## PAHCPC Survey, China 2012

**Sampling and Fieldwork Report**

# Research Center for Contemporary China Peking University

# January, 2013

|  |  |
| --- | --- |
|  |  |
|  |  |
| **Principle Investigator** | Shen Mingming, Professor |
| **Sampling Specialists** | Shen Mingming, Professor  Yan Jie, Associate Professor |
| **Project Director** | Chai Jingjing，Yan Jie |
| **Project Assistant** | Liang Yu, Hu Sihui |

Table of Contents

[**I. Overview** 1](#_Toc348111154)

[1.1 Length of Fieldwork 1](#_Toc348111155)

[1.2 General Outturn 1](#_Toc348111156)

[**II. Sampling** 1](#_Toc348111157)

[2.1 Sample Population Overview 1](#_Toc348111158)

[2.2 Sample Population Exclusions 1](#_Toc348111159)

[2.3 Sampling Method 2](#_Toc348111160)

[2.4 Stratification Method 2](#_Toc348111161)

[2.5 Sampling Unit 3](#_Toc348111162)

[2.6 Sampling Frames 3](#_Toc348111163)

[2.7 Selection Method in each Stage 4](#_Toc348111164)

[2.8 Sample Size 4](#_Toc348111165)

[**III. Fieldwork Report** 5](#_Toc348111166)

[3.1 Supervisor Training 5](#_Toc348111167)

[3.2 Interviewer Training 6](#_Toc348111168)

[3.3 Official Implementation 6](#_Toc348111169)

[3.3.1 Project Team 6](#_Toc348111170)

[3.3.2 Implementation Process 7](#_Toc348111171)

[3.3.3 Survey Verification 8](#_Toc348111172)

[3.4 Turnout at each PSU 10](#_Toc348111173)

[3.5 Reasons for Unsuccessful Visits 12](#_Toc348111174)

[3.6 Difficulties of Project Implementation 13](#_Toc348111175)

[**IV. Database Creation** 14](#_Toc348111176)

[4.1 Data Entry 14](#_Toc348111177)

[4.2 Sampling Verification 14](#_Toc348111178)

[4.3 Data Cleaning 14](#_Toc348111179)

[4.4 Database Creation 14](#_Toc348111180)

[4.5 Weighting 15](#_Toc348111181)

[4.5.1 Base Weight（Weighting Design） 15](#_Toc348111182)

[4.5.2 Post- Stratification and Weighting 17](#_Toc348111183)

[4.6 Compilation of Sampling and Implementation Reports 18](#_Toc348111184)

# **I. Overview**

## 1.1 Length of Fieldwork

Nov. 1 2012—Jan. 17 2013

## 1.2 General Outturn

Target sample size：5638

Sample drawn in the field：5424

Completed and valid interviews：3684

Response Rate：67.9%

# **II. Sampling**

## 2.1 Sample Population Overview

The target population covers adults between the ages of 18 and 70, who reside in all 31 provinces of the Chinese Mainland (Hong Kong, Macao and Taiwan are not included).

## 2.2 Sample Population Exclusions

Individuals who reside in the places listed below were not included in the study:

1. Military residential complexes
2. Residential units in compounds of Central Ministries
3. Embassies and consulates
4. Infrastructural buildings（i.e. Power Stations, Water Stations etc.）
5. Prisons
6. Tourist destinations and religious sites

Individuals with the following characteristics who were residing in valid residential complexes were not included in the study: Residents of Hong Kong and Macau, non-Chinese citizens.

## 2.3 Sampling Method

The sampling plan for the general public uses the “GPS Assisted Area Sampling Method[[1]](#footnote-1).” which incorporates population as a measure of size, stratification and multi-stage PPS (Probabilities Proportional to Size).

## 2.4 Stratification Method

For the purpose of allocating PSUs across all large regions with different levels of economic development, stratification based on 3 official division of regions in China（Coastal,11 provinces; Central, 8 provinces; Western,12 provinces) will be taken as the first step of the sampling process.

In addition, we have taken into consider the disparity between rural and urban areas in China’s social development. In order to lower the disparity between sampling from urban and rural areas in order to fulfill the research goal, this project will do a second stratification based on urban and rural characteristics within the first strata.

Therefore, there are 6 layers in total. In order to obtain a self-weighted sample, number of primary sampling units (PSUs) within each stratum is proportional to the population size of that stratum.

