Health and Hukou Survey Documentation

Samantha A. Vortherms

1 Funding and Institutional Support

The Chinese Resident Health and *Hukou* Survey (CHHS)ran from 2014 until 2016. The survey was funded through funds from a National Science Foundation Doctoral Dissertation Research Improvement Grant (PI: Melanie Manion,co-PI: Samantha A. Vortherms) and a China Medical Board grant (PI: Gordon G. Liu, Junior Collaborator: Samantha A. Vortherms). Institutional support for the Health and Hukou Survey came from the China Center for Health Economic Research at the National School of Development, Peking University. Additional support came from Hunan University. The research program was reviewed by Peking University and approved by the University of Wisconsin–Madison's IRB (on file with authors). The authors are grateful for the entire research team that implemented the survey, including, in particular, Guan Haijing and Yao Yao of the CCHER.

2 Initial Plan and Adjustments

The initial plan for CHHS involved surveying four cities: Beijing, Changsha, Guangzhou, and Hangzhou. These four cities were selected non-randomly to provide explicit comparisons. Beijing is widely recognized as having the most valuable *hukou*, with higher access to prestigious universities and access to higher quality welfare and housing markets. But Beijing is largely unique, with the only possible comparison city being Shanghai. To provides more representation, the team planned on including three provincial capitals, two coastal and one inland, to capture variation in migrant populations. The research team ran into difficulties in two of the four cities and were unable to complete the needed surveys for analysis in those cities. In Guangzhou, rising tensions between the city government and institutions of higher education lead to severe restrictions on academic research programs. Survey enumerators from a different survey program, for example, were detained by police as a means of restricting research activities. Because of these increased risks, I decided to halt data collection. With only 100 surveys completed, we had insufficient data for analysis. In Hangzhou, the university where most of the survey enumerators lived closed housing suddenly well ahead of the G-20 meeting being held there two months after survey deployment. Local police also implemented a city clean up, encouraging migrant workers to leave the city to reduce traffic and "unwanted" populations in the city ahead of the summit. This made data collection impossible in the time frame allowed.

3 Sampling Procedure

Target population: Individuals between the ages of 18 and 60 who have resided in the selected address for the last three months and are not registered local urban residents. Individuals living in institutionalized settings, such as school dorms, factory dorms on factory campuses, hospitals, and prisons are excluded. Private dorms, such as a private apartment housing several factory workers located off campus and in non-factory owned building are included.

Table 1: Beijing Districts and their population break down

District	Total population (2010)	Non-local population	%	Rural Population	%
东城区	919253	219609	23.89	135246	14.71
西城区	1243315	327084	26.31	209425	16.84
朝阳区	3545137	1514822	42.73	1076356	30.36
丰台区	2112162	812713	38.48	676294	32.01
石景山区	616083	206493	33.52	123109	19.98
海淀区	3280670	1256145	38.29	915602	27.90
门头沟区	290476	47275	16.28	93039	32.02
房山区	944832	195099	20.65	495118	52.40
通州区	1184256	435173	36.75	619368	52.30
顺义区	876620	278721	31.79	523445	59.71
昌平区	1660501	847067	51.01	767124	46.19
大兴区	1365112	644057	47.18	837687	61.36
怀柔区	372887	102649	27.53	242136	64.93
平谷区	415958	48883	11.75	239462	57.56
密云县	467680	69438	14.85	305205	65.25
延庆县	317426	39305	12.38	188939	59.52

Beijing pilot study

For the pilot stage in Beijing, the target sample was 200 respondents sampled randomly through a multi-level process in one county in Beijing.

