that the change in each migrant' s preference over location varies a lot, which could even change the sign of the coefficients at the mean value. That means a migrant' s taste in the particular longitude and latitude (which implies temperature and rainfall level) of a location could be very different from another migrant: some people prefer northern and eastern places, but some people prefer southern and western places. The results in the mixed logit model indicate that, on average, when including the variation in personal taste, migrants in China prefer places with more rainfall and cooler temperature.

Next, we review the performance of the standard MNL models with mixed logit models. First, the mixed logit model proves to be a better fit, with slightly higher performance. Second, after adjusting the probability by personally correcting weights (see section 3.5), we can get the estimated population distribution from the two models in Table 3. The mixed logit gives a better estimation, again.

We will interpret the results with this mixed logit model.

5.2. Partial Effects from the Estimation

As shown in the last section, we can calculate the average partial effect of each location feature x_{nj} with respect to this location j, over each individual n.

Considering that we have 469 choices for each individual, the probability for each choice chosen by n is relatively small, and the partial effect, which contains the quadratic form of the probability, could be even smaller. The following chart (Figure 5.1) shows the basic statistics of the average partial effects of each location choice j.

Moreover, we could focus on the spatial distribution of the partial effect, such as for what kind of the choices j are having a positive partial effect of a specific x_{nj}. In Figure 5.2, I grouped the variables whose partial effects on location choices j have more percentiles over the negative values.

(1) *Hukou* Reform: Seen from the result, *hukou* reform makes the implemented location more attractive in most part of China. Except for western provinces and several sub-regions in the middle zones. The estimation of our data also shows that *hukou* reform could not make places such as the CBD zones of Beijing, Tianjin and Shanghai, or rural regions more attractive.

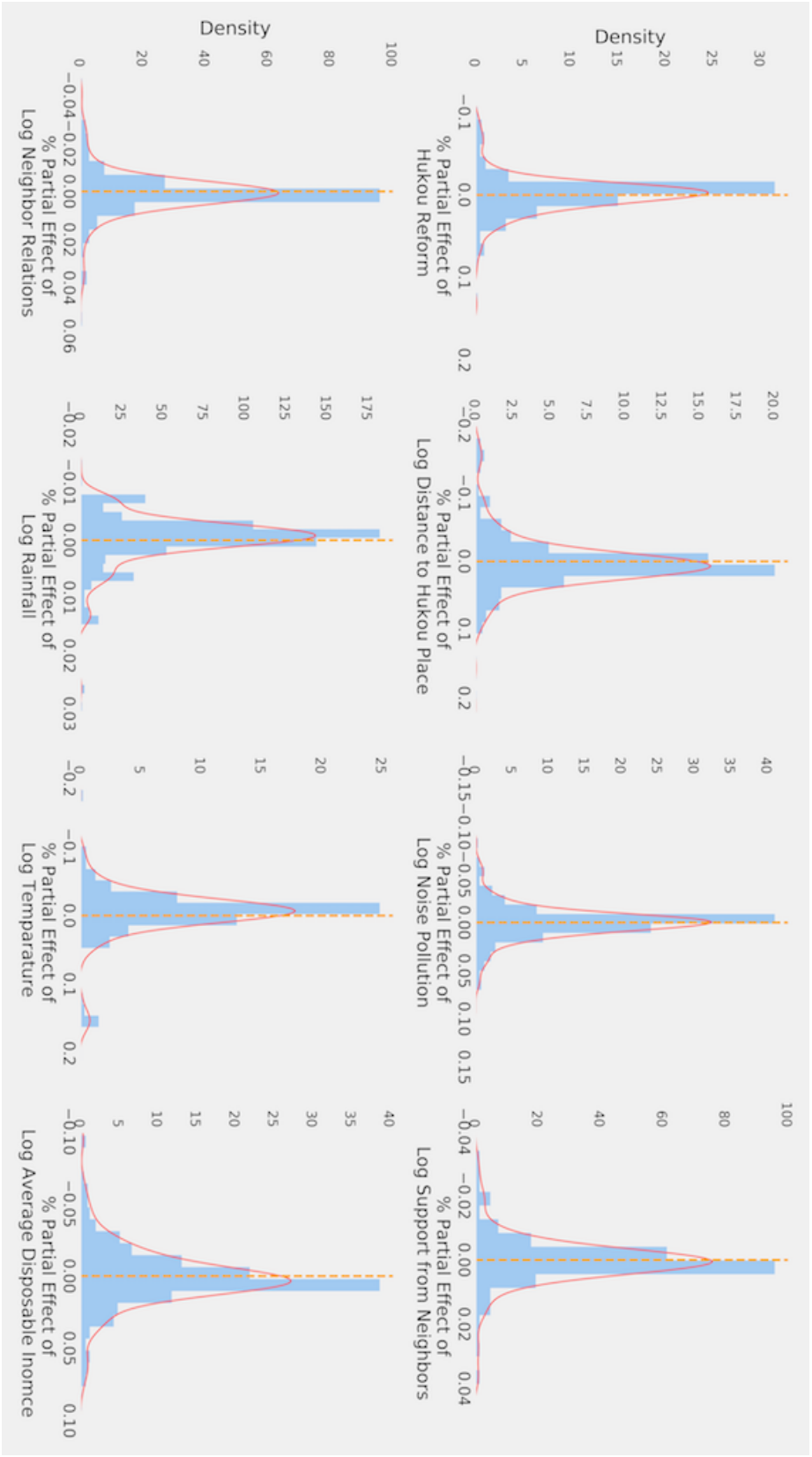


FIGURE 5.1. PARTIAL EFFECTS OF EACH VARIABLE FOR EACH J

Table 2—MNL/Mixed Logit Results for Location Choices in China

For Each Location j	MNL Parameter	MNL t	Mixed Parameter	Mixed t
Log yearly average rainfall	-0.09	-2.04	0.08	1.64
Yearly average temperature	0.03	2.37	-0.03	-1.93
Hukou (trial) reform dummy(implemented=1)	0.41	4.68	0.40	4.52
Hukou at the same prefectural-level city dummy (same=1)	1.17	12.21	1.37	15.11
Log noise pollution	-0.09	-1.48	-0.07	-1.11
Log neighbor relations	-0.49	-4.17	-0.48	-3.75
Log Support from Neighbors	0.38	2.64	0.41	2.62
Log distance (km) to *hukou* location	-2.08	-96.57	-2.20	-94.20
Log average disposable income	1.17	21.82	1.31	6.17
s_{11}	-	-	0.16	3.20
s_{21}	-	-	0.21	14.97
s_{22}	-	-	0.16	0.74
s_{31}	-	-	0.04	0.71
s_{32}	-	-	-0.81	-2.35
s_{33}	-	-	-0.23	-0.18
Number of cases	11783		11783	
Number of choices	469		469	
Log likelihood function	-10815.29		-10796.23	
Log likelihood function at null	-72472.55		-72472.55	
Goodness of Fit 1-LL/LL_null	0.85077		0.85103	

One can think that comparatively, the majority of the population would prefer to live in a well-urbanized, but not super crowded city, if the urban zones become more welcome to the new comers.