**Table 1. The population size of each stratum and their respective PSU**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Strata** | **The population of each strata, 2010**  **(person)** | **The 18-70 population of each strata, 2010**  **(person)** | **Percentage of total population** | **Number of PSUs that should be chosen** | **Number of PSUs that were chosen** | **Total PSUs in each strata** | **fpc\_psu** |
| 11 East Municipal Districts | 330383666 | 236757325 | 0.241 | 14.44 | 15 | 520 | 0.0288462 |
| 13 East county | 189660000 | 135912876 | 0.138 | 8.29 | 8 | 366 | 0.0218579 |
| 21 Central Municipal Districts | 196183034 | 140587369 | 0.143 | 8.58 | 9 | 416 | 0.0216346 |
| 23 Central county | 260701152 | 186821909 | 0.190 | 11.40 | 11 | 478 | 0.0230126 |
| 31 West Municipal Districts | 140102848 | 100399562 | 0.102 | 6.13 | 6 | 288 | 0.0208333 |
| 33 West county | 255360000 | 182994369 | 0.186 | 11.16 | 11 | 787 | 0.0139771 |
| **Total** | **1372390700** | **983473410** | **1** | **60** | **60** | **2855** | **0.021016** |

## 2.5 Sampling Unit

Primary Sampling Units (PSUs):

County level administrative units (municipal districts, county-level cities, counties)

Secondary Sampling Units (SSUs):

Half-square minutes (HSM) of latitude and longitude

Tertiary Sampling Units (TSUs):

Spatial square seconds (SSS), approximately 90m\*90m

Basic Sampling Units:

Dwellings in the sampled units

## 2.6 Sampling Frames

The sampling frame employed by the primary sampling unit will be taken from the name list of all county-level administrative units and population statistics taken from the <National sub-county Population Statistics 2010> (published by the Ministry of Public Security, November, 2011 by Qunzhong Publishing House in Beijing).

A GIS dataset will be established as the sampling frame for this project, which will be based on 1) county level population data from the 2010 Census,[[2]](#footnote-2) 2) the most recent and detailed (paper and electronic) maps, 3) the highest possible resolution images from Google Earth. Based on the abive information, the population density is then calculated for each of the HSMs in county level units.

## 2.7 Selection Method in each Stage

**PSU:** Out of 2,856 counties in China, 60 counties will be chosen by stratified PPS.

**SSU:** Three HSMs will be selected by PPS within each of the selected county.

**TSU:** The measures of size (HSM) used at these stages are the density of the population per sampling unit.

**QSU:**Within each of the selected HSM，the number of SSSs （90m\*90m）is calculated based on the population density, and then selected the SSSs simple randomly.

Trained surveyors equipped with GPS receivers are then sent to locate and enumerate the sampled “spatial square seconds” (SSS). For maintaining equal probabilities of selection across households, all dwellings enumerated in the SSSs will be included in the sample. Using system sampling, we will draw 27 dwellings in each HSM.

**Respondents:** Respondents will be selected from dwellings using the Kish Grid method[[3]](#footnote-3).

## 2.8 Sample Size

To satisfy a confidence level of 95%, with a permissible error of 3%, and taking into consideration factors such as the outcomes of multi-stage sampling (deff), empty responses (caused by reasons such as unqualified individuals, empty residential units, interview refusals, language barriers etc.), a total of 4860 residential units were planned to be selected, with an effective sample size of 3200.

In actuality, 5424 residential units were selected, containing an effective sample size of 3684.

# **III. Fieldwork Report**

## 3.1 Supervisor Training

**3.1.1 Supervisor**

The supervisors for this project are all employees of and were trained by the Research Center for Contemporary China (RCCC) at Peking University.

**3.1.2 Training Period**

Systematic training sessions were held on Oct. 23 2012 and Oct. 26 2012 for the selected candidates at the RCCC meeting room and led by senior researchers in relevant fields.

**3.1.3 Areas Included in Supervisor Training**

* Project Background；
* Basic Interview Techniques；
* Specific Requirements for the Project；
* How to use GPS and the Sample Area Selection Process；
* Address and Interviewee Selection Process；
* Overview and Description of each Question on the Survey；
* Classroom Exercises；
* Project Implementation Procedures；
* Quality Control Procedures；
* Code of Conduct and Safety Protocols。

## 3.2 Interviewer Training

**3.2.1 Interviewer**

The interviewers for the project were all college students in the surveyed area and were trained, according to the Interviewer Manual, by their supervisors.