Stage 1: County Selection

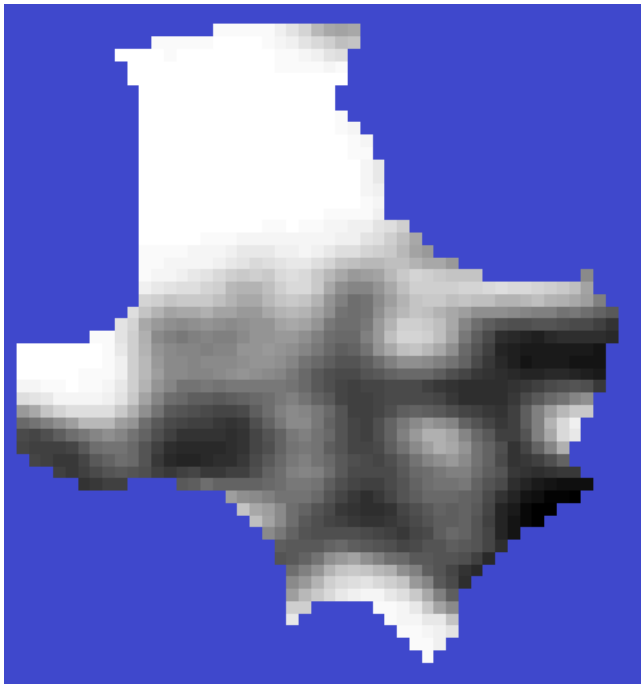
Given the relatively small size of the population, one county was randomly selected for the pilot phase of the study. Each city district and county, the total population, non-local population, and rural populations were enumerated (Table **??**). Given the target population (non local urban populations), city districts or counties with more than 70% urban populations during the 2010 census were excluded. As seen in Table **??**, this means four central city districts were excluded because they had relatively few rural residents. This decision to reduce the scope of possible counties potentially biases the end sample in a downward direction: if one of the primary variables of interest is willingness to pay for *hukou*, non-local residents who live in the city center may have a higher ability to pay for such a status transfer, but will be excluded from the final sample. The desire to ensure a significant number of possible respondents with access to land-use rights drove the decision to focus on the rural threshold above other possible decisions.

Because sampling prioritized non-locals and rural populations as well as the quickly shifting nature of non-local residents, the selection of the county was based on total population, rather than non-local population. With the assistance of Stata 13 and the SAMPLEPPS package, one county was randomly selected with probability proportional to size of total population.

Stage 2: Areas within county

Ideally, in order to go below the county level, we would have a comprehensive list of neighborhoods, units, addresses, or other community organizations with population density information from which to sample from. These types of lists are often incorrect, have significant coverage error through lack of coverage, and generally do not properly reflect the densities of non-local populations, this study does not rely on these formal lists.

Figure 1: Tongzhou District divided by Night Light data



Instead, in order to sample the SSU, the entire area of the county is divided in to 30" square geographical units. The 30" square units were drawn using the NOAA Night Light Intensity data. The Night Light data is a series of dots separated by one geographic minute (of both longitude and latitude) and rates each point based on the average intensity of stable lights at night over the past year on a scale from 0-63. These points mark the center of the SSU (the geographic square is drawn around the point). Figure 1 illustrates the division of the district by the satellite data. Each of the pixels in the picture is one SSU. In the selected county, there are over 1,300 of these SSU. For each square, the higher the light intensity, the higher the population density. In Figure 1 the white areas represent more urbanized areas with higher population density.

Because light intensity is correlated with population, we can use this measure as a proxy for population size. Again with the assistance of Stata, 50 of the SSU were selected with probability proportional to size based on light intensity. In the area selection stage, error is introduced through light bleeding: not all squares that have a non-zero light rating have housing units, either with no housing or with industrial zones. Of the 50 selected, 27 contained housing units. Figure 2 provides an example of two squares (not actually in the sample for privacy reasons) with the longitude/latitude defined square superimposed over the Google map.

Within the selected squares, each building was counted and enumerated. Using both Google Maps and Baidu maps, commercial buildings such as malls were excluded from this list. Using a random number generator, 20-40 buildings were selected (with replacement when the buildings were high rises). Each selected building was then given a random selection table generated using excel's RANDBETWEEN function (Figure 3). Interviewers were trained to randomly select the entrance, floor, and apartment using these random tables. The first row is the number of entrances/floors/apartments and the second row identifies which of the entrances/floors/apartments the enumerators pick. Using the tables in Figure 3 as an example, if the selected building has 4 entrances, the survey interviewers selected the 2nd door and recorded on the sampling sheet this selection. Most entrances, floors, and doors are numbered and interviewers were instructed to number the doors on a given floor with 1 being the first door to the left of

Figure 2: Example of a SSU with Google Map

Figure 3: Random Tables for selection of entrance, floor, apartment door

单元	1	2	3	4	5	6	7	8及以上
选	1	2	3	2	1	5	5	8

层	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20及以上
选	1	1	1	1	2	5	2	5	9	2	9	9	12	2	14	14	13	5	5	8

门	1	2	3	4	5	6	7	8及以上
选	1	2	3	3	2	1	2	8

the stairs. If doors are not numbered, interviewers were instructed to count the doors from left to right, with the left most door from the stairs/elevator being door number 1. When the housing unit was a one story "ping fang," a random table was included for doors in case the individual building had multiple doors. Individual respondents within the house were selected based on a standard Kish tables procedure. Sampling sheets had one of the 8 standard Kish tables and these sampling sheets were randomized then attached to one individual map+random table sheet. Given that Kish table procedures require the household informant to know each member's birthday, should the informant not know, the alternative selection method was the "most recent birthday," where the selected respondent was the person who most recently celebrated their birthday.