(2) Rainfall: The partial effects of rainfall fits well with the annual precipitation map of China, except in some choice location zones that are seen as townships or rural regions, which distorts the matching.[6]

(3) Temperature: Similarly, given a temperature map, we can observe that in some hot regions such as Guangdong, Guangxi, Hubei, Hunan, Zhejiang and Sichuan Provinces, people hold a negative preference towards most of the locations there due to hot temperature in the summer. Conversely, people generally do not favor the cold temperature in some places within provinces such as Hebei and Heilongjiang. [7]

While the mean value of this random coefficient of the mixed logit is negative, with personal tastes varying across regions, we would like to get a sense of whether people prefer colder

[6]Please visit http://www.chinamaps.org/china/china-map-of-precipitation-annual.html to check one example map

[7]An example map could be reached herehttp://www.chinamaps.org/china/china-temperature-map-annual.html

Table 3—Performance of model estimates with true *de facto* population distribution

Type of choices	True distribution %	MNL estimation %	Mixed estimation %
Mega City	3.2498	3.2599	3.2785
Super-large/large City	14.3234	14.1096	14.1010
Medium/small City	12.4625	12.2149	12.1376
Townships	20.1992	21.0806	21.0812
Rural villages	49.7651	49.3350	49.4016
Distance	-	0.0001068	0.0000013

regions or warmer regions that do not experience extreme heat. The results in the standard logit model show that, generally, there is greater probability of someone choosing a warmer place for their living location j. Since the general effect of temperature is positive, we can conclude that, in general, people prefer warmer places over colder places.

(4) Noise: People generally prefer less noise, but in some rural zones or less developed urban zones, the increased level of noise may represent development in that zone, which people may be willing to tolerate, so partial effect of noise on those regions is positive.

In Figure 5.3, I grouped the variables whose partial effects on location choices j have more percentiles over the positive values:

(1) Neighbors: Generally, the two neighbor related variables reflect the opposite direction of urbanization level, especially the data regarding the support from neighbors. The lower the level of support from neighbors, the more urbanized areas they are. Neighbor relations are improved in the southern, middle and northeastern (Dongbei; 东北; dōngběi; historically known as Manchuria) part of China.

(2) Distance to *Hukou* Location: For the most part, people prefer to live in a location close to their *hukou* location. But in provinces such as Neimenggu, Jilin, Heilongjiang, Guizhou, Hebei and Henan, we can see that residents could have come from further origins than other provinces.

(3) Average Disposable Income: Most residents in urbanized areas prefer richer destinations. This factor is not as influential for residents from rural regions.