**3.2.2 Training Period**

Due to the fact that each supervisor departed on different days, there was no uniform training period for the interviewers, but each supervisor was required to perform 1 full day of systematic training for the interviewers.

**3.2.3 Areas included in Interviewer Training**

* Project Background；
* Basic Interviewing Techniques；
* Specific Requirements for the Project；
* Interviewee Selection Process；
* Overview and Description of each Question on the Survey；
* Classroom Exercises；
* Home Interview Procedures；
* Quality Control Procedures；
* Code of Conduct and Safety Protocols。

## 3.3 Official Implementation

### 3.3.1 Project Team

Principle Investigator：Shen Mingming

Project Director：Chai Jingjing，Yan Jie

Project Operation Director：Liang Yu

Supervisors：13 Employees from RCCC

Interviewers：Approximately 227 undergraduate students

Quality Inspectors：Project Operation Director and Supervisors

### 3.3.2 Implementation Process

**（1）Interviewee Address Sampling**

First, supervisors will proceed to the half-square minutes as determined by longitudes and latitudes prescribed by RCCC. Within the half-square minutes, supervisors will be given the relevant longitudes and latitudes to identify and approach the targeted small-grid cluster. Supervisors then will begin the process of address sampling in accordance with sampling protocols as dictated by RCCC. If the process results in more than 60 valid addresses within the small-grid cluster, further selection processes should be performed so as to reduce this number to below 60. On the other hand, if less than 30 valid addresses are given then follow-up samplings should be performed on backup small-grid clusters according to the prescribed order, until one such cluster gives more than 30 addresses. If none of the backup clusters yield more than 30 valid addresses, then a general investigation should be performed on said half-square minutes. If all valid addresses in the half-square minutes amount to more than 60, then further sampling should be performed so as to reduce it to below 60, otherwise the half-square minutes will be deemed invalid.

**（2）Interviewee Sampling**

After interviewers enter a valid address, they will identify all individuals who have resided in the address for more than 30 days, and record them into the Kish grid. All individuals who ages of 18 and 70 are then separated by gender, and then ordered according to age, from oldest to youngest so that the Kish method can be used to select one interviewee.

**（3）Supervisors’ Daily Responsibilities**

* Arranging interviews and ensuring quality
* Leading teams into the targeted communities to perform interviews, collecting surveys, checking amount of completed surveys and comparing it to amount of surveys given out. The reason behind any discrepancy should be sought out and dealt with.
* Completed surveys must be checked daily so that complications can be discovered and noted, preventing similar problems from surfacing in the future. Supervisors must sign their initials on valid surveys. Interviewee names should be recorded on the back of each survey, while exact addresses must be recorded on the front. After the interview, interviewers should deliver completed surveys and interview records to the supervisor within 24 hours.
* During the interview, supervisors must take note of the progression of the interview so as to properly fill out work journals and interview summaries as well as arrange all completed surveys. Forms such as interview progression forms, and completed survey tables must be copied with clear handwriting, error-free.
* Report to RCCC the progression of fieldwork. Any problems that may arise must be reported to RCCC to ask for further instructions.

### 3.3.3 Survey Verification

Supervisors are responsible for on-site supervision of interviewers during the interviewing process as well as the verification of all completed surveys. They are also in charge of ensuring the quality and quantity of surveys, truthfulness of interviews (ensuring that the chosen interviewee is interviewed), information accuracy (correctly recording interviewee’s answers), and completeness of responses (all questions must be asked).

After the interview, interviewers must immediately verify responses of the survey then subsequently sign their initials on the relevant surveys and handing it to their respective supervisors.

Supervisors must check surveys for any problems such that they can be prevented in the future. They must also sign their names on all surveys deemed to be valid.