Full Survey Process

In order to reduce error in the full survey and improve sampling procedures from the pilot survey, I adjusted the sampling plan slightly for the full survey deployment. This section outlines the changes. The same logic was used throughout to maintain consistency but the full survey sample selection involved less coverage error in the third stage and prioritized identification of non-local populations in the PPS process. This design decision came in part because of the different contexts of Guangzhou, Hangzhou, and Changsha: in Guangzhou, the non-local population is expected to be large and the rural population

is expected to be more difficult to identify whereas in Changsha, the reverse is expected to be true.

Stage 1: County Selection

For the county selection, the same general process as Beijing was used. First, counties and districts were enumerated as well as their total population, non-local population (inter- and intra-provincial migrants included but inter-city population excluded), and rural population. Urban central districts and counties that did not have a substantial number of either rural residents or non-local residents, based on an 80% rule (counties were excluded if local urban population was more than 80% as household sampling is inappropriate in these cases). Second, the list of eligible counties was inputted into Stata 13. Using the samplepps package, two counties were randomly selected for inclusion according to the following procedure: first, counties were sorted by non-local population; second, a random seed was set for each city (Stata code: set seed [*random number between 0 and* 2^{31}]); third, two counties were randomly selected with probability proportional to size of non-local population (Stata code: samplepps *sample*, n(2) s(wdpop)). The data used and Stata code, including the random seed used, are available from the author for Hangzhou and Changsha.[1]

Stage 2: Township Selection

The process for township selection followed a similar strategy as the county selection, except that within counties, no townships were excluded. Township present during the 2010 Census were included and made the sample frame. First townships (街道 and 镇) in each county were enumerated based on 2010 census data. Second, the number of non-local residents were enumerated. The definition of non-local at this level, however, is a proxy for real non-local populations: the township level census data I had access to counts all individuals who are not registered at the given address as "non-local." This means that the non-local population includes migrants from other provinces, migrants from other cities within the same province, and individuals who moved across counties within the given city of interest. This last category is not counted as "non-local" for this study and thus this data introduces some error into the sample. Intra-city migration, however, should be correlated with inter-city migration as individuals that move within a city are often migrating for the same jobs as migrants from outside of the city, suggesting that these data inflate the number of non-locals but should remain correlated with the non-local populations of interest.

Similar to the county selection, the list of townships and their non-local populations were entered into Stata 13 and using the same process of setting a random seed and then selecting three townships with probability proportional to size of non-local population was used to select three townships. No townships were excluded from the sample, meaning the final sample should be representative at the county level.

Stage 3: Spatial Sampling within Townships

Unfortunately, no reliable maps or GIS materials are publicly available for the township level. That is, there are no ArcGIS shapefiles available to divide the county to the township level. Additionally maps at the township level were often out of date or too vague to pin-point precise boundaries of townships. One possibility was to map out each of the villages and neighborhoods (村 and 社区) to define the area of the township, but not only are lists of villages and neighborhoods quickly outdated given changing administrative structures, but also village boundaries are not sufficiently distinguishable. For example, in some locations, one village was clearly defined separate from its neighboring villages (Figure 4), but others are spread out with no clear boundary (Figure 5).