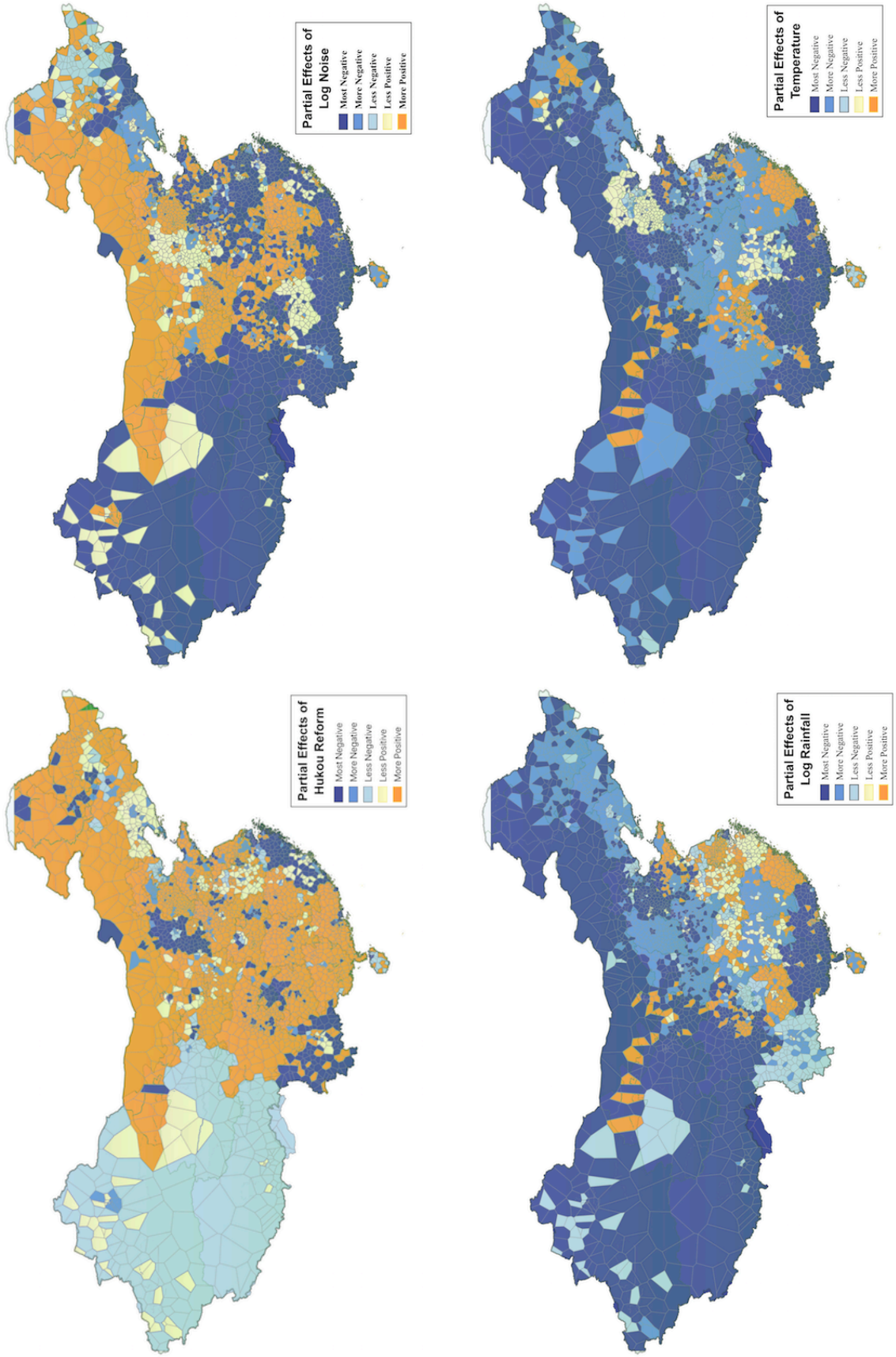
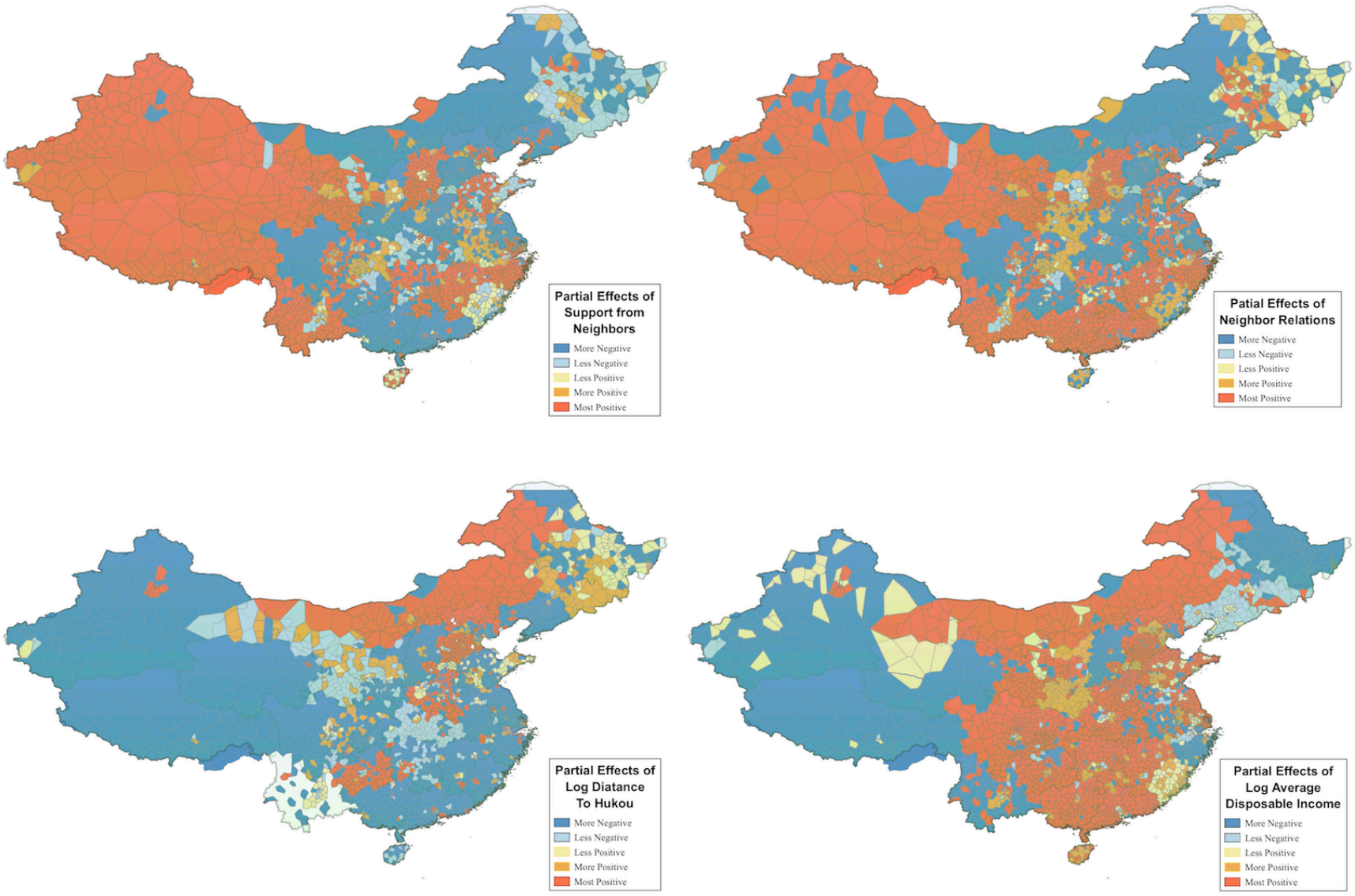


FIGURE 5.2. PARTIAL EFFECTS OF VARIABLES ON THEIR LOCATIONS (1)

Figure 5.3. Partial Effects of Variables on Their Locations (2)



5.3. Effects of Trial-Period Hukou Reform on/with Other Variables

Since *hukou* reform in my model is a dummy variable, to maintain at least the same utility level for each person n at location j after the implementation of *hukou* reform, the change in another variable would be

$$-\frac{\beta^{hk_{reform}}}{\beta_n^k} = \frac{\triangle x_{nj}^k}{1-0} = \triangle x_{nj}^k$$

where β_n^k could be a constant or the random coefficients, depending on the specific variable x_{nj}^k.

The numerator is always the estimated coefficient of the *hukou* reform dummy, which is 0.40375 in the mixed logit model.

First, with the implementation of the *hukou* reform, the percentage change of the noise pollution level was increased by approximately $^{-0.40375}/_{-0.07433} = 5.43$ percentage. Also, the neighbor relations would be worse off by $|^{-0.40375}/_{0.41354}| = 0.98$ percentage. However, the reformed region would experience a more helpful neighbor environment, since the support from neighbors in those regions is increased by $-0.40375/-0.48047 = 0.84$ percent.

Next, we assess the distance of the migrant's *hukou* location to the current living location. The effect is that a reformed region would increase that distance (in kilometers) by $-0.40375/-2.20214 = 0.18$ percentage. The interpretation for this phenomenon should be that a region which has implemented *hukou* reform would be more attractive than it had been previously to migrants from further regions.

Moreover, consider the other dummy variable: holding *hukou* in one' s current location. If both dummy variables are ones, it means that the resident has updated/held his/her *hukou* to their current location, which has implemented *hukou*-reform. Therefore, the change of dummies from zeros to ones (holding other variables constant) would be the change of the utility of each person, which are just the summation of coefficients $0.40375 + 1.37487 = 1.78$ unit of increase. With the reformed *hukou* policy and the ability to update *hukou* location to their current living location, on average, a migrant' s utility is increased.