**Verification process includes checking**

* Whether the interviewer entered the correct address determined by the supervisor;
* Whether the interviewer used the Kish grid to choose interviewees；
* Whether the selected interviewee was interviewed；
* Whether there were empty responses to questions；
* Whether there were incorrect responses；
* Whether there are unclear and logically flawed answers。

## 3.4 Turnout at each PSU

**Table 2. Completed Samples at Each PSU**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PSU id** | **Completed Samples** | **Actual Sampled Addresses** | **Completion Percentage** | **Province name** | **Prefectural**  **City name** | **County name** | **County id** | **Strata id** |
| 1156 | 59 | 104 | 56.7% | 北京市 | 北京市 | 朝阳区 | 110105 | 11 |
| 1206 | 51 | 116 | 44.0% | 天津市 | 天津市 | 滨海新区 | 120116 | 11 |
| 1301 | 51 | 81 | 63.0% | 河北省 | 廊坊市 | 永清县 | 131023 | 13 |
| 1407 | 46 | 69 | 66.7% | 山西省 | 长治市 | 平顺县 | 140425 | 23 |
| 1422 | 59 | 73 | 69.9% | 山西省 | 晋中市 | 平遥县 | 140728 | 23 |
| 1408 | 51 | 62 | 71.0% | 山西省 | 运城市 | 新绛县 | 140825 | 23 |
| 1409 | 44 | 90 | 65.6% | 山西省 | 临汾市 | 洪洞县 | 141024 | 23 |
| 2157 | 67 | 104 | 65.4% | 辽宁省 | 铁岭市 | 昌图县 | 211224 | 13 |
| 2135 | 68 | 93 | 72.0% | 辽宁省 | 朝阳市 | 北票市 | 211381 | 11 |
| 2227 | 56 | 118 | 47.5% | 吉林省 | 吉林市 | 昌邑区 | 220202 | 21 |
| 2349 | 47 | 61 | 65.6% | 黑龙江省 | 齐齐哈尔市 | 甘南县 | 230225 | 23 |
| 2323 | 40 | 56 | 83.9% | 黑龙江省 | 牡丹江市 | 宁安市 | 231084 | 21 |
| 3102 | 56 | 116 | 48.3% | 上海市 | 上海市 | 宝山区 | 310113 | 11 |
| 3110 | 37 | 90 | 41.1% | 上海市 | 上海市 | 松江区 | 310117 | 11 |
| 3236 | 74 | 61 | 86.9% | 江苏省 | 徐州市 | 丰县 | 320321 | 13 |
| 3228 | 75 | 114 | 70.2% | 江苏省 | 连云港市 | 赣榆县 | 320721 | 13 |
| 3203 | 53 | 99 | 75.8% | 江苏省 | 淮安市 | 淮阴区 | 320804 | 11 |
| 3204 | 80 | 97 | 76.3% | 江苏省 | 泰州市 | 兴化市 | 321281 | 11 |
| 3337 | 91 | 122 | 74.6% | 浙江省 | 丽水市 | 莲都区 | 331102 | 11 |
| 3411 | 67 | 105 | 63.8% | 安徽省 | 铜陵市 | 狮子山区 | 340703 | 21 |
| 3438 | 57 | 84 | 67.9% | 安徽省 | 阜阳市 | 太和县 | 341222 | 23 |
| 3505 | 59 | 86 | 68.6% | 福建省 | 厦门市 | 翔安区 | 350213 | 11 |
| 3512 | 65 | 109 | 59.6% | 福建省 | 莆田市 | 城厢区 | 350302 | 11 |
| 3650 | 74 | 108 | 68.5% | 江西省 | 吉安市 | 永丰县 | 360825 | 23 |