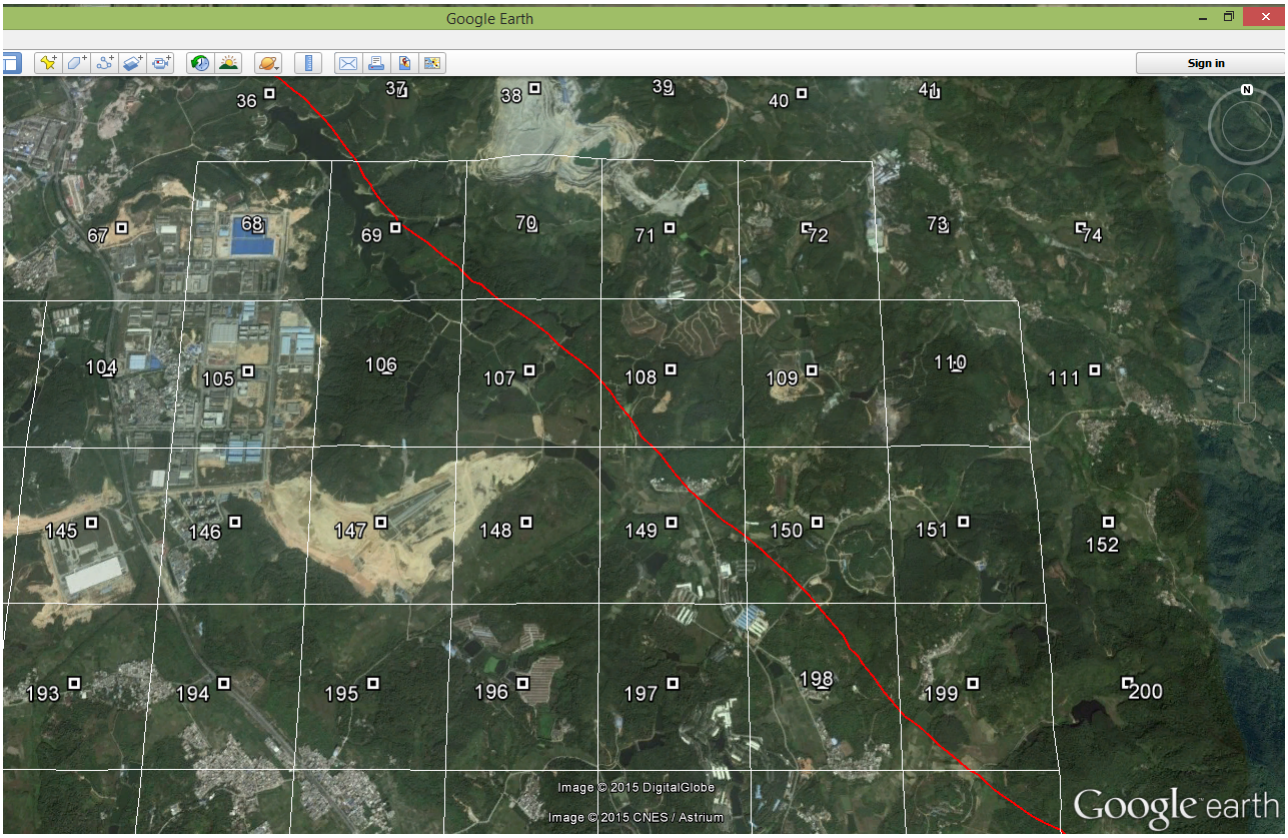
Township villages and neighborhoods were enumerated using lists online. The approximate location of each village was identified using Google Earth by first searching on Baidu Maps then using geographic

[1] Guangzhou county selection followed the same process, but a computer crash prevented me from recording the original random seed numbers.

Figure 4: Village with clear boundary

Figure 5: Villages without clear boundaries

Figure 6: Narrowing down satellite data to township size



features to record the location on Google Earth through an iterative process. Using these village and neighborhood locations as well as published maps online, I drew an approximate outline of the township boundaries on Google Earth. In some locations, this was simple to do given geographic features and the township boundaries corresponding to roads, rivers, and canals. In some locations, however, I used my best judgment for the township boundary. I specifically erred on the side of inclusion, suggesting that the final spatial sample frame may have some over-coverage.

Spatial sampling is most efficient when combined with information related to population density within each spatial unit. Given that population numbers at the township level are limited to census data and data below the township level are often non-existent, I used the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) data on night-light intensity to proxy for population density. I first convert the DMSP-OLS data from raster data (the raw data form available on the National Centers for Environmental Information's website) to point data, which generates a single point identifiable with longitude and latitude coordinates at the center of the 30" GIS square the score averages the data from. These points are then transferred into Google Earth with the assistance of GEPath 1.4.6. I then re-create the original DMSP-OLS 30" grid squares around the point-based data, again with the assistance of GEPath 1.4.6 (add or subtract 15" to each point in both longitude and latitude and run with draw path, suppress the data labels). Using the estimated township boundaries and the grid, I narrowed down the points to each township.

Figure 6 shows an example of the satellite data grid superimposed over a map with a township boundary included. The grid lines match the satellite data points and the diagonal line is the township boundary. Squares that crossed the township boundary were included (again, erring on over-inclusion)

Figure 7: Example of a construction site 1 year and 6 months before survey implementation

while those outside were not. The list of refined points included in a township were then enumerated with their night light intensity data.