Last but not least, we focus on the change of average disposable income of a region with *hukou* reform policy implemented. Given that disposable income has a random coefficient effect, the estimations from the MNL model and Mixed model are not that different on the constant part of the coefficient. I would interpret the effect by the mean value of the coefficient of disposable income. On average, a reformed region would have a percentage change of $^{-0.40375}/_{1.3068} = -0.31$ in the regional average disposable income. This means that in regions where *hukou* reform policy has been implemented the average disposable income will decrease. This is in line with the interpretation of previous results. Regions with *hukou*-reform will become more attractive to migrants. With migrant inflows to newly opened-up regions, these places will become poorer than they had been, previously. However, this does not mean that opening-up policies are making people worse off. The large proportion of newcomers could have been doing basic migrant worker jobs, such as blue-collar jobs, as mentioned in the migrant reports in NBS National Bureau of Statistics (2018)); these roles would earn much lower wages than the average wages/income of urban hired workers.[8]

[8] In 2017, the national average wage for rural migrant workers is 3485 RMB, but the one for the urban hired workers is 6193 RMB, according to NBS reports in 2018. http://www.stats.gov.cn/tjsj/zxfb/201805/t20180515_1599424.html

5.4. *Counterfactual Effects*

Inspired by the 2014 *hukou* reform policy: larger cities (with *de facto* population more than 1 million in the core city part) are more restricted for migrants to register *hukou*, while medium/small sized cities are more open for outsiders to register their *hukou*. I divide all types of the places in China into three groups for the counterfactual effects:

* Mega-, super-large-, large-size cities;
* Medium-, small- size cities and townships
* Rural villages

The aim of *hukou* reform policy is to accelerate urbanization progress, while controlling the population in large cities. In other words, the goal is to find a specific *hukou* reform policy that could reduce the rural population, increase the population in smaller cities, but decrease or maintain the population size in large cities.

To test different possible *hukou* reform policies based on out estimation results, I update the dummy variables of *hukou* reform implementation for each type of j location choice. There are eight cases of counterfactual situations in total, which are listed in Table 4 with the values of the *hukou* reform dummy variables for each case.

Table 4—Dummy Values of Each of the Counterfactual Situations

	Mega-, super-large-, large-size cities	Medium-, small-size cities and townships	Rural Villages
CF1	0	1	0
CF2	0	1	1
CF3	1	1	1
CF4	1	1	0
CF5	1	0	0
CF6	0	0	1
CF7	0	0	0
CF8	1	0	1

The results are shown in the following charts. In Figure 5.4, I grouped the policies which are unlikely to be implemented in practice. In the conditional logic model, if the values of the variables all change in the opposite direction, the change in the probability distributions will be the same. That is, if every place implements the *hukou* reform, or every place does not, the change of the probabilities will look the same, as shown in Figure 5.4. However, given the fact that *hukou* reform is on-going, it is reasonable to respond to both of those scenarios with the same question: how would population distribution change under a policy where every place is reformed?

The basic conclusion shown from this group is that none of the policies could achieve the aim of the 2014 *hukou* reform Opinions. If every place is opened up, mega cities and larger cities would decrease their population, while rural regions would also experience population increases. But the absolute change with respect to each group is very small. If only rural areas are holding the reform, meaning people can move freely to rural places, but not the other direction, China would undergo a de-urbanization. However, only opening up the doors of larger cities, would not push the population to go to smaller urban zones, which are medium-, small- sized cities and townships.

In Figure 5.5, we can observe a group of practical policies. Surprisingly, we find more than one policy that fits our expectations. The policy fit the logic of 2014 Opinions, decreasing the population of mega, super-, large- sized cities, and increasing the medium-, small- sized cities, thus accelerating urbanization progress. More than that, the policy '1 1 0', which means opening up all urban areas while holding the old *hukou* policy in the rural regions, would also accelerate urbanization progress. The difference between the effects of '0 1 0' and '1 1 0' are the following. First, the latter policy may motivate more rural migrants to live in urban areas than the former. However, the population of the super-large-, large-sized cities would increase, rather than decrease. The policy maker needs to establish a desired population distribution in order to make an effective corresponding *hukou* reform policy. If they prefer to accelerate urbanization progress rather than control city congestion, the government should choose '1 1 0'.

How is this policy different from '1 1 1', and what impact will it have if rural areas do not implement a *hukou* reform policy? Let us first look at the remaining two counterfactual effects: '0 1 1', and '1 0 1', with rural regions 'opened up'. In both cases, the rural population will increase. Rural migrants represent the majority of the 'floating population' in China. They may stay in the cities, or move back to their hometown under certain *hukou* policies. The opening-up policies of rural regions should be interpreted differently, rather than by its literal meaning. As introduced before, in the *hukou* reform trial, or even in the present on-going *hukou* reform, opening up rural regions helps rural migrants to maintain rights, while allowing them to move to opening-up urban regions within the cities, to enjoy the same level of social services as other citizens. In reality, *hukou* reform in rural regions cannot be held without the reform of any city regions. Therefore, '0 0 1' counterfactual could not realistically exist. Since rural migrants are worried about losing their rights that are attached to their rural *hukou*, as well as about the instability of living in cities, they might come back. '0' reform in the rural region could be seen as a commitment from the government: rural migrants could update their *hukou* to an post-reform location without losing the rights attached to their original rural *hukou*.

To sum up, unless they are the only regions to hold *hukou* reform in China, the population of mega cities will always decrease. To accelerate urbanization progress, the government should always promise that rural migrants will be able to keep the original rights attached to their rural *hukou*, and open up as many urban regions to them as they can. However, if the goal of the government is to control city congestion, they should follow the 2014 Opinions strategy, while still allowing rural migrants to keep their previous rights from their before-reformed rural *hukou*. In these two cases, the *hukou* reform would typically accelerate urbanization progress in mainland China.