| 3751 | 73 | 119 | 70.6% | 山东省 | 潍坊市 | 安丘市 | 370784 | 11 |
| 3724 | 58 | 77 | 66.2% | 山东省 | 济宁市 | 曲阜市 | 370881 | 11 |
| 3725 | 76 | 63 | 68.3% | 山东省 | 日照市 | 东港区 | 371102 | 11 |
| 3717 | 51 | 96 | 60.4% | 山东省 | 临沂市 | 沂水县 | 371323 | 13 |
| 3729 | 52 | 109 | 69.7% | 山东省 | 临沂市 | 临沭县 | 371329 | 13 |
| 3713 | 84 | 77 | 67.5% | 山东省 | 德州市 | 德城区 | 371402 | 11 |
| 3718 | 43 | 103 | 70.9% | 山东省 | 菏泽市 | 定陶县 | 371727 | 13 |
| 4114 | 82 | 103 | 79.6% | 河南省 | 漯河市 | 郾城区 | 411103 | 21 |
| 4147 | 66 | 83 | 85.5% | 河南省 | 信阳市 | 浉河区 | 411502 | 21 |
| 4139 | 71 | 98 | 67.3% | 河南省 | 周口市 | 西华县 | 411622 | 23 |
| 4240 | 61 | 97 | 78.4% | 湖北省 | 襄樊市 | 宜城市 | 420684 | 21 |
| 4215 | 76 | 71 | 85.9% | 湖北省 | 孝感市 | 云梦县 | 420923 | 23 |
| 4252 | 75 | 85 | 88.2% | 湖北省 | 随州市 | 天门市 | 429006 | 21 |
| 4316 | 88 | 117 | 75.2% | 湖南省 | 长沙市 | 岳麓区 | 430104 | 21 |
| 4353 | 56 | 85 | 74.1% | 湖南省 | 长沙市 | 宁乡县 | 430124 | 23 |
| 4341 | 66 | 85 | 77.6% | 湖南省 | 岳阳市 | 汨罗市 | 430681 | 21 |
| 4330 | 63 | 82 | 68.3% | 湖南省 | 怀化市 | 会同县 | 431225 | 23 |
| 4431 | 74 | 117 | 63.2% | 广东省 | 深圳市 | 龙岗区 | 440307 | 11 |
| 4445 | 33 | 67 | 49.3% | 广东省 | 揭阳市 | 惠来县 | 445224 | 13 |
| 4548 | 94 | 85 | 69.4% | 广西 | 南宁市 | 青秀区 | 450103 | 31 |
| 4546 | 74 | 83 | 89.2% | 广西 | 玉林市 | 陆川县 | 450922 | 33 |
| 4532 | 59 | 116 | 81.0% | 广西 | 贺州市 | 富川瑶族自治县 | 451123 | 33 |
| 5042 | 59 | 77 | 76.6% | 重庆市 |  | 石柱土家族自治县 | 500240 | 33 |
| 5154 | 43 | 103 | 63.1% | 四川省 | 绵阳市 | 梓潼县 | 510725 | 33 |
| 5158 | 58 | 120 | 45.0% | 四川省 | 绵阳市 | 江油市 | 510781 | 31 |
| 5143 | 65 | 62 | 69.4% | 四川省 | 遂宁市 | 船山区 | 510903 | 31 |
| 5160 | 50 | 76 | 76.3% | 四川省 | 内江市 | 资中县 | 511025 | 33 |
| 5144 | 54 | 68 | 73.5% | 四川省 | 乐山市 | 市中区 | 511102 | 31 |
| 5233 | 61 | 87 | 60.9% | 贵州省 | 贵阳市 | 乌当区 | 520112 | 31 |
| 5219 | 53 | 83 | 68.7% | 贵州省 | 安顺市 | 普定县 | 520422 | 33 |
| 5234 | 64 | 93 | 65.6% | 贵州省 | 黔东南 | 凯里市 | 522601 | 31 |
| 5220 | 57 | 92 | 69.6% | 贵州省 | 黔南 | 惠水县 | 522731 | 33 |
| 6159 | 52 | 68 | 76.5% | 陕西省 | 榆林市 | 定边县 | 610825 | 33 |
| 6255 | 58 | 85 | 57.6% | 甘肃省 | 白银市 | 靖远县 | 620421 | 33 |
| 6221 | 49 | 90 | 64.4% | 甘肃省 | 陇南市 | 成县 | 621221 | 33 |
| 6326 | 59 | 84 | 70.2% | 青海省 | 海东 | 互助土族自治县 | 632126 | 33 |
| **Total** | **3684** | **5424** | **67.9%** |  |  |  |  |  |