Given the nature of the grids drawn, not all grid squares included populated areas, i.e. some grid squares contained no houses. In order to reduce sampling error by drawing grid squares with no population, I went through each of the grid squares to classify them as having housing or not. For inclusion, the square had to have at least five recognizable houses. Grid squares with no housing and mountains, such as square 69 in Figure 6, were excluded. Similarly, grid squares with only clearly identifiable factories were also excluded.[2]

I also used discretion in identifying construction zones, which often have high levels of visible light at night, but do not have individuals living there. Construction sites were identified through two means of checking locations. First, when going through to identify housing buildings, locations with the tell-tale yellow cranes were flagged as possible construction sites. Second, I used the time-slider function in Google Earth to identify if the existing structures had been constructed within the last year. The most recent images available were mostly from 6 months before survey implementation.[3] While it might be desirable to capture populations moving in to a new set of high rises and the exclusion of these squares may result in under-coverage of the sample frame, the inclusion of new construction sites or areas under construction within three months of the survey result in over-coverage and sampling error. Therefore, the following rule was used. If a cluster of buildings did not exist one year previously and had cranes 6 months before survey implementation, it is highly unlikely the buildings would be occupied for a full 3 months before survey implementation. Figure 7 shows two pictures of the same construction site. The first image is twelve months before survey implementation and the second is six months before survey implementation.

Within each township, three spatial squares were selected with probability proportional to size with the night light intensity score as the weight. The selection of buildings, entrances, floors, and doors is the same as the Beijing sample. One additional complication arose in Guangzhou and Changsha, which is that some of the selected areas had buildings difficult to identify (Figure 9. In these cases, using Google Earth, I identified GPS coordinates at the front of the building and included it on the sampling sheet (Figure 9). Students were then instructed to download the "GPS Coordinates" application on their cell

[2]While it might have been desirable to include these areas as many migrants live within their factories, the implementation of a survey within factory grounds is not feasible; researchers must obtain formal permission from the factory owners, which is often not granted and requires inside connections to get. Because of this, we define our target population as individuals not living in institutions, which includes factories and schools.

[3]In some instances, even newer images were available, approximately 3 months before survey implementation.

Figure 8: Difficult to identify building

Note: These GPS coordinates are an example and not included in the final sample.

phones to find the GPS coordinates. Once arriving at the coordinates, they were instructed to go to the door closest to where they were standing. Respondent selection within the household is also done with Kish tables.

4 Final Sample

The final sample yielded 484 and 460 respondents in Changsha and Beijing, respectively, with 472 and 436 who completed all of the questions necessary for this analysis.[4] As expected, the two samples have substantively different groups of migrants. The Beijing sample included more inter-provincial migrants, in part because Beijing is a directly administered city and physically smaller than Hunan province, and the Changsha sample included more intra-provincial migrants. Even with the concentration of Hunan migrants in the Changsha sample, there are migrants from almost every province in both samples.[5] The sample is balanced between men and women, with slightly more women than men, primarily due to non-response. Most respondents have rural registration, with 77 percent and 66 percent rural in Beijing and Changsha, respectively. Overall, the sample is highly educated, well above the average, with 31 percent of the Beijing sample and 45 percent of the Changsha sample with some college education,[6] in part because the sample skews young (average age is 35) and in part because of specific neighborhoods randomly selected into the Changsha sample. Almost all of the respondents, 96 percent, are ethnic Han Chinese.

5 Enumerator Training

Enumerators in both Beijing and Changsha were undergraduate students or medical students. We recruited survey enumerators through social media advertisements and through faculty networks. All students received formal training in human subjects protection, sampling procedures, and the questionnaire specifically. Training materials, including handbook and slides for training are available upon request.

[4]There was significant item non-response to the income question, which reduced the sample in both locations.

[5]The only sending regions not present in the samples are Tibet, Yunnan, and Qinghai—three minority-heavy provinces.

[6]Defined as either undergraduate or associates degree program.

Figure 9: GPS Coordinates Screen Shot

Note: These GPS coordinates are an example and not included in the final sample. Students were required to use this software only and not a Chinese version of the software, as China encrypts its GPS codes. The location given and the application used to find them had to correspond with international standard GPS, not Chinese encrypted GPS.

6 Data Collection and Processing

For safety reasons, survey enumerators collected data in pairs. The research team would go to one neighborhood, with the assistance of assigned maps and sampling sheets to complete the final stage of household and individual selection. During the pilot study, we identified a problem with sending two men as one survey team as some respondents felt uncomfortable with two men entering their home. After this issue was identified, each pair was either one man and one woman or two women. Student enumerators were provided with the introduction letter identifying the survey as a Peking University survey program and their personal identification as students.

Respondents were offered a gift with university branding. In the initial pilot program, the gift was a pen and bookmark set. In the full sample survey, we made glass drinking bottles with the CCHER's name and Peking University's logo.

The survey data was collected using a paper-based instrument, with enumerators reading questions and answer categories to respondents. The survey made use of a significant number of show cards to help in understanding. Each experimental treatment was randomized by hand within each paper questionnaire booklet.

Completed questionnaires were digitized with the assistance of Qualtrics with the assistance of the enumerators and additional undergraduate students.