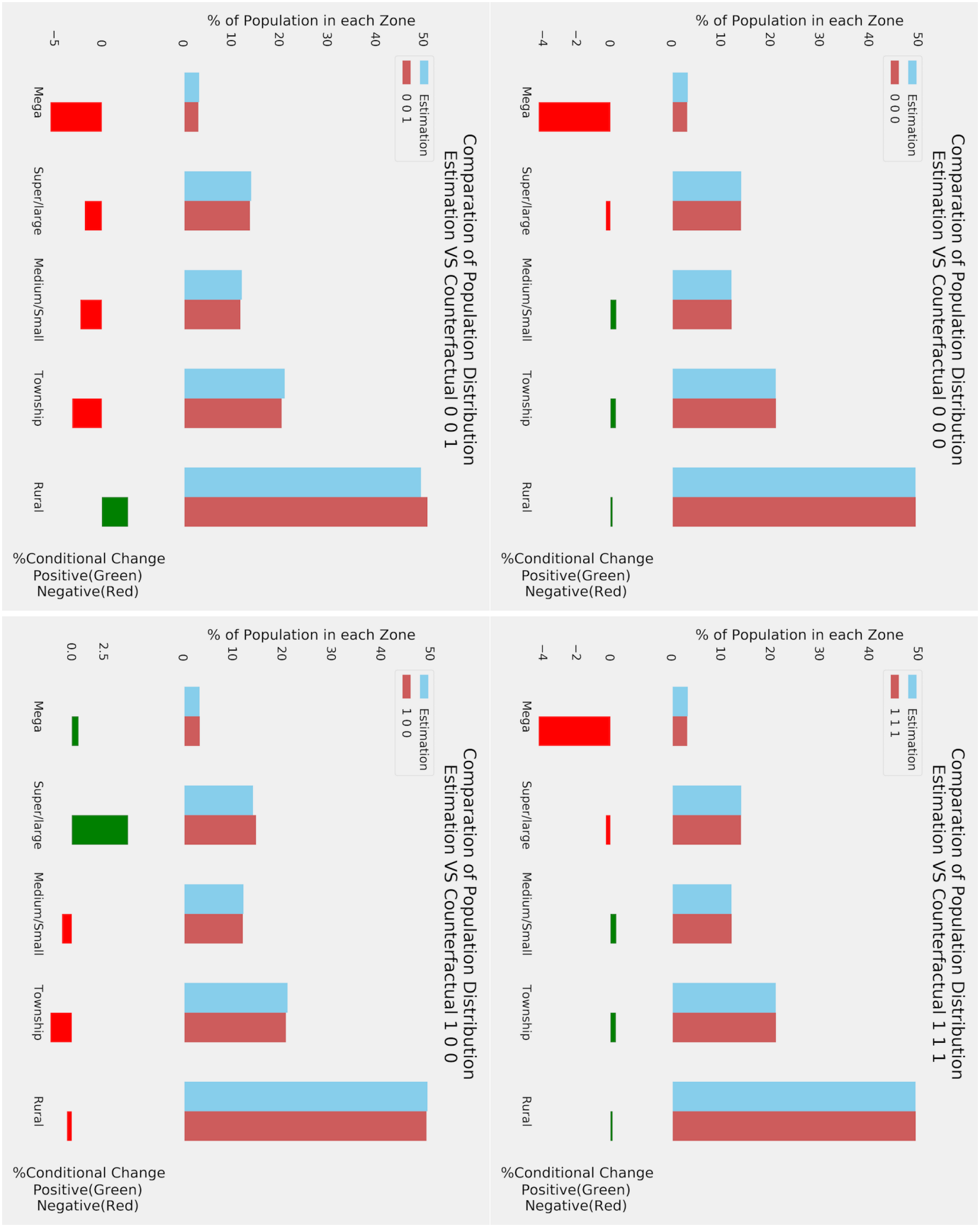
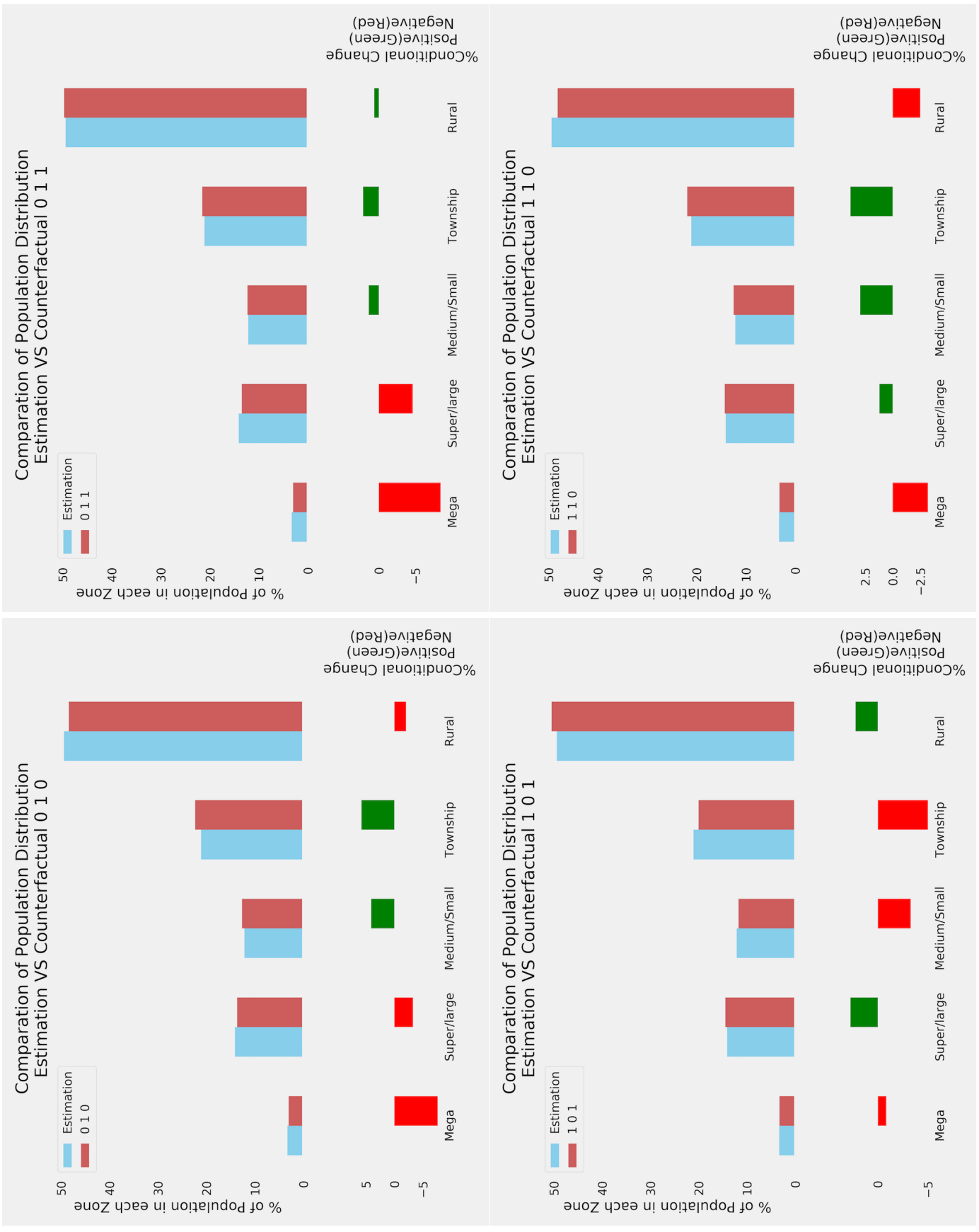


FIGURE 5.4. COMPARISON OF THE RESULTS OF THECOUNTERFACTUAL EFFECTS

FIGURE 5.5. COMPARISON OF THE RESULTS OF THECOUNTERFACTUAL EFFECTS



6. Conclusion

This research studies the qualitative and quantitative effects of the *hukou* reform on the population distribution of different places of China with a large dataset discrete choice model approach by mainly applying CGSS survey data and CN-SAUCS data. The main conclusion demonstrates that *hukou* reform could accelerate urbanization progress, as long as the main rights attached to the before-reform *hukou* remain for rural migrants, and the doors of cities are open to migration inflows. In general, migrants prefer richer places that are not geographically far from their *hukou* origins, with more helpful neighbors. However, at the beginning of reform implementation, the regions that open up would experience a decrease in the average disposable income due to new migration inflows. But the reform also makes the cities more attractive to migrants from further places than before. By applying the above outlined strategies of *hukou* reform policies, the population of the mega cities will decrease, while the population in medium- or small-sized cities will always increase. The population distribution of the urban hierarchy can also be changed under *hukou* reform policy.