## 3.5 Reasons for Unsuccessful Visits

**Table 3. Reasons for unsuccessful visits**

|  |  |  |
| --- | --- | --- |
| **Status of visit** | **Amount** | **Percentage** |
| 1 Completed interview | **3684** | **67.9** |
| 2 Interview refusal at address household | 712 | 13.1 |
| 3 Interview refusal of selected interviewee | 287 | 5.3 |
| 4 Absent interviewee on several occasions | 92 | 1.7 |
| 5 Physical and language barriers/Inability to understand questions due to old age | 38 | .7 |
| 6 Non-contact, unknown eligibility | 610 | 11.2 |
| 14 Interviewed wrong person | 1 | .0 |
| **Total Valid Addresses** | **5424** | **100.0** |
| 7 Empty residential unit | 117 |  |
| 8 No occupants meet requirements | 82 |  |
| 9 Non-residential address | 1 |  |
| 10 No this address | 14 |  |
| **Total Invalid Addresses** | **214** |  |
| **Pre-planned Addresses** | **5638** |  |

## 3.6 Difficulties of Project Implementation

Complications over the course of the project are mainly due to：

（1）Refusal of interview by the whole community. In recent years, the number of high-end residential communities has increased in every city and thus they have a greater chance of being selected in sampling. These communities often require non-residents to register before entering the premises. Property management companies, worrying about residents potentially lodging complaints about letting outside interviewers disturbing their privacy, often deny entry to our field teams.

Furthermore, in normal residential buildings with lax security measures, residential units often have electronic security doors. Interviewers cannot explain the purpose of the project and visit in person but only through electronic speakers, thus greatly diminishing their ability to persuade residents as well as the residents’ perception of the interviewers’ trustworthiness. The problem of refusal of interview by all residents in an entire address or by the interviewee that has been selected is very severe, greatly impeding the work of the interviewers as they can only proceed with their work with the cooperation of the residents. Due to these two circumstances the completion rate of surveys and interviews has decreased dramatically.

（2）Language barriers caused by local dialects caused some trouble in certain southern cities. Although local college students were sought for in most cases, a few selected cases, particularly involving the elderly, were nonetheless subject to communication difficulties.

（3）A few questions on the survey required a certain degree of professional knowledge, further complicating effective communication, thereby effecting the progression of successful interviews. Responding to questions on the survey and interviews required much effort, particularly among interviewees with a lower level of education, as they were unable to understand key terms.

# **IV. Database Creation**

## 4.1 Data Entry

Entry of raw data was performed by specialists at RCCC, using appropriate computer software, beginning Jan. 4 2013. Specialized data entry software can, to a large extent, prevent incorrect data from being entered into the system. Furthermore, to improve the accuracy of entered data, RCCC uses the double data entry method so that two sources of data can be brought together and verified. This task was completed on Jan. 28 2012.

## 4.2 Sampling Verification

Three Peking University graduate students were responsible for verifying the validity of both complete and incomplete surveys. Specifically, they checked whether the interviewees were properly selected and whether the interviews were indeed conducted on the selected individuals.

## 4.3 Data Cleaning

Data cleansing began on Jan. 12 2013, headed by experienced research assistants at RCCC. The process was divided into 3 steps: firstly, whether data was properly entered into the system was verified through the double data entry method; secondly, incorrect data and logical fallacies within the responses were indentified; thirdly, after correcting all mistakes, the variables of the entire data set were double checked by separate individuals for any logical fallacies. This ensures the accuracy of analysis performed on the data in the future. Systemic errors can be identified and corrected according to the reliability of the input code. These tasks were completed on Feb. 6 2013.

## 4.4 Database Creation

On Feb. 6 2013, after finishing data cleansing for the completed surveys, the database creation process was initiated. Steps include changing variable tags and descriptions as well as entering sampling information such as stratification and PSU information. The final database and coding manual were completed on Feb. 8 2013.

## 4.5 Weighting

### 4.5.1 Base Weight（Weighting Design）

On Feb 5 2012, project team members and sampling specialists shared their opinions on data weighting, reaching a consensus on Feb 6 2012. Work on database weighting started immediately afterwards and was completed on Feb 8 2012.

Two weighting methods were used on the database. One of them, base weighting, is shown below, where the reciprocal of each interviewee’s selection probability (including weight of empty responses) is calculated and given the variable name “wt\_base”.

First, with each half-square minutes (SSU) considered one unit; the valid sampling addresses and valid addresses of each half-square minutes are tabulated. Then using the formula below, each address’s selection probability is calculated.



h represents layer number (6 layers total)

represents the number of PSUs in each layer

represents the number of SSUs sampled in each PSU. (For this project, 3 SSUs are sampled in each PSU)

represents each PSU’s size measurement (that is the population in each county level unit in 2010)

 represents each SSU’s size measurement，(that is the number of residential addresses in each half-square minutes)

represents the number of valid sampled addresses in each chosen SSU

c represents the average household size in 2010.

 represents the number of addresses

 represents the population at each layer

Because this project selected only inhabited residential addresses, we need to divide population at each layer by the average household size c to obtain the number of residential addresses and thus be able to calculate each address’s selection probability. The average number of households in 2010 was obtained from the 2010 census, where c equals 3.09.