The two main difficulties and challenges of this study are the large choice data set and the definition of the sections of migration location choices. That is because the heteroskedasticity in entitlements or public services depends on the social services provided by local governments of different places or different urban areas. With the detailed information of the village-level units from the CN-SAUCS, I partitioned China into 469 zones based on their administrative boundaries, urban/rural designation, spatial relationship, and city hierarchy properties.

To get the population distributions with different types of *hukou* reform policies, I applied the standard logit models, mixed logit models and mixed logit with control function approaches. The mixed logit model is chosen as the best fit for its significant estimations. The partial effects and counterfactual effects are calculated based on the results from the mixed logit model. The biased weights of each choice maker, due to the inaccurate population data, were corrected by applying the CN-SAUCS dataset.

There are still some limitations, and the potential to improve upon this research. We would expect to use a panel data or the survey data with the lifetime migration records of respondents in village-level unit details. While it is very challenging to obtain this kind of data, even with the records of the updates of one' s *hukou* types, it might be possible to have the sampled raw data from the census, directly. But the problem will still exist for this census data, since it will have different definitions of administrative boundaries and urban/rural designation for the urbanization progress.

A possible extension from this study would be the analysis of each resident' s average disposable income in *hukou* reformed regions. From this study, we can observe that the average disposable income decreases with the new migration inflows to the *hukou* reformed region, but it would be interesting assess whether or not the wage of rural migrant workers would converge (Cai and Du, 2011).). It would be essential for study further in order to answer if the decrease of the disposable income would still remain in the long run within the reformed region, or if the wage of the rural migrant workers will increase and converge, and under what circumstances the converge would happen.

Moreover, the research warrants further investigation with the application of CN-SAUCS data and other survey data to estimate the relevant migration effect and population distribution

under different policies. The policy makers or other institutions could follow the implications from such research to make their plan for their unique goals.

Finally, extensions of this research could apply other models or methods. This work is the first to study *hukou* reform effects on urbanization progress from both qualitative and quantitative aspects with big and accurate comprehensive spatial census data of China. The results from the standard logit models also show the accuracy, which may highlight the importance of the big and accurate data more than the model itself. However, there is still potential to try different methods with the same data sets to perfect the results. In future research, one could apply a spatial discrete choice model to study the heteroskedasticity induced by spatial dependence (Anselin et al., 2010), or use machine learning approaches to redo the population estimations under different *hukou* policies in China.

References

Anselin, Luc et al. (2010). *Advances in spatial econometrics*. Vol. 40. 5, pp. 253–254. ISBN: 978-3-642-07838-5. DOI: 10.1016/j.regsciurbeco.2010.06.002.

Baidu Maps (2018). *China Urban Research Report 2017 Q4*. Tech. rep.

Beijing Municipal People's Government (2016). *Opinions of the Beijing Municipal People's Government on Further Reform of the Hukou System*. URL: http://zhengce.beijing.gov.cn/library/192/33/50/438650/80771/index.html.

Cai, Fang and Yang Du (2011). "Wage increases, wage convergence, and the Lewis turning point in China". In: *China Economic Review* 22.4, pp. 601–610.

Chan, Kam Wing (2010). "A China Paradox: Migrant Labor Shortage amidst Rural Labor Supply Abundance". In: *Eurasian Geography and Economics* 51.4, pp. 513–530. ISSN: 1538-7216. DOI: 10.2747/1539-7216.51.4.513. URL: http://www.tandfonline.com/doi/abs/10.2747/1539-7216.51.4.513.

– (2013). "China: internal migration". In: *The encyclopedia of global human migration* 2, pp. 980–995. DOI: 10.1002/9781444351071.wbeghm124. URL: http://faculty.washington.edu/kwchan/Chan-migration.pdf.

Chen, Juan, Deborah S Davis, and Pierre F Landry (2017). "Beyond Hukou Reform: Enhancing Human-Centered Urbanization in China Paulson Policy Memorandum". In: February. URL: http://paulsoninstitute.org.cn/wp-content/uploads/2017/02/PPM{_}Beyond-Hukou{_}Chen{_}Davis{_}Landry{_}English.pdf.

Cui, Yuming, Jingjing Meng, and Changrong Lu (2018). "Recent developments in China's labor market: Labor shortage, rising wages and their implications". In: *Review of Development Economics* 22.3, pp. 1217–1238. ISSN: 13636669. DOI: 10.1111/rode.12391. URL: http://doi.wiley.com/10.1111/rode.12391.

Das, Mitali and Papa M N'Diaye (2013). "Chronicle of a Decline Foretold: Has China Reached the Lewis Turning Point?" In: *IMF Working Papers* 13.26, p. 21. ISSN: 1018-5941. DOI: 10.5089/9781475548242.001. URL: http://elibrary.imf.org/view/IMF001/20245-9781475548242/20245-9781475548242/20245-9781475548242.xml.

Friedman, Eli and Sarosh Kuruvilla (2015). "Experimentation and decentralization in China's labor relations". In: *Human Relations* 68.2, pp. 181–195. ISSN: 1741282X. DOI: 10.1177/0018726714552087.

Independent (2012). *Huaxi: The socialist village where everyone is wealthy*. URL: https://www.independent.co.uk/news/world/asia/huaxi-the-socialist-village-where-everyone-is-wealthy-6290583.html.

Liu, Yijiao (2018). "Distinguishing Places and Population in Chinarefered as mainland China in this paper: A Comprehensive Geo-coded Dataset of Census and Administrative Hierarchy".

Mabit, Stefan L (2008). *Sample selection and taste correlation in discrete choice transport modelling*. January. ISBN: 9788773271803.

Magazzini, Laura (2014). *Multiple Choice Models*. Tech. rep., pp. 1–72.

Mcfadden, Daniel (1999). "Chapter 2. sampling and selection 1." In: *Lecture notes of Economics 240B, Second Half*, pp. 1–21.

National People's Congress Standing Committee (1958). *Zhonghua renmin gongheguo zhuxi ling [Regulations of the People's Republic of China on Residence Registration]*. Decree of the Preseident of the People's Republic of China.

NBS National Bureau of Statistics (2017). *2017 China Statistical Yearbook*. Ed. by Dong Guo. Zhongguo Tongji Chubanshe.

– (2018). *The Monitoring Survey of Rural Migrants in China in 2017*. URL: http://www.stats.gov.cn/tjsj/zxfb/201804/t20180427{_}1596389.html.