Next, using the number of valid participants in each address, the probability of selection for each individual is calculated.



q2 is a variable in the database representing the number of valid participants in each address.

Then，the base weight for each individual is calculated



Subsequently, the empty response weight is calculated.



Valid Sampling Size

Valid Response Sample Size

Lastly，the base weight is calculated



The three variables, wt, wt\_respo and wt\_base are all present in the database, and their usage depends on how empty responses will be accounted for. If weighting is used to account for empty responses, then wt\_base can be used directly to add weighting to the variables. If other methods are used, then further adjustments can be performed using wt as a basis.

### 4.5.2 Post- Stratification and Weighting

Post-fieldwork stratification and weighting is based upon statistics found in figures pertaining to population in the 2010 Census. The population is stratified according to the same age groups and gender as displayed in Table 4.

Even though there were two groups where weighting dipped below 0.8, but under this stratification, weighting of most groups remained within the 0.8-1.2 range.

In the database, the post-stratification weighting variable is labeled under wt\_ps, researchers may choose to use either wt\_base or wt\_ps for analysis.

**Table 4. Weighting based on Age and Gender**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Population | | percentage | |
| Age | Male | Female | Male | Female |
| 2010 Census Data |  |  |  |  |
| 18-29 | 136669534 | 134055576 | 0.27349 | 0.27712 |
| 30-39 | 109912926 | 105251236 | 0.21995 | 0.21757 |
| 40-49 | 117385096 | 112963421 | 0.23490 | 0.23352 |
| 50-59 | 81446172 | 78619473 | 0.16298 | 0.16252 |
| 60-70 | 54307502 | 52862474 | 0.10868 | 0.10928 |
| Research Sample |  |  |  |  |
| 18-29 | 334 | 326 | 0.18576 | 0.17285 |
| 30-39 | 330 | 370 | 0.18354 | 0.19618 |
| 40-49 | 431 | 463 | 0.23971 | 0.24549 |
| 50-59 | 331 | 359 | 0.18409 | 0.19035 |
| 60-70 | 372 | 368 | 0.20690 | 0.19512 |
| Weighted (census/sample) |  |  |  |  |
| 18-29 | 409190.2216 | 411213.4233 | 1.47227 | 1.60322 |
| 30-39 | 333069.4727 | 284462.8 | 1.19839 | 1.10905 |
| 40-49 | 272355.2111 | 243981.4708 | 0.97994 | 0.95122 |
| 50-59 | 246060.9426 | 218995.7465 | 0.88533 | 0.85380 |
| 60-70 | 145987.9086 | 143648.0272 | 0.52527 | 0.56005 |

Data Source：Census Office of the State Council，National Bureau of Statistics Population and Employment Statistics Division, *Census of the People’s Republic of China (Vol. 1)*，Beijing：China Statistics Press，2012.4，Table 3-1a.

## 4.6 Compilation of Sampling and Implementation Reports

Sampling and implementation reports were finalized on Jan 9 2013. At this point, all relevant documents and figures have been prepared and can be used for future analysis and research.

1. Landry, F. Pierre and Shen Mingming,. “Reaching Migrants in Survey Research: The Use of The Global Positioning System to Reduce Coverage Bias in China”. *Political Analysis*, 2005,Vol 13, 1-22

   The “GPS Assisted Area Sampling”method was created by the Research Centre for Contemporary China at Peking University. Its main advantage is its ability to solve the complications arising from traditional sampling based on household registration (such as: outdated household registration information, excessive empty households, exclusion of migrant population etc.) as well as its difficulties arising from positioning in area sampling [↑](#footnote-ref-1)
2. Although data from the 2000 Census seemed out of date, but it must be used because that it is the only available national dataset that has migrants information throughout province, city, county and township level units. [↑](#footnote-ref-2)
3. Leslie Kish, *Survey Sampling* (New York: John Wiley & Sons. 1965), Pp.398-399. [↑](#footnote-ref-3)