People.cn (2013). *Counter-Urbanization of Some Regions: People Are in Facor of Rural Hukou*. URL: http://politics.people.com.cn/n/2013/0924/c1001-23009276.html.

Song, Yang (2014). *The Analysis of the Difficulties Strategires of Hukou Reform under the New Situation of Urbanization[Xinxing Chengzhenhua Beijingxia Huji Gaigede Nandianyu Silu Fenxi]*. Tech. rep. National Academy of Development and Strategy, RUC. URL: http://nads.ruc.edu.cn/upfile/file/20141219110233{_}24188.pdf.

Sun, Wenkai (2017). *The Analysis of the Motivation and the Strategies of Hukou Reform(Huji Zhidu Gaige Dongyin Fenxi Ji Jinyibu Gaige Duice)*.

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. Vol. 47, p. 384. ISBN: 9780521747387. DOI: 10.1016/S0898-1221(04)90100-9. URL: http://eml.berkeley.edu/books/choice2.html.

Robustness Checks

A1. Models of Selection–from Standard MNL Models

Many regressions have been done to set up the base-line model (the standard conditional multinomial models) to get the one that applied with the group of variables constructed in this study. Table A1 shows the selective models that tested before. The first MNL included every variable constructed for this study, however, many of the self-reported variables did not provide significant results. Based on the significance levels of the variables in the first MNL model, I excluded the variables with t statistics less than 1 for model 2. Even still, the performance of the noise pollution and safety level variables were not improved that much. Based on model 2, I constructed model 3 (without noise pollution) and the model (without safety level) I applied for this paper. The performance on single variable improved. Also, from the hypothesis tests, model 3 (compared with model 2) had a p value 0.26 from the χ^2_1 distribution, and the model we applied for this study has a $p = 0.27$ from the χ^2_1 distribution. We fail to reject the null hypothesis that the excluded variables in each model is zero.

The model with the noise pollution variable instead of community safety level was chosen because, from the perspective of estimation, the two models behave similarly. But when considered in the real life, I would prefer the noise pollution variable in the observable utility.

Also, the following regression table A1 provides the robustness checks for the standard logit model applied in the paper. It shows that the final estimations included in the study are not affected that much by the presence or absence of the self-reported variables. After running the mixed logit model of the MNL (3), the average housing area variable is not significant enough to be included in the model. That said, in the main structure, I did not include this variable under observed utilities.

A2. Control Function for Endogeneity

In the main content of this paper, I showed the mixed model applied and the corresponding standard logit model for this study. The reason that I did not put the endogeneity test in the main structure is because the results are not significant. Here in the appendix, I will show the results from the control function approach.

The first step is to choose the potential endogenous variables. In the case of the *hukou* analysis, the endogeneity is likely in the decision related variables of one's *hukou* location and the living residential location choices. In the observable variables, the distance between one's *hukou* location and location alternative js might be correlated with the unobserved variables. For example, the distance between one's *hukou* location and location alternative js might not only indicate about how far those two places could be, but also might include the information such as the administrative boundaries between the two locations. Since a person's *hukou* location will mainly decide the level of public services one could receive generally, we should check if the unobservable utilities at location js would affect people's *hukou choice*. If not, we could basically assume that people's utilities at location j is not going to decide people's *hukou choice*, instead, *hukou status* will affect the possible location choices js a person would consider.

Based on methods from Chapter 13 of Train's book(2009), and also based on the mixed logit model I applied in this study, the control function approach has the following structure with the RUM

(A1)
$$U_{nj}^{Mixe\_control} = \beta_n^1 \log rain + \beta_n^2 temprature + \beta^3 d_{location\_reform} + \beta^4 d_{hukou\_at\_same\_city} + \beta^5 \log noise\_pollution + \beta^6 \log neighbor\_relationship + \beta^7 \log neighbor\_helpfulness + \beta^8 \log distance\_to\_hukou + \beta_n^9 \log average\_disposable\_income + \varepsilon_{nj}^1 + \varepsilon_{nj}^2$$

where $\log distance\_to\_hukou$ is considered to be possibly correlated with the first unobservable part of the utility ε_{nj}^1. The instrument variable of $\log distance\_to\_hukou$ is the explanatory variable $d_{location\_reform}$. Assumed that they have the simple linear relationship

(A2)
$$\log distance\_to\_hukou = \gamma d_{location\_reform} + \mu_{nj}$$

and that the control function is the mean of the unobserved utility ε_{nj}^1, which follows

$$\begin{aligned} \varepsilon_{nj}^1 &= \lambda \mu_{nj} + \tilde{\varepsilon_{nj}^1} \\ &= \lambda \mu_{nj} + \sigma_{\varepsilon^1} \eta_{nj} \end{aligned}$$

where $\eta_{nj} \sim iidN(0,1)$. Since I am using the explanatory variable of RUM to be the instrument

(A3)
$$\begin{aligned} \beta^3 d_{location\_reform_{nj}} + \beta^8 \log distance\_to\_hukou_{nj} + \varepsilon_{nj}^1 = \\ \beta^3 d_{location\_reform_{nj}} + \beta^8 \left(\gamma d_{location\_reform_{nj}} + \mu_{nj}\right) + \lambda \mu_{nj} + \sigma_{\varepsilon^1} \eta_{nj} = \\ \left(\beta^3 + \beta^8 \gamma\right) d_{location\_reform_{nj}} + \left(\beta^8 + \lambda\right) \mu_{nj} + \sigma_{\varepsilon^1} \eta_{nj} = \\ \left(\frac{\beta^3}{\gamma} + \beta^8\right) \gamma d_{location\_reform_{nj}} + \left(\beta^8 + \lambda\right) \mu_{nj} + \sigma_{\varepsilon^1} \eta_{nj} \end{aligned}$$

$d_{location\_reform_{nj}}$ is a dummy variable, γ is going to be a scaler in front of this dummy. From the first step of the regression (formula A2), we can have the estimation of the parameter $\hat{\gamma}$, and the estimation of the μ_{nj} which is

(A4)
$$\hat{\mu_{nj}} = distance\_to\_hukou_{nj} - \hat{distance\_to\_hukou_{nj}} = distance\_to\_hukou_{nj} - \hat{\gamma} d_{location\_reform_{nj}}$$

When we plug them back into the formula A3, the equation will naturally become

(A5)
$$\left(\frac{\beta^3}{\hat{\gamma}} + \beta^8\right) \hat{\gamma} d_{location\_reform_{nj}} + \left(\beta^8 + \lambda\right) \hat{\mu_{nj}} + \sigma_{\varepsilon^1} \eta_{nj}$$

. But if we update formula A5 with equation A4

(A6)
$$\begin{aligned} \left(\frac{\beta^3}{\hat{\gamma}} + \beta^8\right) \hat{\gamma} d_{location\_reform} + \left(\beta^8 + \lambda\right) \hat{\mu_{nj}} + \sigma_{\varepsilon^1} \eta_{nj} = \\ \beta^3 d_{location\_reform} + \beta^8 \left(\hat{\gamma} d_{location\_reform} + \hat{\mu_{nj}}\right) + \lambda \hat{\mu_{nj}} + \sigma_{\varepsilon^1} \eta_{nj} = \\ \beta^3 d_{location\_reform} + \beta^8 \left(\hat{\gamma} d_{location\_reform} + distance\_to\_hukou_{nj} - \hat{\gamma} d_{location\_reform_{nj}}\right) + \lambda \hat{\mu_{nj}} + \sigma_{\varepsilon^1} \eta_{nj} = \\ \beta^3 d_{location\_reform} + \beta^8 distance\_to\_hukou_{nj} + \lambda \hat{\mu_{nj}} + \sigma_{\varepsilon^1} \eta_{nj} \end{aligned}$$

We can use either equation A3, or equation A6 to do the simulated estimation of the regression. . However, there is a trade-off. If we apply formula A3, we can not identify the single parameter β^3 and β^8, but we can reduce one dimension to release computation and memory burden for a large dataset mixed logit model, with heavy burden for simulation. If the job is only to check endogeneity, I would prefer to use fewer variables to save computation burden. The RUM under this concern should be

(A7)

$$U_{nj}^{Mixe\_control} = \beta_n^1 \log rain + \beta_n^2 temprature + \tilde{\beta^3}\hat{\gamma} d_{location\_reform_{nj}} + \beta^4 d_{hukou\_at\_same\_city} + \beta^5 \log noise\_pollution + \beta^6 \log neighbor\_relationship + \beta^7 \log neighbor\_helpfulness + \beta_n^9 \log average\_disposable\_income + \tilde{\lambda}\hat{\mu_{nj}} + \sigma_{\varepsilon^1}\eta_{nj} + \varepsilon_{nj}^2$$

where $\tilde{\beta^3} \equiv \left(\frac{\beta^3}{\gamma} + \beta^8\right)$, and $\tilde{\lambda} \equiv \left(\beta^8 + \lambda\right)$. [9]

From the first stage, the regression of $\log distance\_to\_hukou = \gamma d_{location\_reform} + \mu_{nj}$ shows an estimation of γ is 6.7992^{***} with a standard error 0.003. After plugging $\hat{\gamma} = 6.7992$ and data of the residuals of $\hat{\mu_{nj}} = \log distance\_to\_hukou - \hat{\gamma} d_{location\_reform}$, we could have the following results for the regression of A7, which are shown in Table A2.

The above results demonstrate that not only are the estimations of the mixed parameters not significant (except for s_{11}), but also that this setting of the control function approach makes the parameters of temperature and noise pollution less significant. We can not say that endogeneity exists in one's choice of the distance of *hukou* location to their current living location. Rather, we should be more confident in saying that it is the type of hukou that affects the location j where one chooses to live.

[9] To make the results consistently comparable with the results in the main section, I kept the same names of the parameters for the other variables.

Table A1—Sets of MNL Models

For Each Location j	MNL(1) Parameter (t)	MNL(2) Parameter (t)	MNL(3) Parameter (t)	Mixed of (3) Parameter (t)
Log yearly average rainfall	-0.08 (-1.76)	-0.08 (-1.75)	-0.08 (-1.79)	0.14 (2.74)
Yearly average temperature	0.03 (2.46)	0.03 (2.48)	0.03 (2.51)	-0.03 (-2.15)
Hukou (trial) reform dummy(implemented=1)	0.41 (4.65)	0.42 (4.66)	0.42 (4.76)	0.42 (4.87)
average housing areas	-0.01 (-1.57)	-0.01 (-1.55)	-0.01 (-1.59)	-0.001 (-0.35)
Hukou at the same prefectural-level city dummy (same=1)	1.18 (12.24)	1.18 (12.24)	1.18 (12.21)	1.37 (15.19)
Log air pollution	0.04 (0.45)	-	-	-
Log water pollution	0.005 (0.07)	-	-	-
Log noise pollution	-0.09 (-1.13)	-0.07 (-1.12)	-	-
Log general environment	-0.01 (-0.25)	-	-	-
Log food access	0.01 (0.1)	-	-	-
Log infrastructure level	-0.01 (-0.21)	-	-	-
Log safety level	0.08 (1.09)	0.08 (1.1)	0.10 (1.39)	0.08 (1.04)
Log neighbor relations	-0.45 (-3.8)	-0.46 (-3.9)	-0.44 (-3.79)	-0.45 (-3.56)
Log Support from Neighbors	0.35 (2.3)	0.35 (2.38)	0.37 (2.48)	0.41 (2.52)
Log distance to *hukou*	-2.08 (-96.48)	-2.08 (-96.56)	-2.08 (-96.58)	-2.15 (-99.79)
Log average disposable income	1.17 (20.5)	1.16 (21.63)	1.14 (22.73)	1.17 (25.04)
s_{11}	-	-	-	0.42 (6.36)
s_{21}	-	-	-	-0.02 (-0.52)
s_{22}	-	-	-	0.21 (18.29)
Number of cases	11783	11783	11783	11783
Number of choices	469	469	469	469
Log likelihood function	-10813.22	-10813.44	-10814.07	-10733.55
Log likelihood function at null	-72472.55	-72472.55	-72472.55	-72472.55
Goodness of Fit 1-LL/LL_null	0.85	0.85	0.85	0.85

Table A2—Mixed Logit with Control Function Approach Results for Location Choices in China

For Each Location j	Parameter	t
Log yearly average rainfall	0.12	2.50
Yearly average temperature	-0.01	-0.93
Hukou (trial) reform dummy(implemented=1) with scaler $\hat{\gamma}$	-2.14	-79.86
Hukou at the same prefectural-level city dummy (same=1)	1.38	15.14
Log noise pollution	-0.07	-1.10
Log Support from Neighbors	-0.48	-3.72
Log neighbor relations	0.42	2.65
Log average disposable income	1.14	9.56
Residuals $\hat{\mu_{nj}}$	-2.20	-93.76
s_{11}	0.17	3.39
s_{21}	0.20	13.34
s_{22}	0.15	0.52
s_{31}	0.02	0.25
s_{32}	0.04	0.46
s_{33}	-0.64	-0.83
s_{41}	-0.02	-0.04
s_{42}	-0.46	-0.43
s_{43}	-0.02	-0.02
s_{44}	-0.01	-0.01
Number of cases	11783	
Number of choices	469	
Log likelihood function	-11547.04	
Log likelihood function at null	-72472.55	
Goodness of Fit 1-LL/LL_null	0.84067	